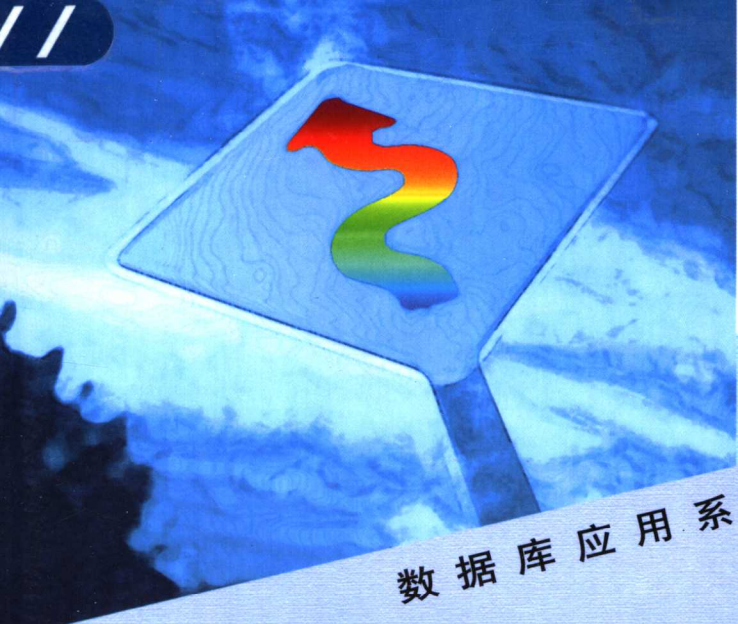




数据库应用系列教材



数据仓库与数据挖掘 原理及应用

王丽珍 周丽华 陈红梅 邹力鹏 编著



科学出版社

www.sciencep.com

数据库应用系列教材

数据仓库与数据挖掘 原理及应用

王丽珍 周丽华 陈红梅 邹力鹏 编著

科学出版社

北 京

内 容 简 介

本书全面深入地介绍了数据仓库、联机分析处理(OLAP)和数据挖掘的基本概念、基本原理和应用技术。全书分成三篇,数据仓库及 OLAP 概念、原理和技术篇的主要内容包括数据仓库的基本概念、体系结构、模型设计、创建和维护,ETL、元数据、数据集市,OLAP 的基本概念、分类、模型设计;数据挖掘技术篇介绍了数据挖掘的基本理论、基本过程、常见模型和算法;工具及实例介绍篇简要介绍了数据仓库产品工具的基本情况,对产品选择和评判进行了一些分析,并较详细地介绍和分析了移动通信业务数据仓库系统。

本书可作为计算机、信息系统等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、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及应用/王丽珍等编著. —北京:科学出版社, 2005

(数据库应用系列教材)

ISBN 7-03-015657-9

I. 数… II. 王… III. ①数据库系统-教材②数据采集-教材 IV. ① TP311.13②TP274

中国版本图书馆 CIP 数据核字(2005)第 058632 号

责任编辑:鞠丽娜 韩 洁/责任校对:柏连海

责任印制:吕春珉/封面设计:三函设计

科 学 出 版 社 出 版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

双 青 印 刷 厂 印 刷

科学出版社发行 各地新华书店经销

*

2005 年 7 月第 一 版 开本: B5 (720×1000)

2005 年 7 月第一次印刷 印张: 19 3/4

印数: 1—4 000 字数: 378 000

定价: 26.00 元

(如有印装质量问题,我社负责调换<环伟>)

销售部电话 010-62136131 编辑部电话 010-62138978-8002 (H106)

《数据库应用系列教材》编委会

主 任 王 珊 中国人民大学

徐洁磐 南京大学

编 委 (按姓氏笔画排序)

马玉书 石油大学 (北京)

王能斌 东南大学

孙志挥 东南大学

许龙飞 暨南大学

李庆忠 山东大学

李昭原 北京航空航天大学

沈钧毅 西安交通大学

邵晓英 宁波大学

邵佩英 中国科学院研究生院

单启成 南京大学

唐世渭 北京大学

聂培尧 山东财政学院

郭景峰 燕山大学

黄上腾 上海交通大学

序

近年来,我国高等教育事业飞跃发展,在校学生人数突飞猛进,与此同时,高校教育改革逐渐冲破旧的计划经济模式,新的模式也正在建立。在这种形势下,旧的教材体系已不能适应新的需要,因此迫切需要建立新的教材体系。基于此种情况,我们以计算机相关专业中的数据库系统教材为依托,组织了一套适应不同需求、不同层次、不同目标的数据库系列教材,其组织依据是:

