21世纪

高等院校计算机系列教材

数据仓库原理、设计与应用

陈京民　编著

中国水利水电出版社
www.waterpub.com.cn

21世纪高等院校计算机系列教材

数据仓库原理、设计与应用

陈京民　编著

中国水利水电出版社

内 容 提 要

本书全面、系统地介绍了数据仓库的原理、开发和应用技术。主要内容包含数据仓库、联机分析处理和数据挖掘的基本概念、体系结构、开发模型、项目规划、创建过程和应用管理，涵盖了数据仓库的完整生命周期。本书力求从务实的角度出发，揭开笼罩在数据仓库、联机分析处理和数据挖掘上面的神秘面纱，使读者能对数据仓库、联机分析处理和数据挖掘有一个正确认识，以推动数据仓库在我国的健康发展。

为使读者能够从各种角度对数据仓库进行全面系统的了解，并满足不同人员的需要，本书在介绍数据仓库、联机分析处理和数据挖掘的原理、设计与应用全过程的同时还介绍了一个超市数据仓库规划、设计与实施的完整过程，并在其中穿插介绍了 SQL Server 2000 中的数据仓库开发工具的具体应用，为读者对数据仓库的了解提供了实际参考框架。本书适合于企业各个层次的管理人员、项目开发人员，也可以作为相关专业本科生和研究生的教材。

本书为授课教师免费提供电子教案，此教案用 PowerPoint 制作，可以任意修改。需要者可以从中国水利水电出版社网站 www.waterpub.com.cn 下载，也可与北京万水电子信息有限公司联系，联系电话：（010）82564395。

图书在版编目（CIP）数据

数据仓库原理、设计与应用 / 陈京民编著. —北京：中国水利水电出版社，2004.3

（21 世纪高等院校计算机系列教材）

ISBN 7-5084-2042-X

Ⅰ. 数… Ⅱ. 陈… Ⅲ. 数据库系统－高等学校－教材 Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字（2004）第 019951 号

书　　名	数据仓库原理、设计与应用
作　　者	陈京民　编著
出版　发行	中国水利水电出版社（北京市三里河路 6 号　100044） 网址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电话：（010）63202266（总机）　68331835（营销中心）　82562819（万水）
经　　售	全国各地新华书店和相关出版物销售网点
排　　版	北京万水电子信息有限公司
印　　刷	北京北医印刷厂
规　　格	787mm×1092mm　16 开本　19.25 印张　440 千字
版　　次	2004 年 4 月第 1 版　2004 年 4 月第 1 次印刷
印　　数	0001—5000 册
定　　价	26.00 元

前　言

信息技术的迅速发展和企业管理决策支持的迫切需要，在短短的几年内将数据仓库（Data Warehouse）从纯粹的理论研究迅速转化为决策支持领域中一种实用性极强的技术。数据仓库的发展将我们从简单的批处理、联机事务处理的信息处理时代带入了联机分析处理、数据仓库和数据挖掘的信息分析时代。这一发展过程具有内在的动力和外在的推力。企业在早期的信息化进程中所构建的联机事务处理系统为企业业务快速、准确地处理提供了基本条件，同时为企业积累了大量有价值的业务信息。但是这些处理只能支持企业的日常业务工作，而对企业的经营管理决策却很少能够提供支持。许多企业的经营管理人员在日趋严重的市场竞争压力下，开始着手建立数据存储——数据集市用于经营管理决策，以应对日益严酷的市场竞争。这些因素最终促进了数据仓库的发展与应用。数据仓库所包含的数据仓库技术、联机分析处理和数据挖掘技术不仅体现了当今世界上最先进的 IT 技术，而且还提供了能够对企业管理决策提供实际支持的系统。

