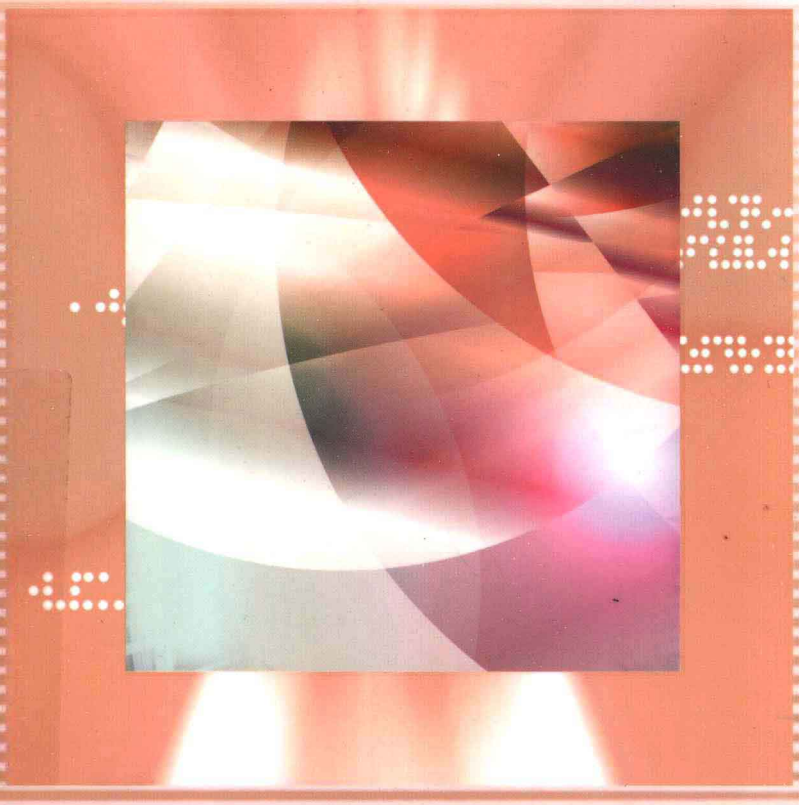
普通高等教育“十一五”国家级规划教材

数据仓库与数据分析教程

Data Warehouse and Data Analysis

王 珊 李翠平 李盛恩 等 编著



高等教育出版社
HIGHER EDUCATION PRESS

普通高等教育“十一五”国家级规划教材

数据仓库与数据分析教程

Shuju Cangku yu Shuju Fenxi Jiaocheng

王珊　李翠平 李盛恩　等　编著

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

数据仓库技术和数据分析技术是信息领域的核心技术之一，是基于海量数据的决策支持系统体系化环境的核心。

本书详尽地介绍了数据仓库和数据分析技术的基本概念和基本原理，建立数据仓库和进行数据分析的方法和过程。全书分为数据仓库技术篇、联机分析处理技术篇、数据挖掘技术篇三部分，共 10 章。附录中介绍了一些典型的数据仓库产品和工具。

本书可以作为高等学校计算机专业、信息管理专业以及其他相关专业本科生和研究生的教材和参考书，也可以作为企事业单位信息管理部门及相关行业从事数据库和数据仓库的研究与开发人员、数据分析人员和管理人员的参考资料。

图书在版编目（CIP）数据

数据仓库与数据分析教程/王珊等编著. —北京：高等教育出版社，2012.8

ISBN 978-7-04-034130-0

Ⅰ.①数… Ⅱ.①王… Ⅲ.①数据库系统-高等学校-教材 Ⅳ.①TP311.13

中国版本图书馆 CIP 数据核字（2012）第 108248 号

策划编辑 倪文慧	责任编辑 倪文慧	封面设计 张 志	版式设计 王艳红
责任校对 陈旭颖	责任印制 张福涛		

出版发行	高等教育出版社	咨询电话	400-810-0598
社 址	北京市西城区德外大街 4 号	网 址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印 刷	人民教育出版社印刷厂	网上订购	http://www.landraco.com
开 本	787×1092 1/16		http://www.landraco.com.cn
印 张	14.5	版 次	2012 年 8 月第 1 版
字 数	360 000	印 次	2012 年 8 月第 1 次印刷
购书热线	010-58581118	定 价	30.00 元

