信息系统工程丛书

# 数据仓库原理与应用

张维明　主编

邓　苏　刘青宝　陈卫东　等编著

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

信息系统工程丛书

# 数据仓库原理与应用

张维明　主编
邓　苏　刘青宝　陈卫东　等编著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

## 内 容 简 介

本书主要介绍数据仓库的概念、基本原理、规划、开发方法以及相关算法。全书共分 8 章，包括数据仓库的发展、技术体系、元数据管理、分析设计方法和开发工具，并对数据挖掘的理论和方法，联机分析等应用技术作了深入的阐述，是一本理论与实践相结合的教材。

本书适合作为本科生高年级教材和研究生教材，也适合于从事信息系统开发的工程技术人员使用。



**图书在版编目(CIP)数据**

数据仓库原理与应用/张维明主编 .—北京：电子工业出版社，2002.3
(信息系统工程丛书)
ISBN 7-5053-6894-X

Ⅰ.数… Ⅱ.张… Ⅲ.数据库系统 Ⅳ.TP311.13

中国版本图书馆 CIP 数据核字(2002)第 015079 号

丛 书 名：信息系统工程丛书
书　　名：**数据仓库原理与应用**
主　　编：张维明
编 著 者：邓　苏　刘青宝　陈卫东　等
策划编辑：秦　梅
责任编辑：贾　贺　吴红梅
排版制作：电子工业出版社计算机排版室监制
印 刷 者：北京牛山世兴印刷厂
装 订 者：三河市路通装订厂
出版发行：电子工业出版社　http://www.phei.com.cn
　　北京市海淀区万寿路 173 信箱　邮编 100036
经　　销：各地新华书店
开　　本：787×1092　1/16　印张：16.25　字数：410 千字
版　　次：2002 年 3 月第 1 版　2002 年 3 月第 1 次印刷
书　　号：ISBN 7-5053-6894-X / TP·3920
印　　数：5 000 册　　定价：30.00 元

# 《信息系统工程丛书》编委会



---

# 《数据仓库原理与应用》编写人员名单

**主编** 张维明

**编著** 邓 苏 刘青宝 陈卫东 李晓林 黄宏斌 戴超凡

# 丛 书 序

从现实世界的角度看，客观世界是由物质、能量和信息三大基本要素组成的，人类的社会生活每时每刻都离不开信息。从远古时代开始，人类就一直在同信息打交道，围绕着信息形成了不同的信息作业，包括信息的采集、存储、表示、传递、加工处理、检索利用和控制等，所有这些环节形成了信息系统，并作为客观世界每个系统的一个子系统或显式或隐式地存在着。

从科学技术的角度看，信息系统是在20世纪中叶由信息科学、计算机科学、管理科学、决策科学、系统科学等学科相互渗透交叉而发展起来的，经过多年的研究目前已经形成了比较完整的独具特色的体系。信息系统工程是80年代出现的以建立信息系统为目标的新兴学科，它是用系统工程的原理、方法来指导信息系统建设与管理的一门工程技术学科，主要研究各级各类信息系统建设和管理中的规律性的问题。它既不是“信息的系统工程”，也不是“信息系统的工程”，而是“信息系统的系统工程”。一般认为，信息系统工程的目标是为以计算机和其它信息技术为手段的各类信息系统提供科学的开发方法、管理手段及有关的工具、标准、规范，通常不包括通信工程、信号处理等具体学科领域的技术。

信息系统工程的研究范围主要包括：

(1)信息系统的基本理论。信息系统的基本观点、认识论和方法论等。

(2)信息系统建模。信息系统概念模型、逻辑模型和物理模型的描述、观察、试验与验证等。

(3)信息系统开发。信息系统建设与管理的概念、方法、评价、规划、工具、标准等一系列相关技术问题和工程问题。

(4)信息系统支撑技术在信息系统中的应用。数据库/数据仓库、网络通信、人机交互、分布计算、决策支持、人工智能等技术如何满足信息系统各层次用户的需求，实现业务管理、信息共享、分析决策等功能，并在组织和人的参与下最终达到信息系统的目标。

(5)信息系统集成。研究系统集成的原则、方法、技术、工具和有关的标准、规范，应用先进的相关技术，将支持各个信息“孤岛”的小运行环境，集成统一在一个大运行环境中，最终形成一体化的信息系统。

