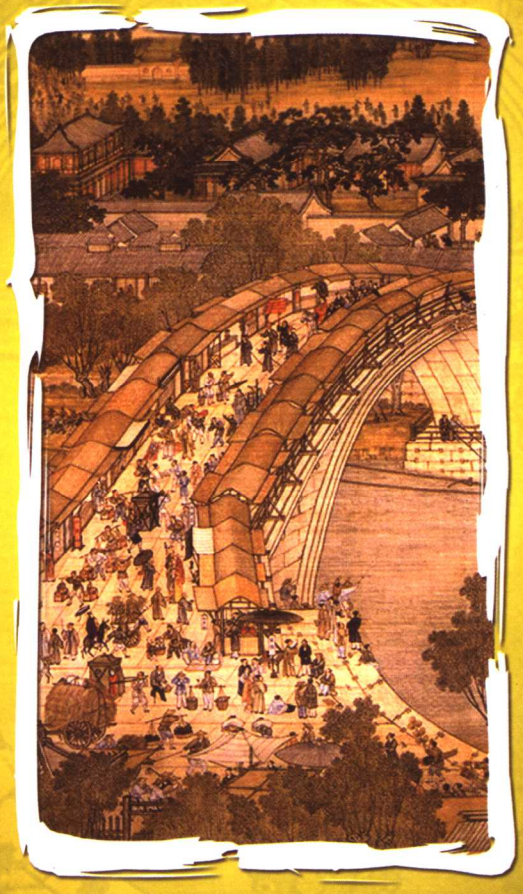
TURING 高等院校计算机教材系列

数据库原理与应用

徐保民　孙丽君　李爱萍　编著



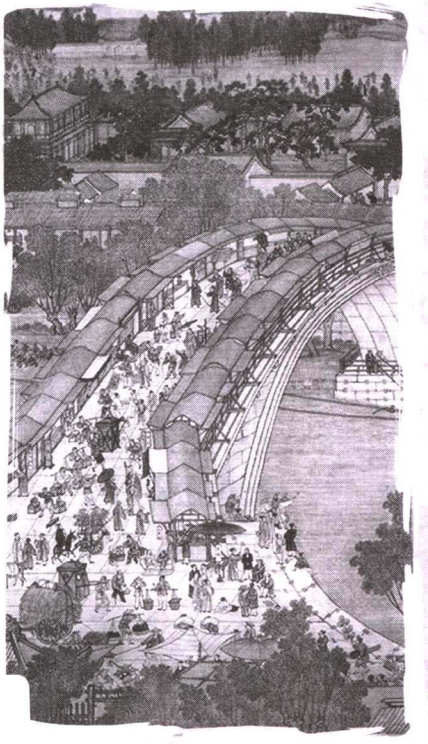
人民邮电出版社
POSTS & TELECOM PRESS

TP311.13/312
2008

TURING 高等院校计算机教材系列

数据库原理与应用

徐保民　孙丽君　李爱萍　编著



人民邮电出版社
北　京

图书在版编目（CIP）数据

数据库原理与应用 / 徐保民，孙丽君，李爱萍编著.
北京：人民邮电出版社，2008.1
（高等院校计算机教材系列）
ISBN 978-7-115-17072-9

Ⅰ. 数… Ⅱ. ①徐…②孙…③李… 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（2007）第 166942 号

内 容 提 要

本书系统全面地阐述了数据库的基本原理及应用。全书内容包括数据库系统概述、关系模型、关系数据库理论、SQL语言、数据库安全与保护、数据库设计、SQL Server 2005数据库管理系统和数据库应用系统开发等。本书内容丰富、语言通俗易懂，注重理论与实践相结合，讲求实用性和先进性。

本书可作为高等院校计算机或相关专业“数据库理论与应用技术”课程的教材，也可作为数据库应用编程人员的参考用书。

高等院校计算机教材系列

数据库原理与应用

◆ 编　　著　徐保民　孙丽君　李爱萍
责任编辑　杨海玲

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街 14 号
邮编　100061　　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市海波印务有限公司印刷
新华书店总店北京发行所经销

◆ 开本：800×1000　1/16
印张：18.75
字数：440 千字　　2008 年 1 月第 1 版
印数：1－4 000 册　　2008 年 1 月河北第 1 次印刷

ISBN 978-7-115-17072-9/TP

定价：32.00 元

读者服务热线：(010)88593802　印装质量热线：(010)67129223

前　言

数据库技术是计算机科学中一个非常重要的部分，它已成为计算机信息系统与应用系统的核心技术和重要基础。因此，作为计算机专业及相关专业的学生，学习和掌握数据库知识是非常必要的。