物 料 号 34130-00

前　言

21 世纪是信息的世纪，是知识的世纪，也是数据爆炸的世纪。数据的快速增长源于媒介类型的极大丰富。社交网站、在线视频、数码摄像、移动通信、电子商务、遥感卫星等，每天都在源源不断地产生着大量的数据。据美国国际数据公司预测，未来十年，全球总体信息量将是现在的 44 倍。如何对这些海量数据进行有效存储、分析和利用，以帮助企业管理人员及时准确地把握市场变化的脉搏，做出正确有效的决策，从而在日趋激烈的市场竞争中立于不败之地，将是技术人员面临的巨大挑战。

数据仓库和数据分析技术就是针对上述问题而产生的一种技术解决方案。数据仓库技术是基于海量数据的决策支持系统体系化环境的核心，是面向主题的、集成的、不可更新的、随时间不断变化的数据集合，主要面向分析型应用，用于支持管理层的决策。数据分析技术是在一定的数据基础上进行分析的方式和方法，主要包括联机分析处理和数据挖掘等内容。其中，联机分析处理从不同的角度、快速灵活地对数据仓库中的数据进行复杂查询和多维分析处理，并以直观易懂的形式将查询和分析结果提供给决策人员；而数据挖掘则是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又潜在有用的信息和知识的过程。

本书详尽地介绍了数据仓库和数据分析技术的基本概念和基本原理，建立数据仓库和进行数据分析的方法和过程。全书分为数据仓库技术篇、联机分析处理技术篇、数据挖掘技术篇三部分，共 10 章。另外，附录中介绍了一些典型的数据仓库产品和工具。

数据仓库技术篇包括第一～第三章，论述了数据仓库与数据库的差别和联系、数据仓库产生的原因、数据仓库中数据的 4 个基本特征及数据仓库的体系结构；介绍了操作数据存储（ODS）的定义、特点、功能及实现机制；讨论了数据仓库中存放的数据内容及其组织形式。

联机分析处理技术篇包括第四～第六章，论述了联机分析处理技术的一些基本概念及其基本内容，包括多维数据模型、多维分析操作、多维查询语言、多维数据展示等；介绍了数据方体的存储、预计算、缩减、索引、查询和维护等相关技术。

数据挖掘技术篇包括第七～第十章，论述了数据挖掘技术的一些基本概念及其算法的组件化思想；介绍了关联规则挖掘的基本概念、经典算法及价值评估方法，序列模式和频繁子图的挖掘方法，决策树、贝叶斯、支持向量机、人工神经网络等分类方法，对象间相似性的度量方法以及基于划分的、基于密度的、基于层次的、基于模型的和基于方格的聚类方法。

附录部分介绍了 IBM 的数据仓库解决方案（包括 IBM DB2 UDB 、IBM WareHouse Manager、IBM DB2 OLAP Server、IBM Cognos、IBM Intelligent Miner）、Oracle 的数据仓库解

决方案、Microsoft SQL Server 的商务智能解决方案、Sybase 的数据仓库解决方案、Sagent 的商务智能应用平台以及 Informatica 的专业 ETL 工具。

本书可以作为高等学校计算机类、信息管理类、数据分析类等各相关专业本科生和研究生的教材和参考书，也可以作为企事业单位信息管理部门以及其他行业的开发者、管理者、设计者、信息分析人员、数据统计人员、科学研究人员的参考资料。为尽可能面向不同的应用者，在编写过程中注意做到既使全书各篇章是一个相互联系的整体，同时又使其能自成一体。读者可以选择其中的某些篇章来阅读。

作者所在的研究组对数据仓库和数据分析技术已进行了长时间的研究，早在 1996 年 7 月 15 日就在《计算机世界报》上发表了一组有关数据仓库的文章，引起了学术界、企业界很大的反响和浓厚的兴趣；1998 年 6 月出版了国内第一本数据仓库方面的著作《数据仓库技术与联机分析处理》，受到了社会各界的好评。这些年随着研究工作的深入，我们对数据仓库和数据分析技术又有了更进一步的理解，为此，在 1998 年著作的基础上，结合国内外数据仓库技术的最新发展和应用需求编写了本书。

本书由王珊主编，王珊、李翠平、李盛恩等编著。参加本书研讨和写作的还有陈红、张磊、王静等。在本书的编写过程中，博士生张应龙以及硕士生方艺璇、万杰、宋少华、耿怡娜、景婉婧、刘虹、赵婷婷、赵琳录等从不同方面分别做了一些工作，在此一并表示诚挚的谢意。

清华大学计算机系的王建勇博士审阅了全书并提出了许多有益的意见，高等教育出版社的有关人员对书稿进行了仔细的编辑加工，在此向他们致以衷心的谢意。

本书在编写过程中，虽然尽可能做到深入浅出，力求概念正确、理论联系实际，但由于数据仓库应用领域很广，发展非常迅速，加之我们水平有限，故书中一定存在许多不足之处，希望同行和广大读者提出批评和建议。

