

高等学校教材· 计算机应用

多维数据分析原理与应用

姚家奕 等编著



清华大学出版社



高等学校教材· 计算机应用



本书特色：

- ◆ 从数据库理论、多维数据分析技术和多维数据分析应用三方面加以介绍
- ◆ 提供从理论到应用的一整套数据仓库 OLAP 解决方案
- ◆ 每章均配有阅读资料和案例分析



ISBN 7-302-08377-0



9 787302 083771 >

定价：20.00 元

高等学校教材·计算机应用

多维数据分析原理与应用

姚家奕 等编著

清华大学出版社
北京

内 容 简 介

数据仓库技术的研究和应用是当前数据管理领域的热点。随着信息化建设在企业、国家政府部门以及社会其他领域的不断普及,数据分析和决策支持系统的开发与建设逐渐纳入到信息化建设的过程中来。面向组织决策层的数据管理技术是决策者挖掘组织内部和组织外部经营管理信息价值的有利工具和方法。

本书从数据仓库理论、多维数据分析技术和多维数据分析应用3个方面,以某市地税局数据仓库的成功实施为背景案例,系统地介绍了基于Microsoft SQL Server 2000的OLAP多维数据引擎——Analysis Services构建数据仓库多维数据集的全过程,旨在为读者提供从理论到应用的一整套数据仓库OLAP解决方案的清晰视图。

本书可以作为信息管理专业和计算机专业本科生或硕士生教材,也可以作为从事数据仓库建设和研究人员的参考书。

版权所有,翻印必究。

本书封面贴有清华大学出版社激光防伪标签,无标签者不得销售。

图书在版编目(CIP)数据

多维数据分析原理与应用/姚家奕等编著. —北京:清华大学出版社,2004.5

高等学校教材·计算机应用

ISBN 7-302-08377-0

I. 多… II. 姚… III. 数据库系统—系统分析—高等学校—教材 IV. TP311.13

中国版本图书馆CIP数据核字(2004)第026139号

出版者:清华大学出版社

<http://www.tup.com.cn>

社总机:010-62770175

地 址:北京清华大学学研大厦

邮 编:100084

客户服务:010-62776969

组稿编辑:丁 岭

文稿编辑:闫红梅

印装者:北京鑫海金奥胶印有限公司

发 行 者:新华书店总店北京发行所

开 本:185×260 印张:12 字数:298千字

版 次:2004年5月第1版 2004年5月第1次印刷

书 号:ISBN 7-302-08377-0/TP·6029

印 数:1~5000

定 价:20.00元

本书如存在文字不清、漏印以及缺页、倒页、脱页等印装质量问题,请与清华大学出版社出版部联系调换。联系电话:(010) 62770175-3103或(010) 62795704

前 言

目前，基于通用数据库技术的联机事务处理（OLTP）系统在大多数现代企业中的应用已经比较成熟，如企业管理信息系统、办公自动化系统、电子商务远程交易系统等。它们为企业信息化迈出了坚实的一步，由此带来的日常业务处理效率的大大提高给企业贴上了信息化的标签。然而，这仅仅是企业信息化的初期。

在 OLTP 系统逐渐走向成熟的进程中，业务高效处理的企业群体日益暴露出来的是决策滞后的新问题。抢先一步的企业已经清楚地认识到：因决策不当造成的人员、设备、资金等的不均衡配置渐渐冲淡了因业务处理效率的提高带来的效益；数百万美元的生意因为你犹豫了一分钟而被决策更快、更准的竞争对手拿走。这些信息技术引发的决策快速问题，常常令企业决策者措手不及。企业运作是一个循环移动的过程，任何一个复杂的企业运作模式都可简化为：业务处理——决策——业务处理。业务处理是执行前次的决策，决策是推进新的业务处理。实际企业中，这个循环链处处存在，小到推销员推销产品，大到整个企业的高利润目标。这个链条的任何一个环节出现问题，都会导致企业的正常运转。OLTP 系统完善的现代竞争机制下，业务处理迅速而高效，决策频率也大大加快，决策者应接不暇是难免的。