本书系统地讲述数据库系统的基本原理，剖析数据库技术的各个领域，详细描述数据库技术的重点和难点，力求做到概念清晰、内容全面和实用性强。

与常规的“数据库原理与应用”方面的书籍相比，本书的特色主要表现在第3章、第6章、第7章和第9章。第3章在关系代数基本内容的基础上，增加了“包上关系操作”；第6章中比较详实地介绍了数据库的存储结构，这部分内容在一般的数据库书籍中介绍得比较少；第7章按照软件工程的原理和方法介绍了数据库设计应该遵循的方法和步骤，并给出设计文档的写作标准，另外还讲述了应用UML技术进行数据库设计的方法和步骤，以及操作的具体方法；第9章介绍了通过ASP.NET技术在.NET环境下开发Web数据库的数据库应用实例，既包括前端界面和后端数据库之间的连接技术，也包括前端开发中存取后端数据库中数据的技术和方法。

全书包括9章和4个附录。

第1章概述了数据库的基本概念、数据库系统的体系结构以及数据库领域的新进展等内容。

第2章从集合论的观点出发，介绍关系数据库的基本理论，即关系数据结构、关系操作及元组关系演算等内容。

第3章从关系的角度出发，介绍关系数据库的数据模型、基本概念及常用的关系数据库对象。

第4章主要对关系数据库的标准语言SQL及其使用进行比较深入的讨论。

第5章主要讨论数据库规范化理论中的函数依赖、多值依赖及1NF到4NF的定义等内容。

第6章讨论数据的物理组织和索引技术，为用户开发应用系统时选择更适合于具体应用的存储和索引技术奠定了基础。

第7章重点讲述概念结构设计、逻辑结构设计和物理设计的方法和步骤，并简单阐述UML在数据库设计中的应用。

第8章讨论数据库的安全性、完整性、并发控制和恢复等数据库保护技术。

第9章介绍通过ASP.NET技术在.NET环境下开发Web数据库的数据库应用实例。

附录A、附录B、附录C和附录D分别简单介绍SQL Server 2005的安装步骤和常用工具的使用、Web数据库的基本概念、.NET框架和有关数据库访问技术，以及ASP.NET应用程序开发工具Web Matrix的安装和使用。

本书对数据库系统基本知识的讲解始终围绕着一个简单的“学生选课”系统逐步展开。同时，在第9章给出了一个综合利用前面各章内容的“图书管理系统”的实现实例，这充分体现了案例

教学的思路，特别适合学生自主学习。

参与本书编写的老师多年来一直从事数据库原理课程的教学，积累了丰富的教学经验，书中的很多内容都是他们教学经验的总结。本书第1章~第6章、附录A、附录B和附录D由徐保民编写，第7章~第9章和附录C由孙丽君编写。全书由李爱萍负责统稿。

本书的出版得到了人民邮电出版社图灵公司及陈贤舜老师的大力支持，北京交通大学计算机与信息技术学院的陈旭东、张宏勋等老师为本书的编写提出了不少的建议，在此一并表示深深的谢意。

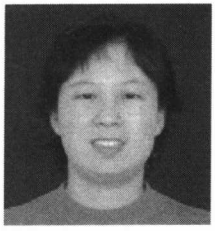
由于编者水平所限，书中难免会存在很多不足和错误，恳请各位读者不吝指正。

此为试读，需要完整PDF请访问：www.ertongbook.com

作 者 简 介

徐保民 1966年出生于河南叶县。2000年毕业于中国科学院计算技术研究所，获博士学位。现任北京交通大学计算机与信息技术学院副教授。2001～2002年在法国国家信息与自动化研究院（INRIA）从事博士后研究，2002～2003年丹麦科技大学(Technical University of Denmark)访问学者，2007～2008年美国中佛罗里达州立大学（University of Central Florida）访问学者。自2006年受邀任International Journal on Computer Science and Information Systems编委。以第一作者在国际、国内学术期刊和会议上发表论文60余篇，出版教材5部、专著1部，其中2005年出版的《数据库系统原理与应用》被评为2006年北京高等教育精品教材。目前的研究领域有生物信息处理、网格计算和计算机支持的协同工作。

孙丽君 1968年5月出生于河南省郑州市。现任河南工业大学电气工程学院副教授，工学博士，硕士生导师，河南省杰出青年科学基金获得者，河南省教育厅学术技术带头人。主要研究方向为信号与信息处理、盲信号处理、自适应技术、数字通信等。近年来以第一作者在国家一级期刊、外文期刊、核心期刊和IEEE重要国际会议上发表学术论文26篇，其中被EI收录7篇、ISTP收录2篇，主编和参编教材4部，先后主持或参与完成省级以上科研项目10余项，目前主持有省部级重大科研项目等2项。