1. 在高等学校中随着老校的调整与改革,新校的不断涌现,过去计划经济的一刀切模式已逐渐改变,各校在培养目标、人才市场定位方面已出现多种模式(如研究型、应用型、开发型等),因此需要有多种不同数据库系统教材以适应不同模式的需求,而现有教材大多只能适应少数模式的需求。

2. 近年来计算机应用飞速发展,计算机与其他专业的交叉应用发展很快,如文科中的数量经济、信息管理、电子商务、财政金融等专业,理工科中的机械、建筑、城市规划、遥感遥测等都急需开设计算机及数据库等相应课程,也需相应的教材,而此方面的合适教材目前较为少见。

3. 随着教学改革的深入,数据库课程自身也需进行改革,它除了需要有主课程外,还需要有若干门配套的辅助性课程与教材,如数据库分析与设计、Web 数据库、数据库应用等课程,以及数据库实验课、实习课以及习题集等配套教材。此外,还需配合使用现代化手段如电子教案及课件等相关音像制品。所有这些教材都需构成一个以数据库主课程为核心的有机组合的系列教材,而此方面的组合教材正是目前所缺少的。

4. 数据库技术本身发展很快,而教材编写相对滞后,同时国内数据库教材又受国外教材影响较大,因此适合国情的本土化教材的建设尤为重要,因此,能编写出既适应目前技术发展水平,又能适应我国经济发展需要的数据库教材是当前之急需。

5. 本系列教材能适应不同模式,不同层次、不同系科(计算机及非计算机专业)的需求,它除追求基本原理的正确性外着重在它的应用性。由于数据库是一门实用性很强的课程,我们希望学生在学了此课程后能在实际应用中发挥作用。

本系列教材正是为适应上面所述的需要而编写的,目前它以计算机及非计算机专业的本科生教材为主,并将逐渐扩充到研究生及大专层次。本系列教材采取开放性组织方式,今后将根据学科发展陆续组织出版数据库领域的优秀图书。

本系列教材的编写人员涉及各个不同层次与专业,有大量实际经验与理论水平,相信这套教材的问世能对数据库教学起一定的促进作用。

《数据库应用系列教材》编委会

2004年9月

前 言

进入信息社会以来,信息技术经历了这样的发展过程:从计算机主机的信息集中处理方式到个人计算机(PC)的信息分布处理形式的转变;从单一的计算机操作系统到计算机互联网络操作的改变;从客户机/服务器(Client/Server)计算体系到多层体系结构计算模式的转变;从单一数据库到大型数据仓库和从局域网到Internet的改变。现代信息技术的发展和现代科学技术的进步,使人类迈入了新的时期——信息化时代。

信息处理技术的发展,使得各类数据、信息急剧增长,给数据的传输、存储带来了许多新的问题,特别是由于各类不同事务产生大量不同类型的数据,这些数据分别被各个时期建立的许多应用系统所使用。人们希望能够看到所有数据和信息的综合情况,而这些数据和信息有许多不能被统一描述,不能被现有应用系统综合使用。针对这一问题,人们设想专门为业务的统计分析建立一个数据中心,它的数据来自联机的事务处理系统、异构的外部数据源、脱机的历史业务数据等,这个数据中心就叫数据仓库。数据仓库技术的应运而生,成为信息技术领域非常热门的话题之一。

数据仓库技术的提出,建立了一种体系化的数据存储环境,将分析决策所需要的大量数据从传统的操作环境中分离出来,使分散、不一致的操作数据转换成集成、统一的信息。企业内不同单位、不同角色的成员都可以在此单一的环境之下,通过运用其中的数据与信息,发现全新的视野和新的问题,产生用于决策的新分析方法。作为决策支持系统的重要组成部分,数据仓库为决策支持系统提供了分析决策所需的数据;OLAP的产生进一步增强了决策支持系统快速、一致和交互性的分析能力,它利用存储在数据仓库中的数据完成各种分析操作,并以直观易懂的形式将分析结果展现给决策分析人员;而数据挖掘是从大量数据中提取或“挖掘”知识,从而实现从“数据→信息→知识”的过程,为企业的管理层提供各种层次的决策支持。