另一个推动因素是决策深度问题。在企业得益于丰富的数据资源的同时，管理者更清楚地认识到，要竞争就决不能满足于简单的数据收集和整理。市场、销售、财务、产品、顾客、服务、供货、存货、销货等方面的数据的内在联系逐渐成为企业各层管理者致力挖掘的焦点。于是，如何从大量的历史数据中有效地提取有用的数据成为各企业所普遍面临的问题。这时，企业管理者已不满足于进行“本月销售额是多少，比上个月增长了多少？”等简单的数据处理，而是要清楚哪些客户会成为我们的利润源泉？哪些产品最有潜力？将会失去哪些市场？哪些地方花费最高？哪些市场运行得最好，为什么？我们进入新市场的胜算有几何，为什么？这些都是辅助决策问题，也是企业要追求的高利润问题。

信息技术引发的问题常常会推动信息技术的前进，正是由于这些困扰企业的实际问题的推动，一种新兴的数据仓库技术诞生了。OLTP 系统积累的大量历史数据为数据仓库提供了数据支持；成熟的通用数据库技术衍生出来的数据仓库专用数据库技术为数据仓库提供了技术背景；对决策质量、效率要求日益提高的现代企业为数据仓库提供了广阔的应用领域。在信息化高度发展的美国企业或企业联合中，大规模的集成数据库系统已经蕴涵着数据仓库思想，只是没有被提出来而已。当决策迟缓渐渐成为企业发展的“瓶颈”时，这种新技术被提上日程，许多专家、学者开始致力于如何应用数据仓库技术辅助企业决策的研究。

本由由北京交通大学经管学院姚家奕负责本书的统稿和定稿，吕希艳负责编写第 1 章和第 2 章，姜海负责编写第 3 章~第 5 章，姚家奕负责编写第 6 章和第 7 章，青岛大学于忠清负责编写第 8 章。

在本书的编写过程中，北京交通大学陈景艳教授对全书的结构和内容提出了许多合理的建议，同时得到多位同行和老师的大力支持，在此一并表示衷心的感谢！书中的“阅读材料”和“案例分析”摘自《计算机世界》和一些网站，在此，也对这些媒体表示感谢！书中内容若有不妥或错误之处，恳请各位读者不吝指正。

本书可以作为信息管理专业和计算机专业本科教材，也可以作为从事数据仓库建设和研究人员的参考书。

目 录

第 1 章 数据仓库体系结构	1
1.1 初识数据仓库	1
1.2 数据仓库解决的问题	1
1.3 一个成功的例子	2
1.4 数据仓库中心——操作型数据还是分析型数据	4
1.5 数据仓库体系结构	5
1.5.1 数据仓库软件工具集	5
1.5.2 体系结构的稳定性	6
1.5.3 维数据结构	6
1.6 数据仓库体系结构的计算模式	7
1.7 以数据为中心	7
1.8 数据仓库 workflow	8
1.9 数据仓库体系结构的基本特点	9
1.10 一个现实的问题	10
1.11 小结	10
阅读资料	11
A 何时需要数据仓库	11
B 数据仓库会带来什么	12
案例分析	13
零售行业数据仓库决策支持系统	13
第 2 章 数据仓库的基本特征	15
2.1 业务系统和决策支持系统	15
2.2 数据仓库的数据源	16
2.3 数据仓库的维	18
2.4 数据仓库的事实数据	20
2.5 数据仓库的多维数据模型	21
2.6 数据立方体	22
2.7 数据立方体中的数据聚合	23
2.8 数据仓库的职业角色	23
2.9 小结	24
阅读资料	25
什么是数据集市	25