《信息系统工程丛书》是由国防科技大学管理学院组织多位专家和科研人员面向信息系统工程专业撰写的教材类图书。作者所在的单位是70年代末在钱学森院士的亲自倡导下建立起来的，在国内最早开设了信息系统工程专业。作者长期从事信息系统工程方面的教学、科研和开发，这套丛书是其多年学术研究和科技开发的成果总结，也是其多年教学工作中的实践积累，从丛书体系的设置到内容的安排，都基本体现了对当今信息系统工程领域前沿技术的把握。

这套丛书准备分批出版，第一批由《信息系统原理与工程》、《信息系统集成技术》、《信息系统建模》、《多媒体信息系统》、《智能协作信息技术》、《数据仓库原理与应用》、《语义信息模型及应用》等7部教材和专著组成，再加上该单位近年已出版的《决策支持系统技术》和《智能决策支持技术》两部研究生教材，基本上已覆盖了上述的信息系统工程主要研究范围。其中：

《信息系统原理与工程》主要介绍信息系统的基本概念、基本原理、技术和设计开发方法。

具体包括信息系统与信息系统工程的基本概念，信息系统中的基础理论、开发方法，结构化系统分析、系统设计和面向对象的分析设计方法，信息系统战略规划，系统实施，信息系统对计划、控制、决策的支持，计算机辅助信息系统开发等。

《信息系统集成技术》主要介绍信息系统集成的基本概念、基本原理和设计开发方法。首先介绍信息系统集成技术的发展，然后从体系结构入手，分网络集成、数据集成和应用集成三个层次展开对信息系统集成的论述，并给出了系统集成的案例。

《信息系统建模》主要介绍信息系统建模的基本概念、基本原理、方法、工程技术与工具。具体包括面向信息系统建模的思想，需求建模，逻辑建模，对象建模，Agent 建模，数据建模，统一建模语言等，是国内第一部按照较完整的体系专门介绍信息系统建模技术的著作。

《多媒体信息系统》主要介绍多媒体信息系统的基本概念、原理、技术和应用，主要内容包括多媒体信息系统的体系结构和数据模型、多媒体数据库和信息管理、多媒体通信和网络、多媒体人机交互与表现技术、原型系统与应用等。

《智能协作信息技术》主要介绍智能协作信息技术及系统的基本概念、基本原理和设计开发方法。具体包括智能协作信息技术的发展概况，智能主体概念、性质、内部结构和实现方法，多智能主体协作的基本原理、实现技术等，还介绍了智能协作信息系统的开发方法和智能协作信息技术在工业、管理、办公自动化等领域的应用。它是国内第一部全面介绍智能协作信息技术和智能协作信息系统的专著。

《数据仓库原理与应用》主要介绍数据仓库的概念、基本原理、规划、开发方法以及相关算法，包括数据仓库的发展、技术体系、元数据管理、分析设计方法和开发工具，并对数据开采的主要理论和方法、联机分析等应用技术作了深入的阐述，是一本理论与实践相结合的教材，是国内较为全面地分析数据仓库、开发数据仓库的书籍。

《语义信息模型及应用》深入到目前信息管理领域的前沿，探讨了语义信息模型的基本概念，并以 XML 为具体实现手段介绍了语义信息模型在信息组织、信息处理、信息服务、信息交换等方面高级应用的原理与实现机制。

除《语义信息模型及应用》以外，丛书中所有教材都作为内印教材或讲义试用过多次，吸收了许多专家学者以及学生的意见。

这套丛书既能够使广大读者从整体上把握知识结构、理清相关技术领域的关系和分类，又能够从中找到每项具体理论、技术、方法、工具的介绍和例解，再加上融合了多项“九五”期间的高水平科研成果，应该使这套丛书具有较高的系统性和实用性。

《信息系统工程丛书》是一套理论与工程实践并重的著作，它不仅可作为相关专业的大学本科生和研究生的系列化教材和参考书，而且也可以为从事信息系统工程的科研人员提供参考。我们相信，这套丛书的出版，将对我国信息系统工程的全面、深入发展起到重要的推动和促进作用。

《信息系统工程丛书》编委会

2001 年 6 月

# 前　言

近年来，在信息技术领域兴起并日益成熟的数据仓库技术成为了研究和应用的热点。从20世纪90年代初数据仓库概念的提出到现在短短的十几年时间，数据仓库便从理论研究进入到实际应用，发展速度十分惊人。实践证明，数据仓库技术在为企业科学地提高决策支持水平，提高企业信息质量和企业的应变能力等方面具有重要意义。