李爱萍 太原理工大学计算机与软件学院软件系教师。1995年于中国矿业大学获得学士学位，2001年于太原理工大学获得硕士学位，2006年于西安电子科技大学获得博士学位。自1995年以来，一直在太原理工大学计算机学院从事教学工作，主要课程有Basic语言、C语言、面向对象程序设计与C++语言、（SQL Server）数据库原理及应用、软件工程导论、软件体系结构、软件质量工程等。目前的主要研究方向包括数据库应用、程序代码自动生成、软件体系结构等。

目　录

第1章 绪 论

数据库是一门专门研究如何科学地组织和存储数据、如何高效地获取和处理数据的技术。本章将从基本问题、特点等方面引出后续各章所涉及的主要知识点，具体内容包括数据库的基本概念、数据库系统的发展、数据库系统的结构及组成等。

1.1 数据库的基本概念

数据库（database，DB）、*数据库管理系统*（database management system，DBMS）和*数据库系统*（database system，DBS）是数据库技术中最基本、最常用的3个术语，它们之间既存在着一定的联系，也存在着一定的区别。

1.1.1 数据库

“数据库”这一术语有很多种解释。从字面上来看，就是存放数据的仓库。从本质上讲，数据库是指数据和数据对象的集合。这种集合可以长期存储，具有确定的数据存储结构，同时能以安全和可靠的方法进行数据的检索和存储。数据对象是指*表*（table）、*视图*（view）、*存储过程*（stored procedure）和*触发器*（trigger）等，这些数据对象将在以后的章节中介绍。

1.1.2 数据库管理系统

数据库管理系统（DBMS）是数据库系统中的一个系统软件，它允许用户对数据库中的数据进行操作，并将操作结果以某种格式返回给用户。从本质上讲，DBMS就是管理数据库中数据集合的系统软件。为了便于用户访问数据库，数据库通常都封装在DBMS中。

数据库访问操作可以归纳为定义、查询及更新（包括插入、删除或修改操作）3大类。目前，最常用的支持数据库访问操作的语言是*结构化查询语言*（structured query language，SQL）。尽管它被称为查询语言，但它也支持数据库定义和数据更新操作。利用SQL语言进行访问操作时，用户程序只需关注“要做什么”，“如何做”则由DBMS完成。

一般来讲，DBMS至少要提供如下功能。

- 允许用户使用专门的*数据定义语言*（data definition language，DDL）建立新的数据库，并说明它的逻辑结构，即*模式*（schema）。

- 允许用户使用专门的数据操纵语言（data manipulation language，DML）进行查询和更新操作。数据操纵语言可以分为两大类：一类嵌入在C、COBOL等高级语言中，这类数据操纵语言本身不能独立使用，因此称为宿主型数据操纵语言；另一类是交互式命令语言，它们的语法简单，且可以独立使用，故又称为自主型或自含型数据操纵语言。
- 支持大数据量的持久存储，并提供数据保护功能。
- 支持多用户对数据的并发存取，保证一个用户的操作不影响另一个用户的操作。

1.1.3 数据库系统和数据库应用系统

*数据库系统*是指一个采用数据库技术的计算机存储系统。广义地讲，数据库系统是由计算机硬件、操作系统、数据库管理系统以及在它支持下建立起来的数据库、应用程序、用户和维护人员组成的一个整体。狭义地讲，数据库系统由数据库、数据库管理系统和用户构成。由此可见，数据库系统的主要组成部分是数据库和数据库管理系统。

*数据库应用系统*是指为特定应用开发的数据库应用软件。它是对数据库中的数据进行处理和加工的软件，是面向特定应用的。例如，基于数据库的管理信息系统、决策支持系统等都属于数据库应用系统。

综上所述，数据库、数据库管理系统和数据库系统是3个不同的概念。*数据库*强调的是数据，*数据库管理系统*是系统软件，而*数据库系统*强调的是整个系统。数据库系统的目的在于维护信息，并在必要时提供协助来获取这些信息。另一方面，用户的目的是使用数据库，而数据库管理系统是帮助达到这一目的的工具和手段。

需要指出的是，人们常常将数据库作为数据库系统的同义词使用，将数据库系统作为数据库管理系统的同义词使用。

1.2 数据库管理系统的发展

数据库技术是20世纪60年代末发展起来的一种数据管理技术。它的出现标志着以应用程序自己管理数据、数据无法共享等为特点的传统手工管理数据阶段的结束。

1.2.1 早期数据库管理系统

最早的数据库管理系统出现于20世纪60年代末。它们都是采用文件系统对数据进行管理，即将数据保存在文件中。应用程序直接操作文件中的数据。显然，文件系统可以用于大数据量的长期存储。但是，它也存在一定的局限性，例如，它不支持多用户或进程对文件的并发访问；对数据定义功能的支持仅局限于文件目录结构的创建；尽管支持数据查询和更新操作，但没有专门的数据操纵语言等。