案例分析.....	28
数据仓库技术在移动通信领域的应用.....	28
第 3 章 联机分析处理系统.....	32
3.1 OLAP 的实质.....	32
3.1.1 OLAP 系统与 OLTP 系统的区别.....	32
3.1.2 OLAP 系统的组成.....	33
3.2 使用维和度量进行数据分析.....	33
3.3 多维视图.....	34
3.4 维表.....	35
3.4.1 维表的分类.....	35
3.4.2 结构维的特点.....	37
3.4.3 星型模型.....	38
3.4.4 雪花模型.....	38
3.4.5 雪花模型与星型模型的对比.....	39
3.5 事实表.....	39
3.6 多维数据集.....	40
3.7 ROLAP、MOLAP 和 HOLAP.....	40
3.7.1 ROLAP.....	41
3.7.2 索引.....	41
3.7.3 MOLAP.....	43
3.7.4 MOLAP 与 ROLAP 的比较.....	44
3.7.5 HOLAP.....	45
3.8 小结.....	45
阅读资料.....	46
基于供应链数据仓库的 OLAP 数据挖掘 (上).....	46
案例分析.....	48
数据仓库与 CRM.....	48
第 4 章 多维数据集的分析与建立.....	51
4.1 多维数据集.....	51
4.1.1 多维数据集的基本结构.....	51
4.1.2 虚拟多维数据结构.....	53
4.1.3 多维数据结构的分区存储.....	54
4.2 OLAP 服务管理的基本术语.....	55
4.2.1 聚合.....	55
4.2.2 分区.....	56
4.2.3 钻取.....	57
4.2.4 角色.....	57

4.2.5 虚拟立方体	58
4.2.6 OLAP 服务控制台	58
4.3 多维数据集结构的更新	59
4.3.1 OLAP 存储方式回顾	59
4.3.2 多维数据集结构的更新方式	60
4.3.3 增量更新	61
4.3.4 刷新更新	62
4.3.5 完整处理	63
4.3.6 刷新共享维	63
4.3.7 检查刷新后的结果	64
4.4 多维扩展语言	64
4.4.1 MDX 语言的五要素	65
4.4.2 MDX 应用示例	67
4.5 小结	69
阅读资料	70
基于供应链数据仓库的 OLAP 数据挖掘 (下)	70
案例分析	72
数据仓库——在“啤酒与尿布”中挖掘 (上)	72
第 5 章 OLAP 数据挖掘技术	74
5.1 OLAP 数据挖掘技术简介	74
5.2 OLAP 多维数据集	74
5.3 数据挖掘的主要功能	76
5.4 期望的 OLAP 挖掘功能	77
5.5 OLAP 数据挖掘的有效实施	78
5.5.1 基于 OLAP 的数据特征和比较	78
5.5.2 基于 OLAP 的关联	79
5.5.3 基于 OLAP 的分类	80
5.5.4 基于 OLAP 的预测	81
5.5.5 基于 OLAP 的聚类分析	82
5.5.6 回滚和比较挖掘分析	82
5.6 小结	83
阅读资料	84
数据挖掘的研究现状	84
案例分析	87
数据仓库——在“啤酒与尿布”中挖掘 (下)	87
第 6 章 Analysis Services 多维数据引擎	90
6.1 启动 Analysis Services	90

6.2	建立数据库和数据源	91
6.2.1	建立数据库结构	91
6.2.2	建立数据源	91
6.3	建立多维数据集	92
6.3.1	向多维数据集添加度量值	92
6.3.2	建立时间维度	92
6.3.3	建立雪花模型维度	94
6.3.4	建立星型模型维度	95
6.3.5	建立父子维度	95
6.3.6	完成多维数据集	96
6.4	编辑多维数据集	96
6.4.1	在多维数据集编辑器内编辑多维数据集	96
6.4.2	向现有多维数据集添加维度	97
6.5	设计存储和处理多维数据集	97
6.6	定义立方体的存取权限	98
6.6.1	创建多维数据集角色	99
6.6.2	创建数据库角色	100
6.7	定义钻取选项	103
6.7.1	启用多维数据集的钻取功能	103
6.7.2	给角色提供钻取权限	103
6.8	小结	105
	阅读资料	106
	SAS 快速建库的方法论	106
	案例分析	108
	财政金融行业的数据仓库决策支持系统	108
第 7 章	i Analyze 智能工具简介	110
7.1	i Analyze 的产生背景和目标	110
7.2	i Analyze 的设计方案和系统需求	110
7.3	i Analyze 的体系结构与访问安全性	111
7.4	i Analyze 工具的操作	112
7.4.1	连接分析服务器	112
7.4.2	界面功能	112
7.5	i Analyze 智能解决方案	127
7.5.1	用户需求和数据源分析	128
7.5.2	设计分析模型	129
7.6	小结	132
	阅读资料	133
	决策树的后期修剪技术	133