从应用角度来看，数据仓库是一种解决问题的方案，而不仅仅是技术和产品。数据仓库的建立是一个决策分析系统实施的过程。一方面，需要根据企业自身的管理特征具体分析，针对性地建立符合自身要求的数据仓库应用系统；另一方面，建立数据仓库需要有数据支持。企业管理信息化程度、数据的真实性、准确性、有效性、规范性是建立数据仓库的基础。只有当企业信息化程度达到一定水平，才能建立有效的数据仓库为企业管理和决策服务。

从技术角度来看，数据仓库以数据库技术作为存储数据和资源管理的手段，以联机分析处理技术和方法作为提取信息的有效手段，以数据挖掘、人工智能中的模型、算法作为发现知识和规律的途径。因此，数据仓库是诸多学科相互交叉、综合应用的技术。本书将从数据仓库技术的原理、开发和应用的角度阐述这些技术的相互作用和相互之间的关系。

本书共分8章，第1章简要概述数据仓库的发生、发展、组成，以及数据仓库的特征；第2章从理论的角度阐述数据仓库的基本原理、结构、元数据、开放信息模型以及关键技术；第3章全面阐述建立数据仓库的分析、设计直至实现过程中的内容、方法和步骤；第4章介绍数据仓库管理的技术和方法、数据管理的方式及工具软件；第5章介绍联机分析处理技术的概念、结构、方法和分析工具；第6章阐述数据挖掘的各类技术；第7章讲述数据挖掘的主要算法；第8章以一个实例说明数据仓库的设计、开发和使用的过程。

在本书编著过程中，陈文伟教授、姚庭宝教授和黄金才同志给予了大力帮助和支持。

由于编者水平有限，书中不免存在一些缺点和欠妥之处，恳请广大读者批评指正。

作　者

# 目　录


**第 1 章　概述** …… (1)
1.1　数据库与决策支持技术的发展 …… (1)
1.1.1　数据库技术的发展 …… (1)
1.1.2　决策支持技术的发展 …… (2)
1.2　数据仓库技术的发展 …… (3)
1.2.1　数据仓库概念的提出 …… (3)
1.2.2　数据仓库的发展 …… (5)
1.2.3　数据仓库技术的兴起 …… (6)
1.2.4　数据仓库的动态 …… (7)
1.3　数据挖掘技术的发展 …… (9)
1.3.1　数据挖掘研究和应用面临的挑战 …… (9)
1.3.2　数据仓库与数据挖掘的关系 …… (10)
1.4　数据仓库未来发展方向 …… (11)
**第 2 章　数据仓库原理** …… (15)
2.1　数据仓库的概念 …… (15)
2.1.1　数据仓库的定义 …… (15)
2.1.2　数据仓库的特征 …… (16)
2.1.3　数据集市 …… (18)
2.2　数据仓库的技术要求 …… (19)
2.3　数据仓库的结构 …… (21)
2.3.1　数据仓库的自顶向下结构 …… (21)
2.3.2　数据仓库的自底向上结构 …… (22)
2.3.3　企业级数据集市结构 …… (23)
2.3.4　数据存储/数据集市结构 …… (24)
2.3.5　分布式数据仓库/数据集市结构 …… (25)
2.3.6　分布式知识管理结构 …… (25)
2.3.7　数据仓库系统的结构 …… (26)
2.3.8　数据仓库的数据组织 …… (28)
2.4　元数据 …… (31)
2.4.1　元数据的由来 …… (31)
2.4.2　元数据的定义 …… (32)
2.4.3　元数据的主要作用 …… (33)
2.4.4　元数据的分类 …… (34)
2.4.5　元数据的标准化 …… (37)
2.4.6　OIM 简介 …… (40)