本书对数据仓库、OLAP、数据挖掘的原理、技术、工具和应用做了全面深入地介绍和分析,对数据仓库、OLAP和数据挖掘的发展及应用前景也进行了细致深入地讨论。全书共三篇,分别是数据仓库及OLAP概念、原理和技术篇、数据挖掘技术篇和工具及实例介绍篇。内容组织的思路为:基本概念→基本原理→实际应用。

本书在结构的组织上,采用引言→主体内容→小结→习题的结构形式。每章后面的习题可作为课后作业。这些习题或者是短问题,用于测试对内容的掌握;

或者是长问题，需要分析思考甚至查阅资料来完成。在内容的介绍上，除理论联系实际外，还使用了大量的图示及实例，使该书具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，都能学懂本书的内容。

讲授这门课程一般需 70 学时左右，因对问题阐述深入浅出，该书既可作为课堂教学的教材，也可供自学时参考。

本书的写作过程也是笔者学习、研讨、提高的过程。在此过程中，笔者对国内外大量的资料进行了归纳和整理，认真学习了各种数据仓库、OLAP 和数据挖掘工具，对参与及主持开发的两个数据仓库系统进行了全面的分析和总结。

在本书写作过程中，所写内容已经过反复研讨，且有些章节可能由某位老师执笔，而由另一位老师修改，因此难以严格划分每个人之工作量。就执笔而言，其分工如下：第 1、6、7、12、13 章由王丽珍执笔，第 3~5 章由周丽华执笔，第 8~10 章由陈红梅执笔，第 2、11 章由邹力鹏执笔。

本书的写作得到南京大学徐洁磐教授的鼓励和支持，徐教授认真审阅了全书，提出了许多宝贵的修改意见。云南大学研究生夏勇、胥玲芳、秦海林、陈涛、熊芸、陈克平、陈杉等为本书的完成做了大量的辅助工作。另外，本书还得到国家自然科学基金（编号 60463004）和云南省自然科学基金（编号 2002F0013M）的资助，在此一并表示衷心的感谢。

由于笔者水平有限，书中错漏和不妥之处在所难免，恳请读者批评指正。

作 者

2005 年 4 月

目 录

第一篇 数据仓库及 OLAP 概念、原理和技术篇

第 1 章 数据仓库基本概念	1
1.1 从数据库到数据仓库	1
1.1.1 蜘蛛网问题	1
1.1.2 事务型系统和分析型系统的分离	4
1.2 什么是数据仓库	6
1.2.1 面向主题	6
1.2.2 集成	7
1.2.3 稳定性	8
1.2.4 随时间而变化	9
1.3 数据仓库的体系结构	9
1.3.1 数据仓库的体系结构	9
1.3.2 数据仓库中的关键名词	10
1.4 数据仓库的数据组织	13
1.4.1 数据仓库的数据组织结构	13
1.4.2 数据粒度与数据分割	14
1.4.3 数据仓库的数据组织形式	15
1.4.4 数据仓库的数据追加和清理	17
1.5 本章小结	19
习题	19
第 2 章 数据仓库中的 ETL 和元数据	20
2.1 ETL	20
2.1.1 ETL 概念	20
2.1.2 ETL 作用	23
2.1.3 ETL 工具	23
2.2 元数据	26
2.2.1 什么是元数据	27
2.2.2 元数据的标准化	31
2.2.3 数据仓库中的元数据管理	32
2.2.4 在数据仓库项目中使用元数据的建议	34

2.3 外部数据.....	35
2.3.1 外部数据和非结构化数据	35
2.3.2 元数据和外部数据	36
2.3.3 外部数据的存储	36
2.3.4 外部数据的管理	37
2.4 本章小结.....	37
习题	38
第3章 数据仓库模型设计	39
3.1 数据仓库模型设计方法概述.....	39
3.2 数据仓库设计的三级数据模型.....	40
3.2.1 概念模型	41
3.2.2 逻辑模型	41
3.2.3 物理模型	41
3.2.4 三种模型之间的关系	41
3.2.5 高级模型、中级模型和低级模型	42
3.3 数据仓库的概念模型设计	43
3.3.1 E-R 模型.....	43
3.3.2 面向对象的分析方法	46
3.4 数据仓库的逻辑模型设计	48
3.4.1 分析主题, 确定当前要装载的主题	48
3.4.2 确定数据粒度的选择	49
3.4.3 确定数据分割策略	53
3.4.4 增加导出字段	54
3.4.5 定义关系模式	54
3.4.6 定义记录系统	55
3.5 数据仓库的物理模型设计	55
3.5.1 存储结构	55
3.5.2 索引策略	59
3.5.3 数据存储策略	65
3.5.4 存储分配优化	67
3.6 数据装载接口设计	68
3.7 本章小结.....	69
习题	69
第4章 数据仓库的建立和维护	71
4.1 数据仓库的投资分析	71
4.1.1 建设数据仓库的必要性	71