案例分析.....	136
加拿大用 Sybase 技术做数据统计.....	136
第 8 章 地税数据仓库.....	138
8.1 地税数据仓库的实施背景.....	138
8.2 实施过程.....	139
8.3 开发环境与目标.....	139
8.4 数据仓库的总体结构模型.....	140
8.5 税款开票数据立方体.....	141
8.5.1 分析目的.....	141
8.5.2 分析模型.....	141
8.6 费入库数据立方体.....	147
8.6.1 分析目的.....	147
8.6.2 分析模型.....	147
8.6.3 表结构和抽取规则.....	149
8.7 小结.....	152
阅读资料.....	153
细说 BI——商业智能.....	153
案例分析.....	155
综合医疗系统中的数据仓库解决方案.....	155
附录 A 数据仓库相关技术常用名词解释.....	166
附录 B 常用的 MDX 函数.....	169
附录 C 国外数据仓库解决方案简介.....	179
参考文献.....	182

第 1 章 数据仓库体系结构

数据仓库技术是随着计算机技术的飞速发展而产生的。传统的数据库技术是单一的数据资源，即数据库为中心，进行事务处理、批处理到决策分析等各种类型的数据处理工作。近年来，由于计算机和网络的广泛应用，计算开始向两个不同的方向拓展，一是广度计算，一是深度计算。广度计算的含义是把计算机的应用范围尽量扩大，同时实现广泛的数据交流，互联网就是广度计算的特征；另一方面就是人们对以往计算机的简单数据操作，提出了更高的要求，希望计算机能够更多地参与数据分析与决策的制定等领域。特别是数据库处理可以大致划分为两大类：操作型处理和分析型处理（或信息型处理）。这种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，从而由原来的以单一数据库为中心的数据环境发展为一种以数据仓库为基础的体系化环境。

1.1 初识数据仓库

按“数据仓库之父”W.H.Inmon 的定义，数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集合，用来支持管理人员的决策。以主题为导向的数据仓库是围绕着企业的基本实体设计的，如政府主管行业的城市数据仓库设计主题可考虑城市的人口总数、人均工资水平、失业人数、税收情况及污染指数等，它们都是内在密切联系的，这样有利于实现数据的关系化、规则化，且可提供动态、多维的数据统计、数据查询，建立关系数据模型，预测发展趋势等，若仅靠现有的在线事务处理系统是无法做到这一点的。数据集成需通过设计实现命名协议、关键字、关系、编码的一致；高度的集成性强调企业必须规划好大量的设计工作，才能真正成功地建立数据仓库。非易失性指数据仓库里的数据不进行实时更新，数据经过复杂的提取过程后定期进入数据仓库。一旦进入，就不能再由用户进行更新。随时间而变化，指数据仓库的设计要按不同时段组织数据，如按月、按季或按年。从历史数据中生成不同级别的汇总数据（或阶段数据），这一点很重要。

1.2 数据仓库解决的问题

数据仓库技术可解决事务处理系统处理不了的决策问题，具有动态集成和综合处理能力。具体能解决如下问题：

(1) 解决“业绩下降 10%”与“业绩上升 15%”的问题

在线事务处理系统进行数据抽取时，传统上以放任自流的态度处理整个软、硬件体系，这会由于层层抽取的不一致性产生严重的“蜘蛛网”（spider web）问题，导致对同一个问

题不同部门的结论不同，且可能相距甚远，如部门甲认为公司业绩下降 10%，而部门乙可能得出公司业绩上升 15%的结论。数据仓库技术通过给数据加上时基，分离原始数据与导出数据，消除同类数据的算法的差异，提高了数据的可信性，有效避免了“蜘蛛网”问题。