**第 3 章　数据仓库的设计** …… (46)
3.1　数据仓库的方法论 …… (46)
3.2　数据仓库规划 …… (48)
3.3　数据仓库体系结构 …… (49)
3.4　数据仓库的技术体系结构 …… (50)
3.5　数据仓库的数据组织 …… (54)
3.5.1　维表和事实表构成的关系型数据仓库 …… (55)
3.5.2　多维数据库数据组织 …… (57)
3.5.3　两种数据组织的等价性 …… (58)
3.5.4　虚拟数据仓库 …… (59)
3.6　数据仓库的粒度 …… (60)
3.6.1　粒度确定 …… (60)
3.6.2　粒度划分示例 …… (62)
3.7　数据仓库开发 …… (63)
3.7.1　定义体系结构 …… (64)
3.7.2　决策者的需求 …… (65)
3.7.3　主题区分析 …… (65)
3.7.4　源系统分析 …… (66)
3.7.5　变换设计 …… (66)
3.7.6　物理数据库设计 …… (67)
3.7.7　最终用户访问方法的设计、定义和开发 …… (67)
3.7.8　数据仓库开发 …… (68)
3.7.9　数据仓库填充和实施 …… (69)
3.7.10　数据仓库的开发流程 …… (69)
3.8　数据仓库解决方案 …… (70)
3.8.1　Sybase 提供的数据仓库解决方案 …… (70)
3.8.2　SAS 提供的数据仓库解决方案 …… (70)
3.8.3　Platinum 提供的数据仓库解决方案 …… (72)
3.8.4　其他解决方案 …… (74)
**第 4 章　数据仓库管理技术** …… (75)
4.1　数据仓库管理的基本问题 …… (75)
4.2　数据仓库中的多维建模技术 …… (76)
4.2.1　多维模型的两种结构 …… (77)
4.2.2　多维建模在决策支持系统中的应用 …… (79)
4.2.3　多维建模面临的挑战 …… (81)
4.3　休眠数据管理 …… (82)
4.3.1　问题的提出 …… (82)
4.3.2　休眠数据对数据仓库的影响 …… (83)
4.3.3　解决方案 …… (83)
4.4　元数据的管理 …… (87)

4.4.1 早期的数据管理:从内部管理到数据字典 …… (87)
4.4.2 企业级中心知识库的管理方法 …… (87)
4.4.3 传统的元数据管理方法 …… (89)
4.4.4 元数据的数据仓库管理功能 …… (90)
4.4.5 数据仓库研究项目和元数据管理介绍 …… (94)
4.4.6 评估元数据的价值 …… (97)
4.4.7 管理元数据 …… (98)
4.5 数据仓库管理工具 …… (98)
**第5章 联机分析处理** …… (101)
5.1 概述 …… (101)
5.1.1 OLAP的出现 …… (101)
5.1.2 OLAP的定义 …… (102)
5.1.3 OLAP的结构 …… (103)
5.1.4 OLAP的一些基本概念 …… (105)
5.1.5 OLAP的基本分析操作 …… (106)
5.1.6 OLAP与OLTP的比较 …… (109)
5.2 多维OLAP与关系OLAP …… (111)
5.2.1 多维数据存储与关系数据存储 …… (111)
5.2.2 OLAP服务器 …… (112)
5.2.3 MOLAP …… (112)
5.2.4 ROLAP …… (113)
5.3 OLAP技术分析 …… (118)
5.3.1 结构分析 …… (118)
5.3.2 数据存储和管理 …… (119)
5.3.3 数据存取 …… (119)
5.3.4 多维模型的实现技术 …… (120)
5.3.5 OLAP的12条准则 …… (122)
5.3.6 OLAP服务器和工具的评价 …… (125)
5.4 实用OLAP技术简介 …… (127)
5.4.1 Oracle OLAP工具 …… (127)
5.4.2 Oracle Express Server技术特色 …… (128)
5.4.3 Informix OLAP工具 …… (134)
**第6章 数据挖掘技术** …… (140)
6.1 数据挖掘概念、方法与任务 …… (140)
6.1.1 基本概念 …… (140)
6.1.2 数据挖掘的任务与分类 …… (142)
6.1.3 数据挖掘的方法和技术 …… (144)
6.1.4 数据挖掘的现状与应用 …… (147)
6.2 关联规则的发现 …… (153)
6.2.1 关联规则简介 …… (153)

6.2.2 关联规则的基本概念 …… (154)
6.2.3 关联规则发现的经典算法 …… (155)
6.2.4 基于聚类的周期关联规则发现算法 CCAR …… (159)
6.2.5 关联规则价值衡量的方法 …… (162)
6.3 公式发现 …… (164)
6.3.1 现状 …… (164)
6.3.2 问题描述 …… (165)
6.3.3 BACON 系统 …… (165)
6.3.4 FDD 系统 …… (168)
6.3.5 Explore 系统 …… (168)
6.4 数据聚类 …… (172)
6.4.1 聚类的概念 …… (172)
6.4.2 SAS 的聚类算法 …… (173)
6.4.3 基于遗传算法的聚类方法 …… (175)
6.4.4 基于随机搜索的聚类算法 …… (176)
6.4.5 聚类算法 BIRCH …… (177)
**第 7 章 数据挖掘算法** …… (182)
7.1 数据挖掘的集合论方法 …… (182)
7.1.1 粗集方法 …… (182)
7.1.2 概念树方法 …… (186)
7.1.3 覆盖正例排斥反例方法 …… (188)
7.2 数据挖掘中的决策树方法 …… (188)
7.2.1 基本原理 …… (188)
7.2.2 ID3 决策树方法 …… (190)
7.2.3 IBLE 决策规则树方法 …… (194)
7.2.4 决策树方法的优点和发展 …… (199)
7.3 数据挖掘中的遗传算法 …… (200)
7.3.1 遗传算法的形成和发展 …… (200)
7.3.2 遗传算法的基本原理 …… (200)
7.3.3 遗传算法的研究方向 …… (204)
7.3.4 基于遗传算法的分类系统 …… (205)
7.3.5 基于混合数据的遗传分类算法 …… (206)
7.4 数据挖掘的神经网络方法 …… (208)
7.4.1 神经网络的理论基础 …… (208)
7.4.2 几个常见神经网络 …… (211)
7.4.3 非线性神经网络的原理及其学习算法 …… (218)
**第 8 章 数据仓库应用** …… (222)
8.1 需求分析 …… (222)
8.1.1 环境分析 …… (223)
8.1.2 业务数据库结构分析 …… (225)