王　珊

2012 年 6 月于中国人民大学

目　录

第一篇　数据仓库技术

第一章　从数据库到数据仓库 ……2
1.1　数据仓库产生的原因 ……2
1.1.1　操作型数据处理 ……2
1.1.2　分析型数据处理 ……3
1.1.3　两种数据处理模式的差别 ……4
1.1.4　数据库系统的局限性 ……5
1.2　数据仓库的基本概念 ……6
1.2.1　主题与面向主题 ……7
1.2.2　数据仓库的其他三个特征 ……11
1.2.3　数据仓库的功能 ……13
1.3　数据仓库的体系结构 ……14
1.3.1　体系结构 ……14
1.3.2　数据集市 ……16
小结 ……20
习题 ……20
第二章　操作数据存储 ……21
2.1　什么是 ODS ……22
2.1.1　ODS 的定义及特点 ……22
2.1.2　ODS 的功能和实现机制 ……22
2.2　DB～ODS～DW 体系结构 ……26
2.2.1　ODS 与 DW ……26
2.2.2　DB～ODS～DW 三层体系结构 ……27
小结 ……29
习题 ……29
第三章　数据仓库中的数据及组织 ……30
3.1　数据仓库中的数据组织 ……30
3.2　数据仓库中数据的追加 ……32
3.3　数据仓库中的元数据 ……33
3.3.1　元数据的定义 ……33
3.3.2　元数据的分类 ……34
3.3.3　元数据管理的标准化 ……36
小结 ……39
习题 ……39

第二篇　联机分析处理技术

第四章　概述及模型 ……42
4.1　OLAP 技术概述 ……42
4.1.1　OLAP 的起源 ……42
4.1.2　OLAP 的定义 ……42
4.1.3　OLAP 与 OLTP 的区别 ……43
4.1.4　OLAP 核心技术 ……43
4.2　多维数据模型 ……44
4.2.1　基本概念 ……44
4.2.2　星形、雪片和事实群模型 ……49
4.3　多维分析操作 ……51
4.3.1　多维分析基础：聚集 ……52
4.3.2　常用多维分析操作 ……53

4.3.3 其他多维分析操作……55
4.3.4 聚集的一些限制……57
4.3.5 水平层次结构和非水平层次结构……59
4.4 多维查询语言……60
4.4.1 MDX 简介……61
4.4.2 MDX 对象模型……62
4.5 多维数据展示……63
4.5.1 三维数据展示……63
4.5.2 高维数据展示……64
小结……65
习题……65
第五章 数据方体的存储、预计算和缩减……66
5.1 数据方体的存储……66
5.1.1 MOLAP……66
5.1.2 ROLAP……70
5.1.3 MOLAP 和 ROLAP 实现机制的比较……73
5.2 数据方体的预计算……75
5.2.1 预计算的相关概念……75
5.2.2 数据方体格结构……76
5.2.3 数据方体格存储方法……77
5.3 完整数据方体的预计算方法……78
5.3.1 流水线算法……78
5.3.2 BUC 算法……80
5.4 部分数据方体的预计算方法……83
5.4.1 BPUS 算法……84
5.4.2 PBS 算法……87
5.5 数据方体的缩减技术……88
5.5.1 Drawf 数据方体……88
5.5.2 Condensed 数据方体……90
5.5.3 Quotient 数据方体……91
小结……93
习题……93
第六章 数据方体的索引、查询和维护……94
6.1 数据方体的索引技术……94
6.1.1 树索引……94
6.1.2 位图索引……99
6.2* 数据方体的查询处理和优化技术……101
6.2.1 子查询划分技术……102
6.2.2 子查询处理及优化技术……105
6.3* 数据方体的维护技术……106
小结……107
习题……107

第三篇 数据挖掘技术

第七章 数据挖掘概述……110
7.1 数据挖掘简介……110
7.1.1 数据挖掘的特点……110
7.1.2 数据挖掘与 KDD……111
7.1.3 数据挖掘与 OLAP……112
7.1.4 数据挖掘与数据仓库……113
7.1.5 数据挖掘的分类……113
7.1.6 数据挖掘的应用……114
7.2 数据挖掘算法的组件化思想……116
7.2.1 模型或模式结构……116
7.2.2 数据挖掘的任务……117
7.2.3 评分函数……118
7.2.4 搜索和优化方法……118
7.2.5 数据管理策略……119
7.2.6 组件化思想的应用……119
小结……120
习题……120
第八章 频繁模式挖掘……121
8.1 频繁项集和关联规则……121
8.1.1 问题描述……122