数据仓库的建立不仅需要有各种建设工具，而且还需要有相应的数据支持，数据仓库的建设必须基于比较完善的信息化构架，只有在一定的信息化基础上，才能进行数据仓库的建设。数据仓库的建设是企业经营管理决策与信息化的结合过程，只有依照企业管理决策的实际需要，才能建设一个支持企业管理决策的数据仓库。数据仓库的建设是各种先进的信息处理技术与企业管理决策结合的过程。只有将 OLAP 技术、数据挖掘技术与数据仓库中庞大的数据相结合，与企业先进的管理决策方法相结合，才能使数据仓库在企业的经营管理决策中发挥巨大的作用。数据仓库的建设成功不仅取决于技术人员对数据仓库开发方法与开发工具的熟练应用，更取决于数据仓库能否得到熟练应用。可以毫不夸张地说，数据仓库的成功关键在于用户的应用情况，而不是数据仓库开发技术的熟练应用。因此，本书在介绍了数据仓库的开发模型和开发方法后，还用相当的篇幅介绍了数据仓库的管理与应用。其中包含了大量的数据仓库应用情况与应用案例，使读者可以了解如何利用数据仓库来降低企业的运营成本，建立更好的客户关系管理，提高产品的质量。

为使读者能够清楚地了解数据仓库的开发，本书介绍了数据仓库开发应用的生命周期。数据仓库的整个开发过程从数据仓库规划分析到设计实施，终结于应用管理，使读者可以了解到数据仓库开发应用的完整周期，以及如何处理在不同阶段中所遇到的问题。为使读者能够通过实际的数据仓库开发应用，以加深对数据仓库与数据挖掘的了解，本书还介绍了超市数据仓库规划、设计和实施实例，并在其中穿插介绍了 SQL Server 2000 在数据仓库开发应用中的实际应用，目的在于使读者能够更深入地了解数据仓库、联机分析处理与数据挖掘技术。

全书共分 9 章。第 1 章主要介绍数据仓库和数据挖掘技术的产生背景、发展、总体结构和使用技术；第 2 章从理论上介绍了数据仓库的开发模型——概念模型、逻辑模型、物

理模型、元数据模型和数据粒度及聚集模型；第 3 章叙述了数据仓库开发应用的完整周期，涉及到数据仓库的开发规划、需求分析、设计、实施、使用及支持等；第 4 章阐述了联机分析技术（OLAP）的基本概念、结构、实施以及 OLAP 工具评价标准；第 5 章详细介绍了传统的数据挖掘技术——统计分析类数据挖掘技术、工具、应用及应用中的问题；第 6 章介绍了现代数据挖掘技术与发展，其中包含了规则类、神经网络类、遗传算法类和粗糙集类型等现代挖掘技术，同时还介绍了知识发现工具与应用，以及文本挖掘、Web 挖掘、可视化数据挖掘、空间数据挖掘和分布式数据挖掘等数据挖掘技术的未来发展；第 7 章从数据仓库的用户、应用案例、运行技术管理、元数据管理、应用中的法律问题以及成本与效益分析等角度说明了数据仓库的应用和管理中的问题；第 8 章和第 9 章分别介绍了数据仓库的开发和应用实例，以及 SQL Server 2000 在数据仓库开发中的具体应用。

参加本书研讨并提供实例资料的还有朱惠云、杜冬军、俞强等。另外，孙春亮为本书的顺利出版做了大量的筹备和组织工作。在此对他们的辛勤工作表示深深的谢意！

由于数据仓库技术正处于日新月异的发展阶段，加之编者水平有限，书中谬误或疏漏之处在所难免，恳请广大读者不吝指教，欢迎联系：

E_mail：cjm20020101@sina.com。

作　者

2004 年 2 月

目　　录

前言
第 1 章　数据仓库与数据挖掘概述 1
1.1　数据仓库的发展与展望 1
1.1.1　从传统数据库到数据仓库 1
1.1.2　数据仓库的定义与基本特性 3
1.1.3　数据仓库的未来发展 7
1.2　数据仓库的体系结构 8
1.2.1　数据仓库的概念结构 8
1.2.2　虚拟数据仓库结构 8
1.2.3　数据集市结构 9
1.2.4　单一数据仓库结构 9
1.2.5　分布式数据仓库结构 10
1.3　数据仓库的参照结构 11
1.3.1　数据仓库基本功能层 11
1.3.2　数据仓库的管理层 18
1.3.3　数据仓库的元数据管理层 19
1.3.4　数据仓库的环境支持层 20
1.4　数据挖掘技术概述 21
1.4.1　数据挖掘的发展 21
1.4.2　数据挖掘的定义 22
1.5　数据挖掘技术与工具 24
1.5.1　常用的数据挖掘技术 24
1.5.2　常用数据挖掘工具 26
1.5.3　数据挖掘工具的评价标准 28
1.5.4　常用数据挖掘工具的选择 29
1.6　数据挖掘的应用 30
1.6.1　数据挖掘与数据仓库 30
1.6.2　数据挖掘过程 31
1.6.3　数据挖掘的用户 35
第 2 章　数据仓库开发模型 36
2.1　数据仓库开发模型概述 36
2.2　数据仓库概念模型 37