8.1.3 数据仓库应用系统的分析主题 …… (226)
8.1.4 数据仓库应用系统的具体要求 …… (227)
8.2 数据仓库应用系统设计 …… (227)
8.2.1 数据仓库应用系统结构 …… (227)
8.2.2 数据模型设计 …… (231)
8.3 数据转移 …… (240)
8.3.1 数据转移方案 …… (240)
8.3.2 数据装载 …… (241)
8.4 创建多维数据集 …… (242)
8.5 小结 …… (243)
**参考文献** …… (244)

# 第1章 概 述

众所周知，管理信息系统早已成功应用于全球的各行各业，并积累了大量的数据，基本上满足了用户对数据存储、查询和统计的需要。可以说，管理信息系统的成功得益于数据库技术的进一步完善。但是，用户目前面临的问题是怎么从大量的数据中获得自己需要的信息，尤其是决策者需要的信息。这些信息不仅仅来自本部门，同时还要考虑所处环境下的全方位信息，而这一点，现有的管理信息系统已经很难完成了。

从1997年开始，全球数据库市场就流传着不景气的说法，各大数据库厂商纷纷在寻找新的增长点。开始，大家都把注意力放在对象关系数据库技术上。而到了1998年，各厂商又纷纷转向数据仓库，使得数据仓库(DW，Data Warehouse)技术在短短的时间内，从思想走向应用，在积累了大量的经验的基础上，又逐步走向成熟和工具化。

本章介绍数据仓库的起源和现状。

## 1.1 数据库与决策支持技术的发展

数据库是决策支持技术的关键，数据库技术的成熟也是数据仓库技术提出的基础，而数据仓库和数据挖掘技术是决策支持新技术，它将成为一体化信息支持系统的核心技术。

### 1.1.1 数据库技术的发展

数据库系统是数据库和数据库管理系统的总称。数据库系统是适合于大量数据的存储和管理的有效方法。

1968年美国IBM公司研制的信息管理系统是著名的层次型数据库系统的典型代表。1969年10月美国CODASYL的数据库任务组提出了网络数据库模型的数据规范，并于1971年4月发表了DBTG报告，正式确定了数据库设计的网络方法(DBTG方法)，从而真正把数据库和文件系统区别开来，为数据库技术奠定了基础。1970年6月E.F.Codd提出了数据库关系模型，开创了数据库的关系方法和数据库规范化理论的研究。20世纪80年代以来，关系型数据库理论日益成熟并得到空前广泛的应用，这一阶段数据库理论和技术主要在两个方面得到了进一步发展。一方面是采用新数据模型(如面向对象数据模型、对象-关系数据模型)构造数据库，将数据库系统从传统的事务处理领域扩展到更广泛的领域，如应用在计算机辅助设计/制造(CAD/CAM)、计算机辅助软件工程(CASE)和地理信息系统(GIS)等领域中，满足对复杂对象的存储和处理要求；另一方面是数据库技术与其他学科的发展高度结合，例如数据库技术与分布处理技术结合导出的分布式数据库，数据库技术与人工智能技术结合导出的演绎数据库、智能数据库和主动数据库，数据库技术与多媒体技术结合导出的多媒体数据库等。

但是，总的来说，在20世纪90年代，数据库技术并没有出现革命性的创新。不论是IBM于1998年9月发布的IBM DB2 UBD5.2，Oracle于1998年11月发布的Oracle 8i，还是Sybase于1998年10月发布的ASE(Adaptive Server Enterprise)11.9.2，以及微软重彩描绘的SQL Server 7.0，尽管其名称各不相同，但都只能说是一般的技术改进。