4.1.2 数据仓库的投资回报分析	72
4.2 数据仓库的开发方法	73
4.2.1 瀑布式开发	73
4.2.2 螺旋式开发	74
4.3 数据仓库的建立过程	74
4.3.1 需求分析	75
4.3.2 数据路线	76
4.3.3 技术路线	77
4.3.4 应用路线	81
4.3.5 数据仓库部署	87
4.3.6 运行维护	88
4.4 数据仓库的维护	88
4.4.1 数据周期	88
4.4.2 参照完整性	89
4.4.3 数据环境信息	90
4.4.4 数据备份与恢复	91
4.5 提高数据仓库性能	92
4.5.1 提高 I/O 性能	92
4.5.2 缩小查询范围	93
4.5.3 采取并行优化技术	93
4.5.4 选择适当的初始化参数	95
4.6 数据仓库的安全性	95
4.6.1 安全类型	96
4.6.2 安全方法	96
4.7 本章小结	100
习题	101
第 5 章 数据仓库与数据集市的关系	102
5.1 什么是数据集市	102
5.2 数据集市的类型	103
5.3 数据集市与数据仓库的区别	104
5.4 数据集市的特点	105
5.5 数据集市的开发方法	106
5.6 数据集市的建立	107
5.7 本章小结	108
习题	108

第 6 章 联机分析处理 (OLAP)	109
6.1 OLAP 概念	109
6.1.1 什么是 OLAP	109
6.1.2 OLAP 的相关基本概念	109
6.1.3 OLAP 和 OLTP 的区别	110
6.2 OLAP 的基本操作	111
6.2.1 数据切片	111
6.2.2 数据切块	113
6.2.3 数据上探/下钻	113
6.2.4 数据旋转	114
6.3 OLAP 分类和体系结构	115
6.3.1 OLAP 的三层客户/服务器结构	115
6.3.2 OLAP 的分类	115
6.3.3 OLAP 的体系结构	116
6.4 基于多维数据库的 OLAP (MOLAP)	118
6.4.1 多维数据库	118
6.4.2 维的分类	119
6.4.3 多维数据库存储	121
6.5 基于关系数据库的 OLAP (ROLAP)	121
6.5.1 维表和事实表	121
6.5.2 星型模型和雪花模型	125
6.5.3 星座模型和雪暴模型	127
6.5.4 ROLAP 与 MOLAP 比较	129
6.5.5 HOLAP	131
6.6 OLAP 的衡量和特性	132
6.6.1 OLAP 的 12 准则	132
6.6.2 OLAP 的简洁准则 (OLAP 的特性)	135
6.7 OLAP 的前端展现方式	136
6.7.1 OLAP 实现架构	136
6.7.2 OLAP 的 Web 呈现方式	137
6.7.3 瘦客户机方式	137
6.7.4 OLAP 的前端展现	137
6.8 OLAP 的发展及展望	140
6.8.1 OLAP 在应用领域的发展趋势	140
6.8.2 OLAP 基于 Web 的应用	142
6.8.3 OLAP 展望	142

6.9 本章小结.....	143
习题	143
第7章 数据仓库的应用前景.....	144
7.1 在电信业的应用前景	144
7.2 在客户服务及营销方面的应用前景.....	146
7.3 在银行领域的应用前景.....	147
7.4 在保险业的应用前景	148
7.5 在图书馆领域的应用前景	148
7.6 成功案例分析	149
7.7 本章小结.....	154
习题	154

第二篇 数据挖掘技术篇

第8章 数据挖掘介绍	155
8.1 数据挖掘概述	155
8.2 数据挖掘分类	157
8.2.1 概述	157
8.2.2 描述性挖掘	158
8.2.3 预测性挖掘	159
8.3 数据挖掘系统	160
8.3.1 数据挖掘系统的结构	160
8.3.2 数据挖掘系统的设计	161
8.3.3 数据挖掘系统的发展	163
8.4 数据预处理	164
8.4.1 概述	164
8.4.2 数据清理	165
8.4.3 数据集成	166
8.4.4 数据变换	166
8.4.5 数据归约	167
8.4.6 属性概念分层的自动生成	169
8.5 数据挖掘与数据仓库	172
8.6 数据挖掘的应用和发展	172
8.6.1 数据挖掘的应用	172
8.6.2 数据挖掘未来研究方向	174
8.7 本章小结.....	174
习题	175