(2) 解决企业环境中多数据源及数据不一致性问题

数据仓库可以通过数据转移工具将位于不同的地理位置、不同平台、不同数据库中的数据按照一定的规则，高度集中在一个数据仓库中，达到充分利用各种数据源的目的。且在构建数据仓库的过程中，可充分考虑企业原环境数据的不一致性问题，将系统中不一致的数据，按数据的一致性原则转移到数据仓库中，从而保证数据仓库中数据的完全一致，这对作出正确的决策是至关重要的。例如，在电信市场，数据仓库可迅速而准确地向客户提供混合销售、留言、呼叫、等待等综合服务。

(3) 充分而高效地利用企业积累的大量历史数据

在传统在线事务处理系统中，历史数据大多被存储在光盘、磁带或其他大容量存储介质中，查询历史数据是费时、费力的事情，进行数据分析时就更不用说了，况且各年的数据可能存储在不同的介质上，导致数据处理效率较低。数据仓库中主要存储的就是历史数据和大量的汇总数据，因而基于历史数据的分析在数据仓库系统中则显得非常方便，且效率显著提高，因为在数据仓库中存储的就是大量预先处理的汇总数据。

(4) 进行辅助决策分析

基于数据挖掘、数据抽取和决策支持上发展起来的数据仓库技术，使得决策支持系统进入实用化阶段。一个具有高效的数据分析功能的数据仓库可能会告诉用户（当然，用户事先并不知道它能做到这些）：如何防止丢失有利的客户？查明哪些客户会离开？为什么会离开？提高周末的票价将会带来什么好处？需要多大成本实现一项新服务？2001年意外伤亡索赔份额为何增长一倍？等等。

1.3 一个成功的例子

本书将以海威决策支持系统（HDC3X）为例，分析数据仓库技术在商业企业信息化中的成功应用。该系统对企业积累的大量业务数据的处理游刃有余，使企业人员能快速、交互、方便有效地从这些大量杂乱无章的数据中获取有意义的信息，决策者能利用现有数据指导企业决策和发掘企业的竞争优势。现简介如下：

系统将现代企业需要的两种不同的数据环境清晰地分为：日常操作型的数据环境和决策支持信息的数据环境。将企业整个数据仓库体系结构划分为 4 层：操作层、数据仓库层、部门层（或数据集市层）及个体层。

数据仓库体系结构的操作层即指日常操作型的数据环境，如企业的管理信息系统、办公自动化系统等。

数据仓库层建构在决策支持信息的数据环境中，包括以下 3 个主要功能。

(1) 数据抽取：负责从外部数据源抽取数据、区分数据、复制数据或重新定义数据格式等，以备装入数据仓库。

(2) 数据支持：负责数据仓库的内部数据服务，提供的服务包括数据存储的组织、数

据的维护、数据的分发、数据仓库的例行维护等,这些工作利用了数据库管理系统(DBMS)的功能。

(3) 工具平台:是面向不同类型的用户的数据仓库的前端。主要由查询生成工具、多维分析工具和数据挖掘工具等工具集组成,以实现决策支持系统的各种要求。

部门层可以认为是数据仓库的具体应用领域,用以存放大量的导出数据。该系统包括市场、销售、客户、商品、信用风险、消费者、库存及人力资源等多个方面导出数据,体现了系统面向主题的设计方案。