## 1.1.2 决策支持技术的发展

1. 从时间角度看

(1) 20 世纪 50 年代到 20 世纪 60 年代:数据处理(Data Processing)阶段。

数据处理是电子计算机应用中最广泛的领域,约占 70%。一个国家的现代化水平越高,数据处理的面越宽,量越大,数据处理所占的比例就越高。

(2) 20 世纪 60 年代到 20 世纪 70 年代:管理信息系统(MIS)阶段。

随着 20 世纪 50 年代到 20 世纪 60 年代数据处理领域应用的成功,20 世纪 60 年代到 20 世纪 70 年代西方国家兴起了管理信息系统的热潮,我国是 20 世纪 70 年代末到 20 世纪 80 年代初才兴起了管理信息系统的应用。

(3) 20 世纪 70 年代到 20 世纪 80 年代:决策支持系统(DSS)阶段。

管理信息系统是在管理科学利用计算机后发展起来的,它使计算机的应用由数值计算领域拓宽到数据处理(非数值计算)领域,使计算机走向社会和家庭。运筹学和系统工程利用计算机后形成了模型辅助决策系统,由于采用的模型主要是数学模型,它辅助决策的能力主要表现在定量分析上。

20 世纪 70 年代初发展起来的决策支持系统把管理信息系统和模型辅助决策系统结合起来,使得数值计算和数据处理融为一体,提高了辅助决策的能力。

(4) 20 世纪 90 年代:智能决策支持系统(IDSS)阶段。

该阶段的主要特征是模型技术、专家系统、数据仓库和数据挖掘技术的全方位的有机集成,使得决策支持技术无论是在体系结构还是在信息处理能力上都产生了较大的变化。

2. 从技术角度看

20 世纪 60 年代末兴起了一个新研究领域——专家系统(ES,Expert System),它是 20 世纪 50 年代人工智能的进一步发展。专家系统是利用专家的知识在计算机上进行推理,达到专家解决问题的能力。1968 年由 E. A. Feigenbanm 等人研制了 DENDRAL 专家系统,用来帮助化学家推断分子结构。1974 年由 E. H. Shortliffe 等人研制的 MYCIN 专家系统,用来诊断和治疗感染性疾病。同一时期,人们还研制出不少其他专家系统。专家系统的出现使人工智能走上了实用化阶段。

专家的知识表现为产生式规则和语义网络等形式。知识的推理是采用符号逻辑中的假言推理。在搜索知识的时候,采用了深度优先或启发式搜索方法。专家系统也是一种很有效的辅助决策系统,它是利用专家的知识,特别是经验知识,经过推理得出辅助决策信息。对于专家知识,不限定它是数值的,更多的是不精确的定性知识,这样,专家系统辅助决策的方式属于定性分析。

专家系统的发展使它逐步深入到各个领域,并取得了很大的经济效益。例如,数字设备公司(DEC)的 XCON 系统,它是为顾客配置计算机系统的专家系统,该专家系统每年为 DEC 公司节省数百万美元。

专家系统和决策支持系统几乎是同时兴起,并各自沿着自己的道路发展起来的,它们都能起到辅助决策的作用,但辅助决策的方式完全不同。专家系统辅助决策的方式是属于定性分析;决策支持系统辅助决策的方式是属于定量分析。如果把这两者结合起来,辅助决策的效果将会大大改善,即达到定性辅助决策和定量辅助决策相结合。这种专家系统和决策支持系统结合形成的系统称为智能决策支持系统,它是决策支持系统的发展方向。

决策支持系统和专家系统的结合，并不是那样容易实现，因为它们各自自成体系，要结合它们将有一些技术难题需要解决。专家系统结构中核心的部分由推理机、知识库和动态数据库三部分组成。知识库存放大量的专家知识；推理机完成对知识的搜索和推理；动态数据库存放已知的事实和推出的结果。专家系统中的动态数据库不同于决策支持系统中的数据库，相对来说，决策支持系统中的数据库是静态数据库。两系统中各部件之间的接口以及两系统的集成，是形成智能决策支持系统的关键。

## 1.2 数据仓库技术的发展

随着社会的发展和技术的进步，信息已成为人类社会中除了物质、能量之外的第三大资源。社会的信息化，使信息量急剧增长，大量的信息来不及组织和处理。面对急剧增长的信息，对数据库系统的应用只停留在查询、检索、统计等几个方面，远远没有发挥数据库中的数据的作用和价值。