2.2.1 概念数据模型 37
2.2.2 规范的数据模型 40
2.2.3 星型模型 41
2.2.4 雪花模型 43
2.3 数据仓库的逻辑模型 43
2.3.1 事实表模型设计 45
2.3.2 维模型设计 47
2.4 数据仓库的物理模型 47
2.4.1 数据仓库物理模型的存储结构 47
2.4.2 数据仓库物理模型的索引构建 48
2.4.3 数据仓库物理模型的优化问题 52
2.5 数据仓库的元数据模型 53
2.5.1 元数据的类型与组成 53
2.5.2 元数据在数据仓库中的作用 55
2.5.3 元数据的收集 58
2.6 数据仓库的粒度和聚集模型 59
2.6.1 数据粒度的划分 60
2.6.2 确定粒度的级别 61
2.6.3 数据仓库的聚集模型确定 62
2.6.4 聚集模型的处理 62
2.6.5 聚集模型的管理 63
第 3 章 数据仓库开发应用过程 64
3.1 数据仓库开发应用的特点 64
3.1.1 数据仓库开发应用的阶段性 64
3.1.2 数据仓库的螺旋式开发方法 65
3.1.3 数据仓库的开发特点 66
3.2 数据仓库的规划 67
3.2.1 选择数据仓库的实现策略 67
3.2.2 确定数据仓库的开发目标和实现范围 68
3.2.3 数据仓库的结构 70
3.2.4 数据仓库使用方案和项目规划预算 71
3.3 数据仓库的概念模型设计 72
3.3.1 概念模型的需求调查 72
3.3.2 概念模型的定义 73
3.3.3 概念模型的分析 76
3.3.4 概念模型的设计 77
3.3.5 概念模型文档与评审 79

3.4 数据仓库的逻辑模型设计 80
3.4.1 分析主题域 81
3.4.2 粒度层次和聚集的确定 82
3.4.3 确定数据分割策略 82
3.4.4 关系模型定义 83
3.4.5 数据仓库的实体定义 83
3.4.6 数据仓库的数据抽取模型 84
3.4.7 数据仓库元数据模型的建立与应用 89
3.4.8 逻辑模型的评审 90
3.5 数据仓库的物理模型设计 91
3.5.1 数据仓库设计的规范 91
3.5.2 确定数据结构的类型 92
3.5.3 数据仓库索引的创建 93
3.5.4 确定数据的存放位置 94
3.5.5 确定存储分配 94
3.5.6 数据仓库物理模型的评审 95
3.6 数据仓库的实施 96
3.6.1 数据仓库与业务处理系统接口的设计 97
3.6.2 数据仓库的创建 97
3.6.3 数据仓库的数据加载、复制与发行 98
3.6.4 数据仓库的中间件设计 99
3.6.5 数据仓库的测试 99
3.7 数据仓库的应用、支持和增强 100
3.7.1 数据仓库的用户培训及支持 100
3.7.2 数据仓库的使用方式 101
3.7.3 数据仓库使用中的数据刷新 102
3.7.4 数据仓库的增强 103
第 4 章 OLAP 技术 105
4.1 OLAP 技术概述 105
4.1.1 OLAP 的发展 105
4.1.2 OLAP 的特性 105
4.2 OLAP 与多维分析 106
4.2.1 多维基本概念 106
4.2.2 多维分析 109
4.2.3 维的层次关系 111
4.2.4 维的类关系 111
4.2.5 OLAP 与数据仓库的关系 112