8.1.2 关联规则分类 ……124
8.1.3 关联规则挖掘的经典算法 Apriori ……125
8.1.4 关联规则挖掘的重要算法 FP-Growth ……133
8.1.5 其他关联规则挖掘方法……135
8.1.6 关联规则的兴趣度……137
8.2 序列模式挖掘 ……138
8.2.1 问题描述 ……138
8.2.2 GSP 算法……140
8.2.3 PrefixSpan 算法……143
8.3 频繁子图挖掘 ……145
8.3.1 问题描述 ……145
8.3.2 基于 Apriori 的宽度优先算法 ……146
8.3.3 基于 FP-Growth 的深度优先搜索算法……147
小结……148
习题……149
第九章 预测建模：分类和回归 ……150
9.1 预测建模简介 ……150
9.1.1 用于预测的模型结构……151
9.1.2 用于预测的评分函数……154
9.1.3 用于预测的搜索和优化策略……154
9.2 决策树分类……155
9.2.1 建树阶段 ……156
9.2.2 剪枝阶段 ……162
9.2.3 分类规则的生成……162
9.2.4 可扩展性问题 ……163
9.2.5 其他问题 ……163
9.3 贝叶斯分类……164
9.3.1 基本概念 ……164
9.3.2 朴素贝叶斯分类 ……166
9.4 支持向量机分类……167
9.4.1 线性可分时的二元分类问题……168
9.4.2 线性不可分时的二元分类问题……171
9.4.3 多元分类问题 ……172
9.4.4 可扩展性问题 ……173
9.5 人工神经网络分类……173
9.5.1 神经网络的组成 ……173
9.5.2 神经网络的分类方法……175
小结……178
习题……178
第十章 描述建模：聚类……180
10.1 聚类分析简介 ……180
10.1.1 对象间的相似性 ……180
10.1.2 其他相似性度量 ……183
10.2 聚类方法概述 ……184
10.2.1 基于划分的聚类方法……185
10.2.2 基于密度的聚类方法……190
10.2.3 基于层次的聚类方法……194
10.2.4 基于模型的聚类方法……201
10.2.5 基于方格的聚类方法……203
小结……204
习题……205
附录 产品与工具 ……206
附录 A IBM 数据仓库解决方案 ……206
附录 B Oracle 数据仓库解决方案 ……209
附录 C Microsoft SQL Server 2005 数据仓库解决方案 ……210
附录 D Sybase 数据仓库解决方案 ……211
附录 E Group 1 Sagent 介绍……211
附录 F Informatica 介绍 ……212
参考文献 ……214

第一篇

数据仓库技术

数据仓库（Data Warehouse，DW）是20世纪80年代中期信息领域中迅速发展起来的数据库新技术。数据仓库的建立，能充分利用已有的数据资源，把数据转换为信息，从中挖掘出知识，提炼成智慧，最终创造出效益。所以，越来越多的企业开始认识到数据仓库应用所带来的好处。

计算机系统中存在着两类不同的数据处理工作：操作型处理和分析型处理，也称作联机事务处理（On-Line Transaction Process，OLTP）和联机分析处理（On-Line Analysis Process，OLAP）。

操作型处理，是指对数据库联机的日常操作，通常是对一个或一组记录的查询和修改，例如火车售票系统、银行通存通兑系统、税务征收管理系统等。这些系统要求快速响应用户请求，对数据的安全性、完整性、事务的一致性、事务吞吐量、数据的备份和恢复等要求很高。

分析型处理，是指对数据的查询和分析操作，通常是对海量的历史数据查询和分析，例如金融风险预测预警系统、证券股市违规分析系统。这些系统要访问的数据量非常大，查询和分析的操作十分复杂。

两种数据处理工作之间的差异，使得传统的数据库技术不能同时满足两类数据的处理要求，数据仓库技术应运而生。

数据仓库是一个复杂的系统，包括数据源，后台数据抽取、转换和加载工具，数据仓库服务器，OLAP服务器以及前台数据分析工具。数据仓库不是一个单一的软件系统，而是由一组相关的软件系统组成的。

第一章　从数据库到数据仓库

本章介绍数据仓库产生的原因，数据仓库的基本概念以及体系结构。读者从中可以了解为什么有了数据库还需要数据仓库，数据仓库与数据库之间的区别与联系是什么，数据仓库的核心技术有哪些，等等。