正如奈斯比特在《大趋势》中所说的："我们正在被信息所淹没，但我们却由于缺乏知识而感到饥饿。"数据库容量的指数增长和对数据库应用的贫乏形成了强烈的反差，导致了大量的数据垃圾。

### 1.2.1 数据仓库概念的提出

众所周知，如何有效地管理公司和企业在运营过程中产生的大量数据和信息一直是 IT 人员面临的重要问题。20 世纪 70 年代出现并被广泛应用的关系型数据库技术为解决这一问题提供了强有力的工具。然而，从 20 世纪 80 年代中期开始，随着市场竞争的加剧，信息系统用户已经不满足于仅仅用计算机去管理日复一日的运营数据，他们更需要的是从这些数据中得到有用的信息，以便于进行决策支持，这种需求使得在 20 世纪 80 年代中后期出现了数据仓库思想的萌芽，为数据仓库概念的最终提出和发展打下了基础。1992 年，W. H. Inmon 在其里程碑式的《建立数据仓库》一书中提出了"数据仓库(DW，Data Warehouse)"的概念，数据仓库的研究和应用得到了广泛的关注。

随着信息处理技术的发展，各类数据、信息急剧增长，给数据的传输、存储都带来许多新问题，特别是由于各类不同事务产生大量不同类型的数据，这些数据分别被许多各个时期建立的应用系统所使用。人们希望能够看到所有数据和信息的综合情况，而这些数据与事务处理有许多不能被原有数据结构描述，不能被现有应用系统综合使用。因此在原有的单一数据库概念的基础上，演化出两种不同的数据组织体系结构，它们是数据仓库和原有的业务数据库(Operational Database)。这两个概念之间存在着许多不同，包括两种不同的用户环境，不同的支持技术，不同的数据量以及不同的使用范围等等。

在数据库技术当前及未来的发展里程中，数据仓库以及基于此技术的商业智能无疑将是大势所趋，从而必将成为兵家必争之地。IBM 的实验室在这方面进行了十多年的研究，并将研究成果发展成为商用产品。IBM 在其 DB2UDB 发布一年后的 1998 年 9 月又发布了其 5.2 版，并于 1998 年 12 月推向中国市场，除了用于联机分析处理(OLAP)的后台服务器 DB2OLAPServer 外，IBM 还提供了一系列相关的产品，包括前端工具，来形成一整套解决方案。

其他数据库厂商在数据仓库领域也毫不示弱，方法各有不同。IBM 在一个通用的数据库系统中实现联机事务处理(OLTP)和联机分析处理(OLAP)。相比之下，Oracle 采取了类似的

方法。Informix 也是如此，在其动态服务器 IDS(Informix Dynamic Server)中提供一系列相关选件，如高级决策支持选件(Advanced Decision Support Option)、OLAP 选件(MetaCube ROLAP Option)、扩展并行选件(Extended Parallel Option)等，并认为这种体系结构严谨，管理方便，索引机制完善，并行处理的效率更高，其中数据仓库和数据库查询所用的 SQL 语句的一致性使用户开发更加简便。而微软则是在其 SQL Server 7.0 中集成了代号为 Plato 的 OLAP 服务器，这种做法不知是否会引起业界对其将 Internet 浏览器 IE 集成在 Windows 操作系统中相类似的起诉。与上述公司不同的是，Sybase 提供了专门的 OLAP 服务器 Sybase IQ，并将其与数据仓库相关工具打包成 Warehouse Studio。

实际上，世界上最大的数据仓库系统当数 NCR 公司建立的基于其 Teradata 数据库、拥有 24 TB 数据量的 WalMart(美国最大的零售连锁店)数据仓库系统，并产生了业界经典的“尿布与啤酒”的故事。和国外应用情况相比，尽管各厂商在数据仓库方面的“演出”都很卖力，但在中国市场上的收效仍很有限。从中国的数据库市场来看，大部分数据库系统的建立是用来进行传统的 OLTP 业务。也有一些企业建立了数据仓库系统，但真正发挥效用的却不多见。和 TCP/IP，SMTP，Java 等相比，业界尚不存在可靠的、完善的、被广泛接受的数据仓库标准，影响了数据仓库项目的实施。