4.3 OLAP 的实施 113
4.4 多维 OLAP 与关系 OLAP 114
4.4.1 多维数据库 114
4.4.2 多维数据库的数据存储 116
4.4.3 多维数据库与数据仓库 116
4.4.4 MOLAP 的创建与功能 117
4.4.5 ROLAP 实现的三个规则 118
4.4.6 ROLAP 的多维表示方法 119
4.4.7 ROLAP 的创建与功能 121
4.5 OLAP 技术评价 122
4.5.1 MOLAP 与 ROLAP 的比较 122
4.5.2 OLAP 的衡量标准 124
4.5.3 OLAP 服务器和工具的评价标准 126
第 5 章 传统数据挖掘技术 128
5.1 传统的统计分析类数据挖掘技术 128
5.1.1 统计与统计类数据挖掘技术 128
5.1.2 数据的聚集与度量技术 129
5.1.3 柱状图数据挖掘技术 129
5.1.4 线性回归数据挖掘技术 131
5.1.5 非线性回归数据挖掘技术 133
5.1.6 聚类数据挖掘技术 133
5.1.7 最近邻数据挖掘技术 140
5.2 统计分析类工具 141
5.2.1 统计类数据挖掘工具 141
5.2.2 统计类数据挖掘的商业分析 142
5.2.3 统计类数据挖掘工具的功能 142
5.2.4 统计类数据挖掘工具——SPSS 143
5.3 统计分析类工具的应用 146
5.3.1 趋势分析 146
5.3.2 时序分析 147
5.3.3 周期分析 147
5.4 统计分析类工具应用的问题 148
5.4.1 统计类数据挖掘的预处理问题 148
5.4.2 统计分析应遵循的基本原则 150
5.4.3 统计分析的步骤 151
5.4.4 统计类数据挖掘的性能问题 151
第 6 章 现代数据挖掘技术与发展 153

6.1 知识挖掘系统的体系结构 153
6.1.1 知识发现的定义 153
6.1.2 知识发现系统的结构 154
6.2 现代挖掘技术及应用 156
6.2.1 规则型现代挖掘技术及应用 156
6.2.2 神经网络型现代挖掘技术 161
6.2.3 遗传算法型现代挖掘技术 166
6.2.4 粗糙集型现代挖掘技术 170
6.2.5 决策树型现代挖掘技术 172
6.3 知识发现的工具与应用 175
6.3.1 知识挖掘工具的系统结构 175
6.3.2 知识挖掘工具运用中的问题 177
6.3.3 知识挖掘的价值 179
6.3.4 现代数据挖掘工具简介 180
6.4 数据挖掘技术的发展 181
6.4.1 文本挖掘 181
6.4.2 Web 挖掘技术 183
6.4.3 可视化数据挖掘技术 186
6.4.4 空间数据挖掘 187
6.4.5 分布式数据挖掘 190
第 7 章 数据仓库的应用与管理 193
7.1 数据仓库的用户 193
7.1.1 数据仓库的用户——信息的使用者与知识的挖掘者 193
7.1.2 信息使用者的数据仓库使用方式 193
7.1.3 知识挖掘者的数据仓库使用方式 194
7.2 数据仓库应用案例 195
7.2.1 分层决策体系 195
7.2.2 数据抽样分析 197
7.2.3 发挥历史数据的经济效益 198
7.2.4 回扣分析 199
7.2.5 客户关系管理 199
7.3 数据仓库的运行技术管理 200
7.3.1 数据加载的一些问题 200
7.3.2 故障恢复管理 201
7.3.3 访问控制与安全管理 201
7.3.4 数据增长的管理 202
7.4 数据仓库的元数据管理 203

