

高等学校计算机教材

数据库教程

施伯乐 丁宝康 杨卫东 编著



电子工业出版社
Publishing House of Electronics Industry
<http://www.phei.com.cn>

复旦名师又一力作

高等学校计算机教材

数据库教程

施伯乐 丁宝康 杨卫东 编著

电子工业出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书详细介绍数据库技术的基本原理、方法和应用技术。全书分成6个部分：概念篇、关系篇、设计篇、对象篇、分布篇和决策篇。内容包括：数据库基本概念、关系运算、SQL语言、数据库管理、规范化设计、ER模型、高级概念建模活动（UML类图）、对象关系数据库、面向对象数据库、分布式数据库、异构多数据源的访问、XML技术、数据仓库、OLAP技术和数据挖掘。

本书内容丰富，叙述流畅，概念严谨清楚，涵盖了目前数据库的实用理论和应用技术。每章后均有适量的习题。与本书配套，还有习题解答和上机指导书等出版物。

本书可作为高等院校计算机、信息技术等专业的数据库课程教材，也可供从事信息领域工作的科技人员和其他有关人员参阅。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

数据库教程 / 施伯乐，丁宝康，杨卫东编著. —北京：电子工业出版社，2004.9

ISBN 7-121-00229-9

I. 数… II. ①施…②丁…③杨… III. 数据库系统—教材 IV. TP311.13

中国版本图书馆CIP数据核字（2004）第081091号

责任编辑：吴金生 张月萍

排版制作：华信卓越公司制作部

印 刷：北京东光印刷厂

出版发行：电子工业出版社

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：26 字数：665千字

印 次：2004年9月第1次印刷

定 价：33.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系。联系电话：（010）68279077。质量投诉请发邮件至 zltz@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

前 言

数据库技术产生于20世纪60年代末。经过30多年的发展,数据库技术已形成较为完整的理论体系和一大批实用系统。数据库管理已经从一种专门的计算机应用发展成现代计算环境的一个核心部分。数据库应用领域也已从数据处理、信息管理及事务处理扩大到计算机辅助设计、人工智能、决策支持系统和网络应用等新的应用领域。数据库系统的推广使用使得计算机应用迅速渗透到国民经济的各个部门和社会的每一个角落,并改变着人们的工作方式和生活方式。

在20世纪末,关系数据库得到了广泛的应用,进入21世纪以后,对象数据库、网络数据库及新决策支持系统等技术将得到推广和普及。这四大板块已成为新时期数据库技术的主流。

为此,本书分为6个部分:

- **概念篇** 介绍数据库系统的基本概念、数据模型和全局结构。
- **关系篇** 介绍关系数据库的关系运算理论、SQL语言和数据库管理技术。
- **设计篇** 介绍关系数据库的模式设计理论、基于ER模型的数据库设计和基于UML类图的面向对象的高级概念建模。
- **对象篇** 介绍对象关系数据库和面向对象数据库的定义语言和查询语言。
- **分布篇** 介绍网络环境下的分布式数据库、异构多数据源的访问和半结构化数据的XML技术。
- **决策篇** 介绍组成新决策支持系统的数据仓库、OLAP技术和数据挖掘技术。

这些内容涵盖了目前数据库的实用理论和应用技术。

全书贯穿了数据库的4条线索:

- 数据库的发展从层次型、网状型、关系型、对象型、网络型到决策型的演变过程。
- 数据模型和数据语言的演变。
- 数据库技术的“用户”和“系统”两个端点。
- 数据库设计从实体联系图、对象联系图到UML类图的发展过程。

本书的主要目的是使读者掌握数据库技术的基本原理、方法和应用技术,能有效地使用现有的数据库管理系统和软件开发工具,掌握数据库结构的设计和数据库应用系统的开发原理,同时也适时地掌握对象数据库、网络数据库、数据仓库、OLAP技术及数据挖掘等新型数据库技术。

本书对某些较深的内容,在章节前加了符号“*”;在每一道例题或定理的结尾处,加了符号“□”,以示与正文的区别。

全书内容丰富新颖,具有理论性、实用性和可操作性。每章后均有适量的习题,以配合对知识点的掌握。教师讲授时应根据需要对内容做适当取舍。

本书可采取如下的教学方法:

- (1) 讲授的重点放在前三篇。对于理论性较强的内容(目录中加*的节),可做适当压缩。第

2 章的重点是关系代数,第 5 章的重点是函数依赖和范式。

(2) 对于后三篇的内容,教师可做适当引导,或布置学生自学,或扼要介绍。

(3) 上机环境可采用 PowerBuilder 9.0 或 SQL Server 等系统,使学生具有一定的上机操作能力,以巩固所学的知识。

为了满足大家对使用本书作为主教材的教和学的要求,我们还组织了习题解答和上机指导书,随之一起出版。

本书第 13 章由王预执笔,第 12 章和第 14 章由杨卫东执笔,其余章节由丁宝康执笔。最后由施伯乐、丁宝康和杨卫东一起对全稿进行了修改和定稿。参与本书编写的还有陈长洪、薛剑虹、王晓雯、徐美娟、许建军、严勇、曾宇昆、杨卫稼等老师。复旦大学教务处和信息学院给予热情的支持、鼓励和指导,使得本书得以顺利面世。在此一并表示衷心的感谢。

限于水平,书中欠妥之处,敬请广大读者和专家批评指正。对本书的意见请按地址 dn@citiz.net 反馈给我们,谢谢。

施伯乐 丁宝康 杨卫东
2004 年 5 月于复旦大学

目 录

第 1 部分 概念篇	1
第 1 章 数据库基本概念	1
1.1 数据管理技术的发展	1
1.1.1 人工管理阶段	1
1.1.2 文件系统阶段	2
1.1.3 数据库阶段	3
1.1.4 高级数据库阶段	5
1.1.5 数据库的基本术语	7
1.2 数据描述	8
1.2.1 概念设计中的数据描述	8
1.2.2 逻辑设计中的数据描述	9
1.2.3 存储介质层次及数据描述	9
1.2.4 数据联系的描述	11
1.3 数据抽象的级别	13
1.3.1 数据抽象的过程	13
1.3.2 概念模型	14
1.3.3 逻辑模型	15
1.3.4 外部模型	18
1.3.5 内部模型	19
1.3.6 高度的数据独立性	19
1.4 数据库管理系统	20
1.4.1 DBMS 的工作模式	20
1.4.2 DBMS 的主要功能	21
1.5 数据库系统	22
1.5.1 DBS 的组成	22
1.5.2 DBS 的全局结构	24
1.5.3 应用程序的演变	26
1.5.4 DBS 的效益	26
1.6 小结	27
习题 1	28
第 2 部分 关系篇	31
第 2 章 关系运算理论	31
2.1 关系模型的基本概念	31
2.1.1 基本术语	31
2.1.2 关系的定义和性质	32
2.1.3 三类完整性规则	32

2.1.4	关系模型的形式定义和优点	34
2.1.5	关系查询语言和关系运算	34
2.2	关系代数	35
2.2.1	关系代数的 5 个基本操作	35
2.2.2	关系代数的 4 个组合操作	36
2.2.3	关系代数运算的应用实例	39
2.2.4	关系代数的两个扩充操作	40
* 2.3	关系演算	42
2.3.1	元组关系演算	42
2.3.2	域关系演算	45
2.3.3	关系运算的安全约束和等价性	46
* 2.4	关系逻辑	47
2.4.1	关系逻辑的成分	48
2.4.2	从关系代数到关系逻辑的转换	49
2.4.3	递归过程	51
2.4.4	关系逻辑与关系代数的差异	52
2.5	关系代数表达式的优化	52
2.5.1	关系代数表达式的优化问题	53
2.5.2	关系代数表达式的等价变换规则	53
2.5.3	关系代数表达式的启发式优化算法	55
2.6	小结	58
	习题 2	59
第 3 章	SQL 语言	63
3.1	SQL 简介	63
3.1.1	SQL 发展史	63
3.1.2	SQL 数据库的体系结构	64
3.1.3	SQL 的组成	65
3.1.4	SQL 的特点	65
3.2	SQL 的数据定义	65
3.2.1	SQL 模式的创建和撤销	65
3.2.2	基本数据类型	66
3.2.3	基本表的创建、修改和撤销	67
3.2.4	索引的创建和撤销	69
3.3	SQL 的数据查询	70
3.3.1	SELECT 查询语句的基本结构	70
3.3.2	SELECT 语句的使用技术	71
3.3.3	聚合函数	74
3.3.4	SELECT 语句完整的句法	74