第9章 描述性挖掘	176
9.1 特征与比较描述	176
9.1.1 特征与比较描述概述	176
9.1.2 面向属性归纳	177
9.1.3 特征与比较规则	181
9.2 关联规则挖掘	184
9.2.1 关联规则的基本概念	184
9.2.2 Apriori 算法	186
9.2.3 FP-growth 算法	191
9.3 聚类分析	195
9.3.1 聚类分析的基本概念	195
9.3.2 基于划分的聚类算法	201
9.3.3 基于密度的聚类算法	204
9.4 本章小结	208
习题	208
第10章 分类与预测	210
10.1 决策树分类算法	212
10.1.1 什么是决策树	212
10.1.2 决策树的建立	213
10.1.3 由决策树提取分类规则	218
10.1.4 对新对象分类	219
10.2 神经网络	219
10.2.1 前馈神经网络结构	219
10.2.2 神经网络学习	221
10.2.3 神经网络分类	225
10.3 回归分析	226
10.3.1 一元回归分析	226
10.3.2 多元回归分析	230
10.3.3 非线性回归	231
10.4 本章小结	233
习题	233

第三篇 工具及实例介绍篇

第11章 数据仓库工具介绍	235
11.1 数据仓库产品选择	235

11.1.1 数据仓库产品组成	235
11.1.2 数据仓库产品应具备的关键技术	236
11.1.3 数据仓库产品现状	237
11.1.4 如何选取数据仓库工具	237
11.2 常用数据仓库产品简介	238
11.2.1 Oracle 9i	238
11.2.2 NCR TeraData	239
11.2.3 IBM DB2	240
11.2.4 Informix	242
11.3 本章小结	245
习题	245
第 12 章 Cognos 介绍	246
12.1 Cognos 公司 BI 主要产品介绍	246
12.1.1 数据查询和即席报表生成工具	247
12.1.2 模型建立工具	252
12.1.3 在线分析处理及展现工具	257
12.2 Cognos 应用例子	259
12.2.1 报表的生成	259
12.2.2 Cube 的构造	265
12.3 本章小结	269
习题	269
第 13 章 移动通信业务数据仓库系统	271
13.1 系统介绍	271
13.1.1 系统建设的原则和目标	271
13.1.2 系统结构和功能	272
13.2 系统模型设计	274
13.2.1 概念模型设计	274
13.2.2 逻辑模型设计	279
13.2.3 物理模型设计 (PDM)	283
13.3 数据装载接口设计	286
13.3.1 概述	286
13.3.2 源数据分析	286
13.3.3 ETL	287
13.4 数据仓库的维护	288
13.4.1 数据周期	288
13.4.2 参照完整性	289

13.4.3 数据备份与恢复	289
13.5 前端分析展示	292
13.5.1 概述	292
13.5.2 前端分析展示设计及实现	293
13.5.3 Demo 演示	293
13.6 本章小结	297
习题	297
主要参考文献	299

第一篇 数据仓库及 OLAP 概念、原理和技术篇

第 1 章 数据仓库基本概念

近十几年,随着科学技术飞速的发展,经济和社会都取得了极大的进步,与此同时,在各个领域产生了大量的数据,如人类对太空的探索,银行每天的巨额交易数据。显然在这些数据中蕴藏着丰富的信息,如何处理这些数据得到有益的信息,人们进行了有益的探索。计算机技术的迅速发展使得处理数据成为可能,这就推动了数据库技术的极大发展,但是面对不断增加如潮水般的数据,人们不再满足于数据库的查询功能,提出了深层次问题:能不能从数据中提取信息或者知识为决策服务。就数据库技术而言已经显得无能为力了,这就急需有新的方法来处理这些海量般的数据。在这种情况下,数据库逐步发展到了数据仓库。世界上最早的数据仓库是 NCR 公司为全美、也是全世界最大的连锁超市集团 Wal★Mart 在 1981 年建立的,而最早将数据仓库提升到理论高度进行分析并提出数据仓库这个概念的则是著名学者 W.H.Inmon,他对数据仓库所下的定义是:数据仓库是面向主题的、集成的、稳定的、随时间变化的数据集合,用于支持管理决策过程。由此可见,数据仓库是一个综合的解决方案,主要用来帮助企业有关主管部门和业务人员做出更符合业务发展规律的决策。