1.1　数据仓库产生的原因

顾名思义，数据仓库就是存放数据的地方。数据库也是存放数据的地方。人们自然要问，有了数据库为什么还要数据仓库？数据仓库与数据库有什么不同？它们之间的关系是怎样的？本节从数据仓库产生的原因来阐述这些问题。

数据是企业或机构的重要资源，企业或机构的运营过程可以说是数据的收集、整理、加工、存储和检索的过程。数据处理可以大致地划分为两大类：操作型处理和分析型处理。

操作型处理主要完成数据的收集、整理、存储、查询和增、删、改操作等，主要由一般工作人员和基层管理人员完成。分析型处理是对数据的再加工，往往要访问大量的历史数据，进行复杂的统计分析，从中获取信息，因此也称为信息型处理，主要由中高级管理人员完成。

1.1.1　操作型数据处理

联机事务处理系统是操作型数据处理的典型例子，是数据库系统的主要应用。

联机事务处理系统的主要功能是对事务进行处理，快速地响应客户的服务要求使企业的业务处理自动化。其主要性能指标是事务处理效率和事务吞吐率，每个事务处理的时间越快越好，单位时间能完成的事务数量越多越好。联机事务处理系统是数据库的主要应用之一，其基本架构如图 1.1 所示。其中，数据库管理系统（DBMS）是联机事务处理系统的主要组成部分。数据库管理系统是一种通用的系统软件，用于对数据进行有效的存储、管理和存取，是应用系统赖以运行的平台。应用系统是企业根据自己的需要开发的应用软件，用于

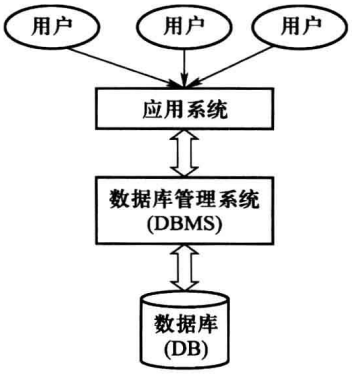


图 1.1　联机事务处理系统架构

处理日常业务。工作人员一般通过应用系统来完成日常工作。例如，银行的储蓄系统，对每一笔存款或取款交易，工作人员根据用户的要求在前台通过应用系统完成对某个账户的存款和取款操作，后台由数据库管理系统完成对数据库数据的增、删、改操作。

为了有效地对事务进行处理，数据库管理系统在技术和管理上采取了很多措施。

首先，数据库系统中严格地定义了事务的概念。所谓**事务**是用户定义的一个数据库操作序列，这些操作要么全做、要么全不做，是一个不可分割的工作单位。例如，在关系数据库中，一个事务可以是一条 SQL 语句、一组 SQL 语句或整个程序。需要注意的是，事务和程序是两个概念。一般地讲，一个程序中包含多个事务。

其次，数据库管理系统采用日志、备份等恢复技术和并发控制技术来保证事务的原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）和持续性（Durability），这 4 个特性简称为 ACID 特性（ACID properties）；采用索引技术来快速地定位数据；采用并行技术提高处理能力和系统的扩展性等。

在联机事务处理环境中，事务一般都是短事务，存取的数据量很少，需要处理的时间很短。数据库管理系统采用封锁技术提高并发度，允许多个用户同时使用数据库、使用系统资源，提高了事务的吞吐量。

在数据库应用系统的设计中，广泛采用了关系规范化理论。每个表一般都要达到第 3 范式或 BC 范式，消除了表中属性间的部分依赖和传递依赖，各个属性只依赖于主码。每个表表示一个自然的对象，例如一个实体、一个概念或者实体间的一种联系。基本消除了数据冗余，这样，在处理事务时，不会存取冗余的数据，缩短事务的处理时间。

在数据库中一般只存储最近、最新的数据。对于历史数据，一般只保存当年的数据，往年的数据被转移到其他数据库或者从数据库中卸出保存，减少数据库中的数据量，以便快速处理对当前数据库的增、删、改操作。

联机事务处理系统除了完成对业务的自动化处理外，还包含了一些简单的查询统计功能，例如，输出生产日报、月报、年报等。

联机事务处理系统基本满足了业务处理的快速响应需求，保证了数据的安全性和完整性。对一般工作人员日常的业务处理和普通管理人员一般的管理工作提供了很好的支持，因此得到了广泛的应用。