3.4	SQL 数据查询中的限制和规定	75
3.4.1	SELECT 语句中的规定	76
3.4.2	条件表达式中的比较操作	77
3.4.3	嵌套查询的改进写法	80
3.4.4	基本表的连接操作	81
* 3.4.5	SQL3 中的递归查询	82
3.5	数据更新	83
3.5.1	数据插入	83
3.5.2	数据删除	84
3.5.3	数据修改	85
3.6	视图	86
3.6.1	视图的创建和撤销	86
3.6.2	对视图的操作	87
3.7	嵌入式 SQL	87
3.7.1	嵌入式 SQL 的实现方式	88
3.7.2	嵌入式 SQL 的使用规定	88
3.7.3	嵌入式 SQL 的使用技术	90
3.7.4	动态 SQL 语句	93
3.8	小结	94
	习题 3	94
第 4 章	数据库管理	98
4.1	事务	98
4.1.1	事务的定义	98
4.1.2	事务的 ACID 性质	99
4.2	数据库的恢复	100
4.2.1	典型的恢复策略	100
4.2.2	故障类型和恢复方法	100
4.2.3	检查点技术	101
4.2.4	SQL 对事务的支持	102
4.3	数据库的并发控制	103
4.3.1	并发操作带来的 4 个问题	103
4.3.2	封锁技术	106
4.3.3	封锁带来的问题	108
4.3.4	并发操作的调度	110
4.3.5	SQL 对并发处理的支持	111
4.4	数据库的完整性	113
4.4.1	完整性子系统	113
4.4.2	SQL 中的完整性约束	113

4.4.3	约束可延迟性	117
* 4.4.4	SQL3 中的触发器	118
4.5	数据库的安全性	121
4.5.1	安全性问题	121
4.5.2	SQL 中的安全性机制	122
4.5.3	常用的安全性措施	124
4.6	小结	127
习题 4		128
第 3 部分	设计篇	131
第 5 章	关系模式设计理论	131
5.1	关系模式的设计准则	131
5.1.1	关系模式的冗余和异常问题	131
5.1.2	关系模式的非形式化设计准则	132
5.2	函数依赖	133
5.2.1	函数依赖的定义	133
5.2.2	FD 的闭包	135
5.2.3	FD 的推理规则	135
5.2.4	FD 和关键码的联系	136
5.2.5	属性集的闭包	137
5.2.6	FD 集的最小依赖集	137
5.3	关系模式的分解特性	138
5.3.1	关系模式的分解	138
5.3.2	无损分解	139
5.3.3	模式分解的优缺点	141
5.3.4	无损分解的测试方法	141
5.3.5	保持函数依赖的分解	143
5.3.6	模式分解与模式等价问题	144
5.4	范式	144
5.4.1	第一范式 (1NF)	144
5.4.2	第二范式 (2NF)	145
5.4.3	第三范式 (3NF)	146
5.4.4	BCNF (Boyce-Codd NF)	147
5.4.5	分解成 BCNF 模式集的分解算法	148
5.4.6	分解成 3NF 模式集的合成算法	148
5.4.7	模式设计方法小结	149
5.5	其他数据依赖和范式	149
5.5.1	多值依赖	149
5.5.2	关于 FD 和 MVD 的推理规则集	151