1.1 从数据库到数据仓库

1.1.1 蜘蛛网问题

在市场经济的激烈竞争中,信息对于企业的生存和发展起着至关重要的作用。企业对信息的需求是多方面的,为了避免企业中各部门或各用户间的冲突和简化用户的数据视图,一种称为“抽取程序”的方法被广泛地应用。比如,市场部人员通常只关心企业的销售、市场策划方面的信息,而不注重企业的研发、生产等其他环节。因此,将销售、市场策划方面的信息抽取出来单独建立部门级的数据库很有必要,这样可以提高数据的访问效率。在部门级数据的基础上可能还要被继续执行抽取程序,以建立个人级的数据库。比如,专门负责制作公司财务报表的数据人员,常常需要从财务部门的数据库系统中抽取数据。又如,部门经理可能经常抽取常用的数据到本地,有针对性的建立个人级数据库就显得尤为重要。

随着数据的逐层抽取,很可能最终导致系统内的数据间形成了错综复杂的网状结构,如图 1.1 所示,人们形象地称其为“蜘蛛网”。一个大型的公司每天进行上万次的数据抽取很普遍。这种演变不是人为制造的,而是自然演变的结果。企业的规模越大,“蜘蛛网”问题就越严重。

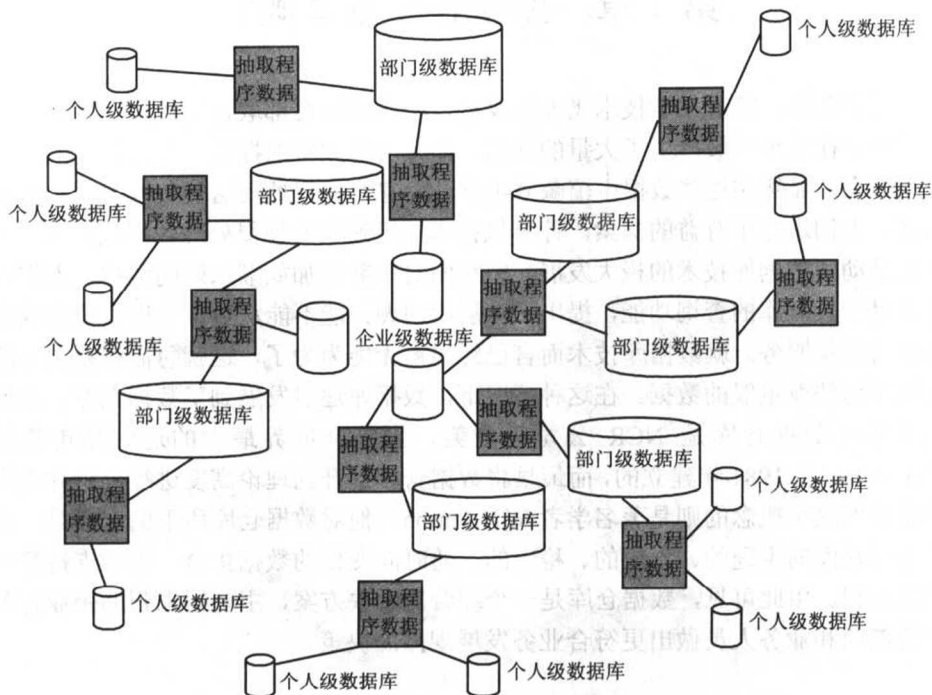


图 1.1 企业中存在的“蜘蛛网”现象

虽然网上的任意两个节点的数据可能归根结底是从一个原始库中抽取出来的,但其数据没有统一的时间基准,因而错综复杂的抽取与访问将产生很多的问题,主要有以下几个方面。

1. 数据分析的结果缺乏可靠性

图 1.2 中展示了某企业的市场部和计划部对项目 I 是否具有市场前景的分析过程和结果。市场部认为“项目 I 的市场前景很好”,而计划部却得到截然相反的结果——“项目 I 没有市场前景”。作为企业的最终决策者,将如何根据这样的结论进行决策呢?

为什么分析同一个企业数据库中的数据,却得到截然相反的结论呢?

首先,两部门可能抽取数据的内容不同。比如,市场部抽取的是项目 I 在大