1.1.2 分析型数据处理

分析型数据处理的典型例子是决策支持系统（Decision Support System，DSS）。决策支持系统需要具备的基本功能是建立各种数学模型，对数据进行统计分析，得出有用的信息作为决策的依据和基础。

企业的中高层管理人员经常要对数据进行分析，摸清企业的运行状态和运行规律。例如，某产品的销售经理希望通过调整该产品在各零售店的分配数量来扩大其销售量。他首先要查询历史数据库中各零售店最近若干年（例如 5 年）内每天的销售记录，通过统计运算计算出近 5 年来各店的年度销售量，再通过对年度销售量的比较确定销售量增长较快的零售店。为了进一

步分析增长的原因，他还会计算出每月的销售量，判断增长的原因是否与季节有关，还要分析是否是其他产品的销售带动了目标产品的销售。他还会到其他部门获取 5 年来该产品的促销计划，确定销售量的增长与促销的关系。经过综合分析后，确定每个零售店对这种商品的分配数量。

上面的简单例子说明，分析型数据处理是不同于操作型数据处理的。分析型数据处理需要访问大量的当前和历史数据，进行复杂的计算，既需要本部门的数据也会需要其他部门的数据；甚至是竞争对手的数据。

1.1.3 两种数据处理模式的差别

通过上面的论述，可以发现，操作型数据处理与分析型数据处理是两种不同的操作，表 1.1 中列出操作型数据与分析型数据之间的主要差别。

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
当前数据	历史数据
可更新	不可更新
操作需求事先可知道	操作需求事先不知道
生命周期符合 SDLC（软件开发生命周期）	完全不同的生命周期
对性能要求高	对性能要求宽松
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
面向业务处理	面向分析挖掘
一次操作数据量小，计算简单	一次操作数据量大，计算复杂
支持日常操作	支持管理需求

操作型数据处理主要用于企业的日常事务处理工作。数据库中存放的是细节的数据。例如，零售店的数据库中存放了每个商品每次销售的情况，包括日期、时间、商品名称、单价、销售数量等，有的还包括购买者信息。数据库中存放的数据是当前的，反映的是最近一次修改后的结果，例如，当前商品的库存量。对数据的操作是数据库的增加、删除、修改和查询操作，数据库中的数据可以修改，以反映最新的结果。数据的组织以方便事务处理、提高事务处理的性能为主要目标。

分析型数据处理主要用于企业的管理工作，数据库中一般存放的是历史数据和综合数据，例如，零售店的数据库中存放了多年来积累的每种商品每天的销售情况，还存放了综合的数据，如每个月每种商品的销售总量。对数据的操作主要是查询和统计分析操作，需要涉及大量数据。数据的组织以方便分析为主要目标，所以不同部门的数据会存放在一起。为了提高查询速度还允许某种程度的数据冗余，数据分析一般需要很长的处理时间。

1.1.4 数据库系统的局限性

传统的数据库系统在操作型数据处理应用中取得了巨大的成功，那么，能否将它应用到分析型数据处理呢？答案是否定的，主要原因包括以下几点。

1．数据的分散

企业开发的联机事务处理系统一般只需要与本部门业务有关的当前数据，而对整个企业范围内的集成应用考虑很少，企业内部各事务处理的应用之间实际上几乎都是独立的，造成了当前绝大部分企业内数据的真正状况是分散而非集成的。

出现这种现象有多种原因。有的原因是设计方面的，例如，系统设计人员为了减少系统开发费用和加快开发进度，总是采用简单而“有效”的设计方案，这种“有效”仅指对解决当前面临的问题有效，而不能保证对以后新出现的问题仍然有效。有的原因是经济方面的，当经费有限时，企业总是考虑先对关键的业务活动建立应用系统，然后再逐步建立其他业务的信息处理系统。还有的原因是历史、地理方面的，例如，某个大公司由分散在各地的多个子公司组成，等等。

2．“蜘蛛网”问题

解决数据分散的一种方法是对数据进行集成。在联机事务处理系统出现不久，就开始出现一种称为“抽取”处理的程序。它从各个分散的数据库中选择符合条件的数据并把它们汇合到一个新的文件或数据库中。由于抽取程序能将数据从联机事务处理系统中转移出来，对转移出来的数据进行分析时不会影响联机事务处理系统的效率，因此，受到了程序员的喜爱，得到了大量的应用。