5.5.3	第四范式 (4NF)	152
5.5.4	嵌入多值依赖	152
5.5.5	连接依赖和第五范式	153
5.6	小结	154
	习题 5	155
第 6 章	基于 ER 模型的数据库设计	159
6.1	数据库工程与数据库系统生存期	159
6.1.1	规划阶段	159
6.1.2	需求分析阶段	160
6.1.3	概念设计阶段	160
6.1.4	逻辑设计阶段	162
6.1.5	物理设计阶段	163
6.1.6	数据库的实现	163
6.1.7	数据库的运行与维护	165
6.2	ER 模型	166
6.2.1	ER 模型的基本元素	166
6.2.2	属性的分类	168
6.2.3	联系的设计	171
6.2.4	ER 模型的操作	175
6.2.5	采用 ER 模型的数据库概念设计步骤	176
6.3	ER 模型到关系模型的转换	184
6.3.1	ER 图转换成关系模式集的算法	184
6.3.2	采用 ER 模型的逻辑设计步骤	186
6.4	ER 模型实例分析	188
6.4.1	库存管理信息系统的 ER 模型及转换	188
6.4.2	人事管理信息系统的 ER 模型	189
6.4.3	住院管理信息系统的 ER 模型	190
6.4.4	公司车队信息系统的 ER 模型	192
6.5	增强的 ER 模型	193
6.5.1	弱实体	193
6.5.2	子类实体与超类实体	194
6.6	小结	196
	习题 6	196
第 7 章	面向对象的高级概念建模	199
7.1	面向对象的数据类型系统	199
7.2	对象联系图	200
7.2.1	对象联系图的成分	200

7.2.2 数据的概化 / 特化	202
7.3 UML 类图	203
7.3.1 统一建模语言 (UML) 概述	203
7.3.2 用类图表达类和关联	203
7.3.3 用类图表达关联类	206
7.3.4 用类图表达概化 / 特化	207
7.3.5 用类图表达聚合	209
7.4 小结	209
习题 7	210
第 4 部分 对象篇	211
第 8 章 对象关系数据库 ORDB	211
8.1 关系模型的发展历程	211
8.1.1 从关系模型到后关系模型	211
8.1.2 从后关系模型到对象关系模型	213
8.2 ORDB 的定义语言	214
8.2.1 对象关系数据模型的定义	214
8.2.2 数据类型的定义	214
8.2.3 继承性的定义	215
8.2.4 引用类型的定义	217
8.2.5 SQL3 中的定义语言	219
8.3 ORDB 的查询语言	220
8.3.1 对 SELECT 语句的新规定	220
8.3.2 嵌套与解除嵌套	222
8.3.3 复合值的创建和查询	223
8.3.4 Oracle 中查询的两种技术	223
8.4 函数和过程	225
8.4.1 SQL 函数和过程	225
8.4.2 外部语言程序	226
8.4.3 过程的构造	226
8.5 小结	229
习题 8	229
第 9 章 面向对象数据库 OODB	231
9.1 OODBS 的基本概念	231
9.1.1 ODMG 标准	231
9.1.2 OODBS 的定义	232
9.1.3 OODB 的基本概念	232

9.2	ODMG 对象模型	236
9.2.1	对象和文字	237
9.2.2	接口、类和继承	238
9.2.3	类外延、关键码和工厂对象	239
9.3	ODMG ODL	240
9.4	ODMG OQL	241
9.4.1	OQL 中的 SELECT 语句	241
9.4.2	OQL 表达式的附加格式	243
9.4.3	OQL 中对象的赋值和建立	245
9.5	C++ 语言的绑定	246
9.6	OODB、ORDB 与 RDB 的比较	247
9.6.1	OODB 与 RDB 在概念设计上的区别	247
9.6.2	OODB 与 ORDB 的比较	248
9.7	小结	249
	习题 9	249
第 5 部分 分布篇		251
第 10 章 分布式数据库		251
10.1	DDBS 的定义和特点	251
10.1.1	从集中式、分散式到分布式	251
10.1.2	DDBS 的定义	252
10.1.3	DDBS 的特点	253
10.1.4	DDBS 的优缺点	254
10.1.5	DDBS 的分类	255
10.2	分布式数据存储	256
10.2.1	数据分片	256
10.2.2	数据分配	257
10.3	DDB 的体系结构	258
10.3.1	体系结构	259
10.3.2	分布透明性	260
10.4	DDBMS	262
10.4.1	DDBS 的组成	262
10.4.2	DDBMS 的功能	263
10.4.3	DDBMS 的组成	263
10.4.4	DDBMS 的同构性程度和局部自治性程度	263
10.4.5	FDBS 的异构性	265
10.4.6	FDBS 的 5 层模式结构	265
10.5	分布式查询处理	266
10.5.1	查询代价的估算方法	266