在市场分析方面,能使市场管理人员通过分析新产品的投放情况,不断优化产品定价和促销方式,并结合市场促销渠道和市场投资回报率等影响市场成败的因素,提出全面开拓市场,增加利润和提高影响力的多种解决方案供决策者选择;在销售分析方面,允许销售人员追踪企业的收入进度、比较自己的库存、其他库存和以前的库存,从而对来自市场的反馈,反应更加迅速和准确;在客户分析方面,提供强大的分析工具,能处理大量的产品、顾客的数据细节,对用户进行分类和分析,使理解顾客购买模型成为可能。在此基础上,商家可以提供更有针对性的商品,以最少的花费,满足老顾客和潜在顾客的要求;在商品分析方面,借助对销售、产品分类、价格和购买趋势等分析,制订有效的商品配比计划,并提供图表和统计报告来评价这个计划的成功与否,而且,基于数据仓库技术的海威决策支持系统还可以进行需求预测、商品潜力、地区购买力以及其他各式各样的商业因素分析,例如:地区周消费量、最佳卖点、季节性消费和异常消费等,进而提供一切可能的商机;在信用风险分析方面,该数据仓库解决方案能使企业的信用评测部门从客户交易、现金循环周期、风险得分、非常规购销模型等分析中得出日常经营的风险及即将面临的风险。通过这些信息,管理者能够提前做出准备,作出科学准确的决策。这样,当他们面临这些风险时,便可游刃有余;在消费者分析方面,通过对消费者市场,如:居民生活水平、消费能力、品牌忠诚度、企业形象等的分析,管理者可以更加准确地对未来消费市场的热点、冷点、转向、走势等行情进行预测,然后信心十足地做出决策;在库存分析方面,不仅能有效地控制资产和商品流通量,对比销售与成本的差价,分析销售百分比供其他分析使用,而且通过对库存历史数据的大量即时查询分析挖掘隐性的数据,辅助完成对仓储式销售的最有利的厂址选择、再分配的最佳方案和端到端传递的最优路线等的决策,使库存经营整体协调达到最优;在人力资源分析方面,数据仓库通过对超时和工作量的分析,糟糕的业绩和优秀人员的鉴别,人力满员和人力紧缺的预测等,辅助管理者进行人力资源的最佳配置。

数据仓库体系结构的个体层借助于一个或多个导出数据完成大多数的启发式分析,实际上它更倾向于执行支持,供最高层管理者使用。

基于数据仓库技术的商业企业决策支持系统解决方案,能有效地解决如何管理商业企业中浩如烟海的数据,以及如何从中提取有用的信息等问题,把企业网络中不同信息环境中的商业数据集中起来,生成不同级别的数据,汇总到集成的数据库中,并提供各种数据挖掘、抽取和分析技术,能有效地辅助企业管理者运筹帷幄,决胜千里,实现在商海竞争中不败的雄心壮志。该决策支持系统是数据仓库技术在商业企业中成功应用的一个典型例子。

1.4 数据仓库中心——操作型数据还是分析型数据

近年来，科学技术飞速发展。许多几年前还不知为何物的技术如今已经风靡全球，其中包括利用电子数据交换(EDI)和因特网商务(电子商务)技术进行的即时交易。当前的市场形势促使公司企业保持旺盛活力。其关键就在于适时掌握准确信息，利用这些信息作出正确决策以及最终及时贯彻这些决策。

对于公司企业来说，获得和利用信息的方式就是建立覆盖公司所有部门的企业综合信息系统。拥有了这种企业综合信息系统的公司更有可能在现代经济中保持竞争优势和盈利能力。而没能成功建立起这种系统的公司则可能始终在现代经济中苦苦挣扎。企业综合信息系统需要使用多种技术、不同类型的计算和操作系统以及各类数据仓库。要想在一个综合系统中把所有部分结合为一个整体，首先需要对各种应用程序和基础结构作周密规划，以使公司提出的各项任务都能高效率完成。这种系统的核心是存储于各个数据源中的所有信息的可访问性。这些信息必须是一致、准确和经过准确定义的，它们源于两种类型的数据：操作型数据和分析型数据。

操作型数据是处于不断变换和更新之中的，属于动态数据。订单输入数据库中的当前订单就是这类数据的一个例子。操作数据所代表的是某一时间点的当前信息。操作数据可以表明待发订单的状况、活期存款账户的当前余额或当前脱销产品的数量。这类数据可以表明某件事情的现状，而且它们在任何时候都是处于变化之中的。

另一方面，分析型数据是历史数据，通常都不会随着时间的推移而发生变化，因此属于静态数据。只有在原始信息错误的情况下，分析数据才会有变动。某个时刻的销售额是最终数据，不可逆的。在这个时刻，信息变成了静态，因此可以从动态数据源迁移到静态数据源。分析数据可用于查看跨时间段的信息。例如，我们可以查看1月份的总销售额或者开发人员在过去6个月里的工资变化。操作型数据与分析型数据的主要区别如表1-1所示。