抽取程序解决了人们对数据的渴求，但也带来了“蜘蛛网”问题。起初只是抽取，随后是抽取之上的抽取，接着是在此基础上的再次抽取，这种不加控制的连续抽取最终导致企业的数据间形成了错综复杂的网状结构，人们形象地称为“蜘蛛网”。企业的规模越大，“蜘蛛网”问题就越严重。虽然“蜘蛛网”上任意两个节点的数据归根结底可能是从一个原始数据库中抽取出来的，但它们的数据没有统一的时间基准，抽取算法各不相同，抽取级别也不相同，并且可能参考不同的外部数据。因而对同一问题的分析，不同节点会产生不同甚至截然相反的结果，使决策者感到迷惑，无所适从。

3．数据不一致问题

由于前述的数据分散和“蜘蛛网”等问题，导致了多个应用间的数据不一致。这些数据不一致的形式是多种多样的，举例如下：

（1）同一字段在不同应用中具有不同的数据类型。例如，“性别”字段在 A 应用中的值为“M/F”，在 B 应用中的值为“0/1”，在 C 应用中又为“Male/Female”。

（2）同一字段在不同应用中具有不同的名字。例如，“余额”字段在 A 应用中的名称为“balance”，在 B 应用中名称为“bal”，在 C 应用中又变成了“currbal”。

（3）同名字段，不同含义。例如，“重量”字段在 A 应用中表示人的体重，在 B 应用中表示汽车的重量。

为了将这些不一致的数据集成起来，首先必须对它们进行转换，消除不一致之后才能供分

析使用。数据的不一致是多种多样的，对每种情况都必须专门处理，因此，这是一项很繁重的工作。

4．数据动态集成问题

由于每次分析都进行数据集成的开销太大，一些应用仅在开始时对所需的数据进行了集成，以后就一直以这部分集成的数据作为分析的基础，不再与数据源发生联系，这种方式的集成称为静态集成。静态集成最大的缺点在于，如果在数据集成后数据源中的数据发生了改变，这些变化将不能反映给决策者，导致决策者使用的是过时的数据。

对于决策者来说，虽然并不要求随时准确地探知系统内的任何数据变化，但也不希望他所分析的是几个月以前的情况。如果每做一次分析都要进行一次这样的集成，将会导致极低的处理效率。因此，集成数据必须以一定的周期（例如 24 小时）进行刷新，这种方式的集成称为动态集成。显然，联机事务处理系统不具备动态集成的能力。决策支持系统对数据集成的迫切需要可能是数据仓库技术出现的重要动因之一。

5．历史数据问题

联机事务处理一般只需要当前数据，在数据库中一般也只存储短期内的数据，且不同数据的保存期限也不一样，即使有一些历史数据保存下来了，往往也被束之高阁，没有得到充分利用。但对于决策分析而言，历史数据是相当重要的，许多分析方法必须以大量的历史数据为依托。没有对历史数据的详细分析，是难以把握企业的发展趋势的。

6．数据的综合问题

在联机事务处理系统中积累了大量的细节数据，一般而言，决策支持系统并不对这些细节数据进行分析。这主要有两个原因：一是细节数据数量太大，会严重影响分析的效率；二是太多的细节数据不利于分析人员将注意力集中于有用的信息上。因此，在分析前，往往需要对细节数据进行不同程度的综合。而联机事务处理系统不具备这种综合能力，根据规范化理论，这种综合还往往因为是一种数据冗余而被加以限制。

以上这些问题表明，在操作型数据处理的应用环境中直接构建分析型数据处理应用是一种失败的尝试。

数据仓库本质上是对存在的这些问题的回答。但是数据仓库的主要驱动力并不是改正过去的缺点，而是市场商业经营行为的改变，市场竞争要求捕获和分析事务级的业务数据。建立在事务处理环境上的分析系统无法达到这一要求。要提高分析和决策的效率和有效性，人们认为分析型处理及其数据必须与操作型处理及其数据相分离，必须把分析型数据从事务处理环境中提取出来，按照决策支持系统处理的需要进行重新组织，建立单独的分析型处理环境。数据仓库正是为了构建这种新的分析型处理环境而出现的一种数据存储和组织技术。

1.2 数据仓库的基本概念

社会的需求极大地推动了技术的发展。人们逐渐尝试对数据库中的数据进行再加工，形成

一个综合的、面向分析型的环境，以更好地支持决策分析。数据仓库的思想逐渐开始形成。但对于什么是数据仓库，许多人提出了如下不同的看法。

- “数据仓库是作为决策支持系统服务基础的分析型数据库，用来存放大容量的只读数据，为制定决策提供所需的信息”。
- “数据仓库是与操作型系统相分离的、基于标准企业模型集成的、带有时间属性的（即与企业定义的时间区段相关）、面向主题（subject-oriented）及不可更新的数据集合”。