目前在信息管理方面存在的普遍性问题包括：

1.“数据太多，信息不足”的现状

随着数据库技术的发展，各企业积累并存放了大量业务数据，但能够为企业提供辅助决策的信息太少，需要改变目前现状。

2.异构环境的数据源

由于市场竞争激烈，新产品周期缩短，如何综合利用分散的异构环境数据源，及时得到准确的信息是使企业取得成功的关键。

3.事务处理环境不适宜 DSS 应用

(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；在分析处理环境中，用户的行为模式与此完全不同，某个 DSS 应用程序可能需要连续使用几个小时，从而消耗大量的系统资源。将具有如此不同处理性能的两种应用放在同一个环境中运行显然是不适当的。

(2) 数据集成问题：DSS 需要集成的数据，全面而正确的数据是有效地分析和决策的首要前提，相关数据收集得越完整，得到的结果就越可靠。当前绝大多数企业内数据的真正状况是分散而非集成的，造成这种分散的原因有多种，主要有事务处理应用分散，数据不一致问题，外部数据和非结构化数据问题等。

(3) 数据动态集成问题：静态集成的最大缺点在于，如果在数据集成后数据源中数据发生了变化，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。集成数据必须以一定的周期(例如 24h)进行刷新，我们称其为动态集成。显然，事务处理系统不具备动态集成的能力。

(4) 历史数据问题：事务处理一般只需要当前数据，在数据库中一般也是存储短期数据，而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，也被束之高阁，未得到充分利用。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，没有历史数据的详细分析，是难以把握企业的发展趋势的。DSS 对数据在空间和时间的广度上都有了更高的要求，而事务处理环境难以满足这些要求。

(5) 数据的综合问题:在事务处理系统中积累了大量的细节数据,一般而言,DSS并不对这些细节数据进行分析。在分析前,往往需要对细节数据进行不同程度的综合,而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。

要提高分析和决策的效率和有效性,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DSS处理的需要进行重新组织,建立单独的分析处理环境。数据仓库正是为了构建这种新的分析处理环境而出现的一种数据存储和组织技术。

随着市场竞争的加剧和信息社会需求的发展,从大量数据中提取(检索、查询等)制定市场策略的信息就显得越来越重要了。这种需求既要求联机服务,又涉及大量用于决策的数据,而传统的数据库系统已无法满足这种需求。这具体体现在三个方面:① 历史数据量很大。② 辅助决策信息涉及许多部门的数据,而不同系统的数据难以集成。③ 由于访问数据的能力不足,它对大量数据的访问性能明显下降。

## 1.2.2 数据仓库的发展

1. 信息管理思想的飞跃

二十多年来,大量新技术、新思路涌现出来并被用于关系数据库系统的开发和实现,如客户机/服务器(C/S)体系结构、存储过程、多线索并发内核、异步I/O、代价优化等等,这一切足以使得关系数据库系统的处理能力毫不逊色于传统封闭的数据库系统。而关系数据库在访问逻辑和应用上所带来的好处则远远不止这些,SQL(the Structured Query Language)的使用已成为一个不可阻挡的潮流,加上近些年来计算机硬件的处理能力呈数量级的递增,关系数据库最终成为联机事务处理系统的主宰。整个20世纪80年代到20世纪90年代初,联机事务处理一直是数据库应用的主流,然而,应用在不断地进步。当联机事务处理系统应用到一定阶段的时候,企业家们便发现单靠拥有联机事务处理系统已经不足以获得市场竞争的优势,他们需要对其自身业务的运作以及整个市场相关行业的态势进行分析,给出有利的决策,这种决策需要对大量的业务数据包括历史业务数据进行分析才能得到。在如今这样激烈的市场竞争环境下,这种基于业务数据的决策分析比以往任何时候都显得更为重要,我们把它称之为联机分析处理。如果说传统联机事务处理强调的是更新数据库——向数据库中添加信息,那么联机分析处理就是从数据库中获取信息和利用信息。因此,著名的数据仓库专家Ralph Kimball写道:"我们花了二十多年的时间将数据放入数据库,如今是该将它们拿出来的时候了。"

事实上,将大量的业务数据应用于分析和统计原本是一个非常简单和自然的想法。但在实际的操作中,人们却发现要获得有用的信息并非如想像的那么容易。

(1) 所有联机事务处理强调的是密集的数据更新处理性能和系统的可靠性,并不关心数据查询的方便与快捷。联机分析和事务处理对系统的要求不同,同一个数据库在理论上难以做到两全。

(2) 业务数据往往被存放于分散的异构环境中,不易统一查询访问,而且还有大量的历史数据处于脱机状态,形同虚设。

(3) 业务数据的模式针对事务处理系统设计,数据的格式和描述方式并不适合非计算机专业人员进行业务上的分析和统计。

因此有人感叹,20年前查询不到数据是因为数据太少了,而今天查询不到数据是因为数