7.4.1 元数据的存储、管理与维护 203
7.4.2 元数据的用户与使用方法 204
7.4.3 元数据管理模型 206
7.5 数据仓库应用中的法律问题 208
7.5.1 数据的隐私权问题 209
7.5.2 数据隐私权的处理 209
7.6 数据仓库的成本与效益分析 211
7.6.1 数据仓库的投资回报的定量分析 211
7.6.2 数据仓库的投资回报的定性分析 212
第 8 章 数据仓库开发实例 214
8.1 超市销售数据仓库的规划与分析 214
8.1.1 超市销售数据仓库的需求分析 214
8.1.2 超市销售数据仓库 E-R 模型的构造 215
8.1.3 超市数据仓库事实表模型 216
8.1.4 超市数据仓库维表模型设计 218
8.1.5 超市数据仓库模型的关键字设计 223
8.1.6 超市数据仓库的元数据设计 225
8.2 数据仓库开发工具简介 228
8.2.1 数据仓库开发工具 228
8.2.2 SQL Server 数据仓库开发应用工具 230
8.3 SQL Server 的数据仓库创建 232
8.3.1 创建数据库 233
8.3.2 创建表 234
8.4 SQL Server 数据仓库事实表与多维数据集的建立 235
8.4.1 Analysis Manager 数据库的创建与数据源确定 235
8.4.2 SQL Server 数据仓库的维创建 239
8.4.3 SQL Server 的多维数据集创建 246
第 9 章 数据仓库应用实例 253
9.1 数据仓库的数据加载与钻取 253
9.1.1 数据仓库的数据加载 253
9.1.2 超市数据仓库系统的数据加载 256
9.1.3 多维数据集的更新 262
9.1.4 数据仓库的钻取访问 267
9.1.5 数据仓库的多维表达式 MDX 应用 270
9.2 数据挖掘模型的设计 272
9.2.1 数据挖掘对象的分析 272
9.2.2 数据挖掘模型与相关数据的准备 273

9.2.3 数据挖掘模型的应用 276
9.3 SQL Server 中的数据挖掘工具 276
9.3.1 决策类数据挖掘工具的应用 277
9.3.2 聚类分析数据挖掘工具的应用 282
9.4 数据仓库客户端界面的设计 287
9.4.1 客户端界面展现内容的设计 287
9.4.2 客户端界面展现工具的选择 288
9.4.3 Excel 展现界面的实现 289
参考文献 294

第 1 章　数据仓库与数据挖掘概述

随着信息技术的不断推广和应用，许多企业都已经在使用管理信息系统处理管理事务和日常业务。这些管理信息系统为企业积累了大量的信息，此时，企业管理者开始考虑如何利用这些信息海洋对企业的管理决策提供支持。因此，在信息处理中，产生了与传统数据库有很大差异的数据环境要求和从这些海洋数据中获取特殊知识的工具的需要。

1.1　数据仓库的发展与展望

传统数据库在日常的管理事务处理中获得了巨大的成功，但是对管理人员的决策分析要求却无法满足。因为，管理人员常常希望能够通过对组织中的大量数据进行分析，了解业务的发展趋势。而传统数据库只保留了当前的业务处理信息，缺乏决策分析所需要的大量历史信息。为满足管理人员的决策分析需要，就需要在数据库的基础上产生适应决策分析的数据环境——数据仓库（DW，Data Warehouse）。

1.1.1　从传统数据库到数据仓库

随着市场竞争的加剧，信息系统的用户已经不满足于仅仅用计算机去处理每天所发生的事务数据，而是需要信息——能够支持决策的信息，去帮助管理决策。这就需要一种能够将日常业务处理中所收集到的各种数据转变为具有商业价值的信息的技术，但是传统数据库系统无法承担这一责任。因为传统数据库的处理方式和决策分析中的数据需求不相称，导致传统数据库无法支持决策分析活动。这些不相称性主要表现在决策处理中的系统响应问题、决策数据需求的问题和决策数据操作的问题。

1. 决策处理的系统响应问题

在传统的事务处理系统中，用户对系统和数据库的要求是数据存取频率要高，操作时间要快。用户的业务处理操作请求往往在很短的时间内就能完成，这就使系统在多用户的情况下，也可以保持较高的系统响应时间。

但在决策分析处理中，用户对系统和数据的要求则发生了很大的变化。有的决策问题处理请求，可能会导致系统长达数小时的运行；有的决策分析问题的解决，则需要遍历数据库中的大部分数据。这些操作必然要消耗大量的系统资源，这是对业务处理实时响应的事务联机处理系统所无法忍受的。

2. 决策数据需求的问题

在进行决策分析时，需要有全面、正确的集成数据，这些集成数据不仅包含企业内部各部门的有关数据，而且还包含企业外部的，甚至竞争对手的相关数据。但是在传统数据库中，只存储了本部门的事务处理数据，而没有与决策问题有关的集成数据，更没有企业

外部的数据。如果将数据的集成交给决策分析程序处理，将大大增加决策分析系统的负担，使原来执行时间冗长的系统运行时间进一步加长，用户更加难以接受。而且，每次用户进行一次决策分析，都需要进行一次数据的集成，将大大降低系统的运行效率。如果数据库能够完成数据的集成，就可以大大提高决策系统的运行效率。