表 1-1 操作型数据与分析型数据的区别

操作型数据	分析型数据
表示业务处理的动态情况	表示业务处理的静态情况
在存取的一瞬间是正确的	代表过去的数据
可更新，由录入人员或经过专门培训的输入事务而更新	不可更新，终端用户的访问权限常常是只读的
处理细节问题	受到更多关注的是结论性的数据，是综合的，或是提炼的
操作需求事先可知，系统可按预计的工作量进行优化	操作需求事先不知道，永远不知道下一步用户要做什么
有许多事务，每个事务影响数据的一小部分	有数目不多的一些查询，每个查询可访问大量数据
对性能要求高	对性能要求宽松
面向应用，支持日常操作	面向分析，支持管理需求
用户不必理解数据库，只是输入数据	用户需要理解数据库，以从数据中得出有意义的结论

分析数据通常采自操作数据，可用来对公司在某个时期的运营状况做全面分析。这种数据必须准确、可访问，并且能以实用的格式显示出来。SQL Server 2000 Analysis Services 将使用户拥有建立此类数据的能力。用于分析的数据可以包括企业内部数据和企业外购数据。SQL Server 2000 可以访问各种数据源的数据，把它们转变为一种统一的格式，校验数据的完整性，最终把数据存储到一个联机分析处理服务器 Analysis Services 中供使用者方便访问。通过对象的链接和嵌入数据库(OLE DB)接口——这是一种 API 接口，可访问任何数据库中的任何信息——可以建立 SQL Server 数据仓库和多维数据中心，从而按需求访问任何信息。

1.5 数据仓库体系结构

数据仓库不是软件产品也不是应用程序，它是系统体系结构。体系结构是按照优先原则对方法进行的安排。这些原则使得通过客户机、网络和数据库软件而执行业务过程成为可能。而这些业务过程又为处理基本业务的规则提供了启示。假如数据仓库市场有各种有用的和有创意的软件产品，而它们都宣称自己是解决方案，那么就有必要考虑一下了。毫无疑问，这些软件产品都有值得注意的优点。尽管许多产品都暗示自己是体系结构，因为没有一种产品是孤立的。但是，关键问题依然存在，即有一种产品本身就是体系结构。

作为体系结构，数据仓库包含了许多产品，如图 1-1 所示。每一种产品都有除数据仓库操作以外的功能。简单地说，数据仓库体系结构是蓝图，是一种安排或示意。这种体系结构提供基础设施，使得传达客户信息的企业应用程序得以实现。数据仓库系统体系结构提供一种模式，在这种模式中，应用程序之间互相紧密连接，而且与硬件、操作系统、数据库、网络及接口软件集成起来，并与业务过程交叉引用。

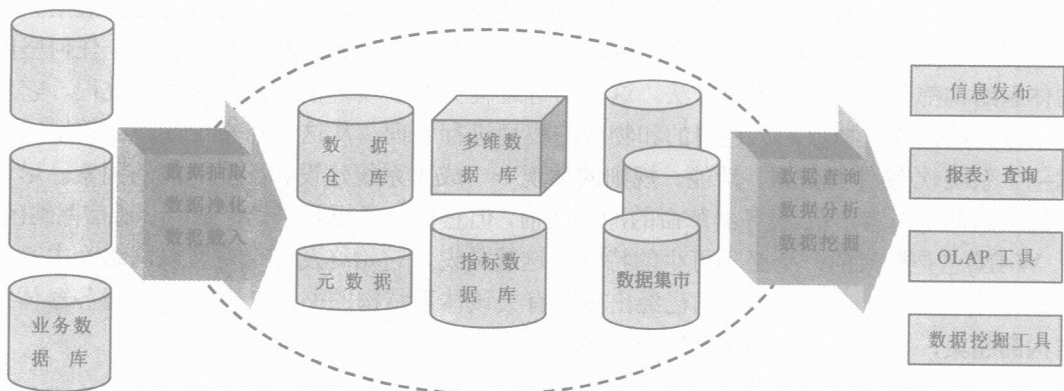


图 1-1 数据仓库的体系结构

1.5.1 数据仓库软件工具集

简单地说，数据仓库软件的工具集包括：一个数据库引擎，这个数据库引擎包括已经