10.5.2	基于半连接的优化策略	267
10.5.3	基于连接的优化方法	269
10.6	分布式数据库中的并发控制和恢复技术	270
10.6.1	DDB 中的问题	270
10.6.2	基于数据项识别拷贝的分布式并发控制	270
10.6.3	基于投票方法的分布式并发控制	271
10.6.4	分布式恢复	272
10.7	小结	272
习题 10		273
第 11 章	异构多数据源的访问	275
11.1	中间件	275
11.1.1	中间件的定义	275
11.1.2	中间件的作用	276
11.2	ODBC 结构	276
11.2.1	ODBC 概念	276
11.2.2	ODBC 的体系结构	277
11.2.3	ODBC 的特性	282
11.3	ODBC 接口	282
11.3.1	ODBC 应用程序的基本流程	282
11.3.2	ODBC 句柄	282
11.3.3	数据源的连接与断开	285
11.3.4	SQL 语句的执行	286
11.3.5	查询结果的获取	287
11.4	ODBC 的符合性级别	288
11.4.1	API 符合性的三个级别	288
11.4.2	SQL 符合性的三个级别	289
11.4.3	ODBC API 与 SQL CLI 之间的协调	290
11.4.4	SQL CLI 与嵌入式 SQL 的比较	291
11.4.5	典型的数据库应用系统开发工具	291
11.5	JDBC 结构	292
11.5.1	JDBC 的提出	292
11.5.2	JDBC 的基本功能	292
11.5.3	JDBC 数据库设计方法	293
11.5.4	保持一致性的措施	294
11.5.5	JDBC 驱动程序	295
11.6	JDBC API	296
11.6.1	JDBC API 的目标	296
11.6.2	JDBC API 接口概貌	297

11.6.3	JDBC 的接口和类	299
11.6.4	JDBC 数据库应用程序的编写	300
11.7	小结	303
习题 11		303
第 12 章	XML 技术	305
12.1	XML 概述	305
12.1.1	XML 的诞生	305
12.1.2	XML 文档	306
12.1.3	文档类型定义 (DTD)	308
12.1.4	XML 模式	310
12.2	XML 编程接口	319
12.2.1	文档对象模型 (DOM)	319
12.2.2	简单的应用程序设计接口 (SAX)	323
12.3	常用的 XML 查询语言 XQuery	326
12.3.1	XQuery 的基本功能	326
12.3.2	XQuery 的基本概念	327
12.3.3	简单查询	328
12.3.4	各种类型的查询	330
12.4	小结	334
习题 12		334
第 6 部分	决策篇	337
第 13 章	数据仓库	337
13.1	DW 概述	337
13.1.1	从 DB 到 DW 的演变	337
13.1.2	DB 数据和 DW 数据的区别	338
13.1.3	DW 的定义和特点	338
13.1.4	DW 的类型	339
13.2	DW 的组织结构	340
13.2.1	DW 的数据组织结构	340
13.2.2	粒度与分割	341
13.2.3	DWS 的结构	341
13.2.4	DW 的运行结构	343
13.3	DW 存储的多维数据模型	343
13.3.1	多维立方体	343
13.3.2	星形模式	343
13.3.3	雪花模式	345
13.3.4	事实星座模式	346