在决策数据的集成中还需要解决数据混乱问题。企业数据混乱的原因多种多样，有的是企业经营活动造成的，例如，企业进行兼并活动后，被兼并企业的信息系统与兼并企业的系统不兼容，数据无法共享。有的是系统开发的历史原因所造成的，例如，在系统开发中，由于资金的缺乏，只考虑了一些关键系统的开发，而对其他系统未予考虑，使决策数据无法集成。面对这些混乱的数据，还可能在决策分析应用中发生数据的不一致性。同一实体的属性在不同的应用系统中，可能有不同的数据类型、不同的字段名称。例如，职工的性别在人事系统中可能用逻辑值“M”和“F”表示，在财务系统中可能用数字“0”和“1”表示。或者同名的字段在不同的应用中有不同的含义，表示了不同实体的不同属性。例如，名称为“GH”的字段名称在人事系统中表示为职工的“工号”，但是在销售管理系统中却表示为“购货号”。这样在使用这些数据进行决策之前，必须对这些数据进行分析，确认其真实含义。

在决策分析中，系统常常需要从数据库中抽取数据、查找有用的数据，然后将这些数据导入其他文件或数据库中，供用户使用。这些被抽取出来的数据，有可能被其他用户再次抽取。由于这种不加限制的数据连续抽取，使企业的数据空间构成了一个错综复杂的数据“蜘蛛网”，即形成了自然演化体系结构。在这个数据“蜘蛛”网中，有可能两个节点上的数据来自于同一个原始数据库。但是由于数据的抽取时间、抽取方法、抽取级别等方面的差异，可能使这两个节点的数据不一致。这样，在对同一问题的决策分析中，由于数据的出发基准不同，可能导致截然相反的结果。也就是说，由于决策分析过程中所形成的自然演化体系，造成了数据可信度的降低，必然导致数据转化为信息的不可行和不可信，使企业无法将大量宝贵的信息资源转化为企业的核心竞争力。

数据的集成还涉及到外部数据与非结构化数据的应用问题。决策分析中经常要用到系统外的数据，例如行业的统计报告、咨询公司的市场调查分析数据。这些数据必须经过格式、类型的转换，才能被决策系统应用。许多系统在对数据进行一次集成以后，就与原来的数据源断绝了联系。这样在决策分析中，所分析的数据可能是几个月前甚至是一年以前的，其结果必然导致决策的失误。因此在决策分析系统中要求数据能够进行定期的、及时的更新，数据的更新期可能是一天，也可能是一周，而传统数据库系统缺乏数据动态更新的能力。

为完成事务处理的需要，传统数据库中的数据一般只保留当前的数据。但是对于决策分析而言，历史的、长期的数据却具有重要的意义。利用历史数据可以对未来的发展进行正确的预测，但是传统数据库却无法长期保留大量的历史数据。

在决策分析过程中，决策人员往往需要的并不是非常详细的数据，而是一些经过汇总、概括的数据。但在传统数据库中为支持日常的事务处理需要，只保留一些非常详细的数据，这对决策分析十分不利。

3. 决策数据操作的问题

在对数据的操作方式上，事务处理系统远远不能满足决策人员的需要。事务处理系统的结构基本上是一种典型的固定结构体系，操作人员只能使用系统所提供的有限参数进行数据操作，用户对数据的访问受到很大的限制。而决策分析人员则往往希望以专业用户的身份而不是参数用户的身份对数据进行操作，他们往往希望能够用各种工具对数据进行多种形式的操作，希望数据操作的结果能以商业智能的方式表达出来。而传统的业务处理系统只能以标准的固定报表方式为用户提供信息，使用户很难理解信息的内涵，无法用于管理决策。

由于系统响应、决策数据需求和决策数据操作等问题的影响，使企业无法使用现有的事务处理系统去解决决策分析的需要。因此，决策分析需要一个能够不受传统事务处理的约束、能够高效率处理决策分析数据的环境，数据仓库正是可以满足这一要求的数据存储和数据组织技术。