这些观点都或多或少道出了数据仓库及其数据的特点，如为制定决策服务、面向主题、数据的不可更新等。

如前所述，传统的数据库系统主要面向以操作型处理为主的联机事务处理应用，无法满足决策时的分析型处理需求。操作型处理和分析型处理在本质上存在很大差异，这种差异也导致了它们对数据有着不同的需求。“数据仓库之父”W.H.Inmon 在其《Building the Data Warehouse》一书中，指出数据仓库中的数据应具备以下 4 个基本特征。

（1）数据仓库的数据是面向主题的；

（2）数据仓库的数据是集成的；

（3）数据仓库的数据是不可更新的；

（4）数据仓库的数据是随时间不断变化的。

并且给出了数据仓库的定义：数据仓库是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合，用以更好地支持企业或组织的决策分析处理。

下面着重来介绍数据仓库数据的 4 个基本特征。

1.2.1 主题与面向主题

与传统数据库面向 OLTP 应用进行数据组织的特点相比较，数据仓库中的数据是面向 OLAP 应用，按照主题进行组织的。什么是主题（subject）呢？主题是一个抽象的概念，是在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象。在逻辑意义上，它是对应企业中某一宏观分析领域所涉及的分析对象。

面向主题（subject-oriented）的数据组织方式，就是在较高层次上对分析对象的数据的一个完整、一致的描述，能完整、统一地刻画各个分析对象所涉及企业的各项数据，以及数据之间的联系。所谓较高层次是相对面向应用的数据组织方式而言的，是指按照主题进行数据组织的方式具有更高的数据抽象级别。

为了更好地理解主题与面向主题的概念，下面用例子来详细说明面向主题的数据组织方式与传统的面向应用的数据组织方式有什么不同。

1．传统的面向应用的数据组织方式

例如，一家采用“会员制”经营方式的商场，按业务已建立起采购、销售、库存管理以及人事管理子系统，并按照其业务处理要求，建立了各自的数据库模式。

采购子系统：

订单（订单号，供应商号，总金额，日期）

订单细则（订单号，商品号，类别，单价，数量）

供应商（供应商号，供应商名，地址，电话）

销售子系统：

顾客（顾客号，姓名，性别，年龄，文化程度，地址，电话）

销售（员工号，顾客号，商品号，数量，单价，日期）

库存管理子系统:

领料单（领料单号，领料人，商品号，数量，日期）

进料单（进料单号，订单号，进料人，收料人，日期）

库存（商品号，库房号，库存量，日期）

库房（库房号，仓库管理员，地点，库存商品描述）

人事管理子系统:

员工（员工号，姓名，性别，年龄，文化程度，部门号）

部门（部门号，部门名称，部门主管，电话）

以上述数据模式为例，先来总结一下传统的面向应用的数据组织方式的特点。

（1）面向应用进行数据组织，首先对企业中相关的组织、部门等进行详细调查，收集数据库的基础数据及其处理过程。调查的重点是“数据”和“处理”，在进行数据组织时要充分了解企业的部门组织结构，考虑企业各部门的业务活动特点。

（2）面向应用进行数据组织，应反映一个企业内数据的动态特征，即它要便于表达企业各部门内的数据流动情况以及部门间的数据输入/输出关系，通俗地讲是要表达每个部门的实际业务处理的数据流程：即从哪儿获取输入数据、在部门内进行什么样的数据处理，以及向什么地方输出数据。按照实际应用即业务处理流程来组织数据，其主要目的是为了进行联机事务处理，以提高日常业务处理的速度和准确性等，提高服务质量。

（3）这种数据组织方式生成的各项数据库模式与企业中实际的业务处理流程中所涉及的单据或文档有很好的对应关系，这种对应关系使得数据库模式具有很强的操作性，因而可以较好地在这些数据库模式上建立起各项实际的应用处理。如库存管理中的“领料单”、“进料单”和“库存”等是实际管理中存在的单据或报表，并且其各项内容也是相互对应的。在有些应用中，这种数据组织方式只是对企业业务活动所涉及数据的存储介质的改变，即从纸介质到磁介质的转变。

（4）面向应用进行数据组织的方式并没有体现数据库这一概念提出的原本意图：数据与数据处理的分离，即要将数据从数据处理或应用中抽象、解放出来，组织成一个和具体的应用相独立的数据世界。

所以，实际的数据库建设由于偏重对联机事务处理的支持，而将数据应用逻辑与数据在一