13.4 DW 的数据获取与管理	347
13.4.1 DW 的数据获取	347
13.4.2 DW 的数据管理	348
13.5 DW 的设计和发展阶段	350
13.5.1 DW 设计的原则	350
13.5.2 DW 设计的步骤	351
13.5.3 DW 的发展阶段	353
13.6 小结	355
习题 13	355
第 14 章 联机分析处理技术	357
14.1 OLAP 概述	357
14.1.1 OLAP 的定义	357
14.1.2 OLAP 准则	358
14.1.3 OLAP 的基本概念	360
14.1.4 OLAP 与 OLTP 之间的比较	361
14.2 OLAP 的数据组织	362
14.2.1 MOLAP	362
14.2.2 ROLAP	362
14.2.3 HOLAP	363
14.2.4 OLAP 数据的处理方式	363
14.3 OLAP 的多维数据分析	364
14.3.1 切片和切块	364
14.3.2 钻取	366
14.3.3 旋转	367
14.3.4 OLAP 应用开发实例	368
14.3.5 广义 OLAP 操作	369
14.4 OLAP 的数据索引技术	370
14.4.1 位图索引	370
14.4.2 连接索引	371
14.5 基于 Web 的 OLAP 系统结构	372
14.6 小结	374
习题 14	375
第 15 章 数据挖掘	376
15.1 DM 概述	376
15.1.1 DM 的由来	376
15.1.2 DM 的定义	377
15.1.3 DM 与 DW 的联系与区别	377

15.1.4 DM 与 OLAP 的联系与区别	378
15.2 DM 过程	379
15.3 DM 的关联分析方法	381
15.3.1 DM 的分析方法	381
15.3.2 关联规则的定义	382
15.3.3 关联规则的分类	383
15.3.4 关联规则的挖掘算法	383
15.3.5 多层和多维关联规则的挖掘	384
15.3.6 关联规则价值衡量的方法	385
15.4 DM 的其他分析方法	386
15.4.1 序列模式分析方法	386
15.4.2 分类分析方法	387
15.4.3 聚类分析方法	389
15.5 DM 的应用领域	390
15.5.1 DM 的应用行业	390
15.5.2 商业化的 DM 工具	392
15.6 新决策支持系统概述	393
15.6.1 新 DSS 的结构图	393
15.6.2 新 DSS 的成功实例	394
15.6.3 新 DSS 与传统 DSS 的比较	395
15.6.4 综合 DSS 的结构图	397
15.7 小结	398
习题 15	399
参考文献	400

第 1 部分 概 念 篇

第 1 章 数据库基本概念

今天,“数据库”(Database, 简记为 DB)这个名词已是家喻户晓,成为社会生活中不可缺少的一部分。例如,我们可能会去银行取款,可能需要预定飞机票,可能网上购物等等。数据库的应用领域已遍及现代社会的每一个角落。

为了学习数据库的基本原理、方法和应用,我们必须从传统数据库技术的基本概念学起。为此,本章先回顾数据管理技术的发展过程,然后介绍数据库中的基本概念,以使读者对数据库概貌有个基本的了解。

1.1 数据管理技术的发展

使用计算机以后,数据处理的速度和规模无论是相对于手工方式还是相对于机械方式都是无可比拟的。随着数据处理量的增长,产生了数据管理技术。数据管理技术的发展,与计算机硬件(主要是外部存储器)、系统软件及计算机应用的范围有着密切的联系。数据管理技术的发展经历了人工管理、文件系统、数据库阶段和高级数据库阶段。下面分别介绍。

1.1.1 人工管理阶段

在这一阶段(20 世纪 50 年代中期以前),计算机主要用于科学计算,其他工作还没有展开。外部存储器只有磁带、卡片和纸带等,还没有磁盘等字节存取存储设备。软件只有汇编语言,尚无数据管理方面的软件。数据处理的方式基本上是批处理。这个时期的数据管理有下列特点:

① 数据不保存在计算机内。计算机主要用于计算,一般不需要长期保存数据。在进行某一课题计算时,将原始数据随程序一起输入内存,运算处理后将结果数据输出。随着计算任务的完成,用户作业退出计算机系统,数据空间随着程序空间一起被释放。

② 没有专用的软件对数据进行管理。每个应用程序都要包括存储结构、存取方法及输入输出方式等内容。程序中的存取子程序随着存储结构的改变而改变,因而数据与程序不具有独立性。存储结构改变时,应用程序必须改变。此时,由于程序直接面向存储结构,因此数据的逻辑结构与物理结构没有区别。

③ 只有程序(Program)的概念,没有文件(File)的概念。数据的组织方式必须由程序员自行设计与安排。

④ 数据面向程序。即一组数据对应一个程序。