4. 数据仓库与传统数据库的对比

数据仓库虽然是从数据库发展而来的，但是两者在许多方面都存在着相当大的差异，如表 1-1 所示。从数据存储内容看，数据库只存放当前值，而数据仓库则存放历史值；数据库中数据的目标是面向业务操作人员，提供事务处理的支持，而数据仓库则是面向中高层管理人员，提供决策支持。数据库内的数据是动态变化的，只要有业务发生，数据就会被更新，而数据仓库则是静态的历史数据，只能定期添加。数据库中的数据结构比较复杂，用各种数据结构来满足业务处理系统的需要，而数据仓库中的数据结构则较为简单。数据库中数据的访问频率高，但是访问数据的量少；而数据仓库的访问频率低，但是数据访问量要远高于数据库。数据库在访问数据时要求响应速度很快，其响应时间一般要求在数秒以内，而数据仓库的响应时间则可能长达数小时。

表 1-1　数据仓库与数据库对比表

对比内容	数据库	数据仓库
数据内容	当前值	历史的、存档的、归纳的、计算的数据
数据目标	面向业务操作程序、重复处理	面向主题域、管理决策分析应用
数据特性	动态变化、按字段更新	静态、不能直接更新、只定时添加
数据结构	高度结构化、复杂、适合操作计算	简单、适合分析
使用频率	高	中到低
数据访问量	每个事务只访问少量记录	有的事务可能要访问大量记录
对响应时间的要求	以秒为单位计量	以秒、分钟，甚至小时为计量单位

1.1.2　数据仓库的定义与基本特性

在数据仓库的发展过程中，许多人对此做出了贡献。其中，Devlin 和 Murphy 在 1988 年发表了关于数据仓库论述的最早文章。而 William H.Inmon 在 1993 年所写的论著

《Building the Data Warehouse》则首先系统地阐述了关于数据仓库的思想、理论，为数据仓库的发展奠定了历史基石。在文中，他将数据仓库定义为：

“一个面向主题的、集成的、随时间变化的、非易失性数据的集合，用于支持管理层的决策过程”。

从 W. H. Inmon 关于数据仓库的定义中可以发现，数据仓库具有这样一些重要的特性：面向主题性、数据集成性、数据的时变性、数据的非易失性、数据的集合性和支持决策作用。

1. *面向主题性*

面向主题性表示了数据仓库中数据组织的基本原则，数据仓库中的所有数据都是围绕着某一主题组织展开的。由于数据仓库的用户大多是企业的管理决策者，这些人所面对的往往是一些比较抽象的、层次较高的管理分析对象。例如，企业中的客户、产品、供应商等都可以作为主题看待。从信息管理的角度看，主题就是在一个较高的管理层次上对信息系统中的数据按照某一具体的管理对象进行综合、归类所形成的分析对象。而从数据组织的角度看，主题就是一些数据集合，这些数据集合对分析对象做了比较完整的、一致的描述，这种描述不仅涉及到数据自身，而且还涉及到数据之间的联系。

数据仓库的创建、使用都是围绕着主题实现的。因此，我们必须了解如何按照决策分析来抽取主题；所抽取出的主题应该包含哪些数据内容；这些数据内容应该如何组织。在进行主题抽取时，必须按照决策分析对象进行。例如，在企业销售管理中的管理人员所关心的是：本企业哪些产品销售量大、利润高；哪些客户采购的产品数量多；竞争对手的哪些产品对本企业产品构成威胁。根据这些管理决策的分析对象，就可以抽取出“产品”、“客户”等主题。

确定主题以后，需要确定主题应该包含的数据。此时，应该注意不能将围绕主题的数据与业务处理系统中的数据相混淆。例如“产品”主题在销售业务处理系统中已有数据存在，但是这些数据未必都能用于数据仓库。因为在业务处理系统中，数据组织的目的在于如何能够更加有效地处理产品的销售业务。因此，可能采用“产品定单”、“产品销售细则”、“产品库存”、“客户”等数据来描述产品的销售活动。但是在对产品销售所进行的决策分析中，分析哪些客户订购产品量大时，只有客户才是所需要分析的对象。而“产品定单”、“产品销售细则”、“产品库存”等数据只是业务处理系统中的业务操作数据。但是仅仅使用业务处理系统中的“客户”数据，又不能完成对“客户”的分析，因为还需要了解客户的产品采购量、最后一次采购时间、购买竞争对手的产品等数据。这就需要围绕“客户”这一主题重新进行数据的组织。在围绕“客户”主题进行数据组织时，不适合决策分析要求的数据可能需要抛弃。例如，“产品库存”对客户的产品采购量没有直接的影响，就不需要在数据仓库中出现。有的则要将关于某一主题的、散落在其他业务处理系统中的信息组织进来。例如，客户的“信用”信息存在于财务处理系统中，在进行客户的产品采购分析时，需要了解这一信息，就要将其组织进来。有的信息则可能存在于企业的外部系统中，在决策分析中需要使用，也要将其组织到所分析的主题中。例如，客户购买竞争对手产品的信息是从企业的销售代理商或市场调查公司那里所获取的，不是企业内部的数据，但是也需要组织到“客户”主题中。

在主题的数据组织中应该注意，不同的主题之间可能会出现相互重叠的信息。例如，“客户”主题与“产品”主题在产品购买信息方面有相互重叠的信息。这种重叠信息往往来源于两个主题之间的联系，例如，“客户”主题与“产品”主题在产品购买信息方面的相互重叠是源于与客户和产品都有关的销售业务处理系统。这种主题间的重叠是逻辑上的重叠，而不是同一数据内容的物理存储重复。

主题在数据仓库中可以用多维数据库方式进行存储。如果主题的存储量大，用多维数据库存储时，处理效率将降低。为提高处理效率，可以采用关系数据库方式进行存储。应该注意，主题只是逻辑上的一个概念，一个主题在数据仓库中存储时可能需要几个表来实现。此时，这些表之间的相互联系需要通过表的主键来实现，这些主键就构成了主题的公共主键。实际存储的主题数据是需要经过综合处理的，而不再是业务处理系统中的详细数据。

在主题的划分中，必须保证每一个主题的独立性。也就是说每一个主题要有独立的内涵和明确的界线。在划分主题时，应该保证在对主题进行分析时所需要的数据都可以在此主题内找到。如果对主题进行分析时，涉及到主题外的其他数据，就需要考虑将这些数据组织到主题中，以保证主题的完备性。

由于主题是在较高层次上的数据抽象，这就使面向主题的数据组织可以独立于数据的处理逻辑，可以很方便地在这种数据环境上进行管理决策的分析处理。

2. 数据集成性

数据仓库的集成性是指根据决策分析的要求，将分散于各处的源数据进行抽取、筛选、清理、综合等工作，使数据仓库中的数据具有集成性。

数据仓库所需要的数据不像业务处理系统那样直接从业务发生地获取，而是从业务处理系统里获取。这里所指的业务处理系统可以包含这样一些系统：传统的以客户机/服务器为基本框架的在线事务处理系统（OLTP）、从早期事务处理系统发展起来的企业资源计划（ERP）和企业业务流程重组（BPR）以及基于因特网的电子商务（EC）。这些业务处理系统中的数据往往与业务处理联系在一起，只为业务的日常处理服务，而不是为管理决策分析服务。这样，数据仓库在从业务处理系统那里获取数据时，并不能将源数据库中的数据直接加载到数据仓库中，而是需要进行一系列的数据预处理，即数据的抽取、筛选、清理、综合等集成工作。也就是说，首先要从源数据库中挑选出数据仓库所需要的数据，然后将这些来自不同数据库中的数据按照某一标准进行统一，即将不同数据源中的数据的单位、字长与内容按照数据仓库的要求统一起来，消除源数据中字段的同名异义、异名同义现象，这些工作通称为数据的清理。在将源数据加载进数据仓库后，即源数据装入数据仓库后，还需要将数据仓库中的数据进行某种程度的综合，即根据决策分析的需要对这些数据进行概括、聚集处理。

3. 数据的时变性

数据仓库的时变性，就是数据应该随着时间的推移而发生变化。尽管数据仓库中的数据并不像业务数据库那样要反映业务处理的实时状况，但是数据也不能长期不变，如果依据 10 年前的数据进行决策分析，那决策所带来的后果将是十分可怕的。因此，数据仓库必须能够不断捕捉主题的变化数据，将那些变化的数据追加到数据仓库中去，也就是说在数