尽管基于文件系统实现的数据管理系统能够将数据单独组织成数据文件而存储在外存中，由文件系统统一管理，并保证应用程序能够方便地访问这些数据，但它仍不是一种理想的数据管理方式。其主要问题是：一个数据文件基本上对应于一个具体的应用程序，即数据文件之间缺乏必

此为试读，需要完整PDF请访问：www.ertongbook.com

要的联系；当不同的应用程序所需要的数据有部分重叠时，必须建立各自的数据文件，而不能共享具有相同部分的数据，更谈不上支持多用户对数据的并发访问。

20世纪60年代后期以来，多个应用程序相互覆盖地共享数据的要求越来越强烈。同时，在处理方式上，联机实时处理的要求也越来越多，并开始提出和考虑多用户并发访问共享数据的需求。在此背景下，出现了使用不同数据模型描述数据库中信息结构的数据库管理系统。主要的数据模型有图1-1所示的基于树结构的层次模型和图1-2所示的基于图结构的网状模型。

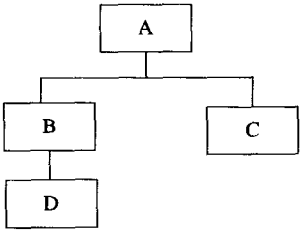


图1-1 层次模型示意图

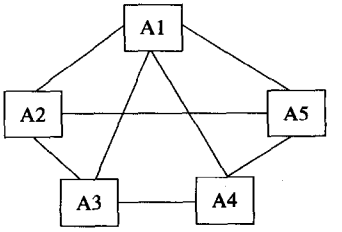


图1-2 网状模型示意图

层次模型由处于不同层次的结点组成，每一个结点为一个描述实体（实体就是现实世界中任何客观存在的事物）的记录类型。每个记录类型可包含若干个描述实体属性（属性是指实体所具有的某一方面的性质或特性）的字段。结点之间的连线表示记录类型间的联系。

网状模型将数据组织成图结构。结构中每个结点代表一个数据记录类型，每个记录类型可包含若干字段，结点间的连线描述不同结点间的联系。显然，网状模型是一种比层次模型更具普遍性的结构，它摆脱了层次模型所固有的一些限制，如允许一个结点有多个父结点、允许两个结点之间有多种联系等。

基于层次模型和网状模型的数据库管理系统的数据操纵语言是一次一条记录的过程化语言。所谓过程化语言是指用户不仅要了解“要做什么”，而且要指出“怎么做”。用户必须使用某种高级程序设计语言编写程序，一步一步地引导程序按照预先定义的存取路径来访问数据库，最终达到要访问的目标数据。在访问数据库时，每次只能存取一条记录。若该记录不满足要求，就沿着存取路径查找下一条记录。因此，即使是写一个简单的查询程序，用户也需花费很大气力。

总之，早期数据库管理系统的一个共同问题是它们不支持像SQL语言这样的高级查询语言。

1.2.2 关系数据库系统

1. 关系数据库的发展历程

1970年，IBM公司圣何塞研究中心的研究员E. F. Codd发表了著名的论文*A Relational Model of Data for Large Shared Data Banks*（大型共享数据库的关系数据模型），自此以后，数据库管理系统有了重大改进。Codd提出了关系数据模型的概念，即数据库管理系统应该将数据组织成二维表（也称为关系）的形式呈现给用户。使用关系数据模型的开发人员不必关心数据的存储结构，并可以使用高级语言来描述其查询，因此，可以极大地提高数据库应用系统开发人员的工作效率。

关系数据模型的主要特点如下。

- 关系模型的概念单一，实体以及实体之间的联系都用关系来表示。
- 以关系代数为基础，易于形式化表示。
- 数据独立性强，数据的物理存储和存取路径对用户隐蔽。
- 关系数据库语言是非过程化的，这样可以将用户从通过编程一步一步引导查询操作执行的过程中解脱出来，大大降低了用户编程的难度。

基于关系数据模型的数据库就是*关系数据库*（relational database，RDB）。它以表的形式将数据提供给用户，且所有的数据库操作都是利用已保存在数据库中的表来产生新的表。

关系数据库的发展历程可以分为3个阶段。

第一阶段从20世纪70年代初期E. F. Codd提出关系模型后开始，这一阶段奠定了关系模型的理论基础。人们研究了关系数据库语言，并开发出了关系数据库管理系统的一些原型。其中，IBM公司的System R和加州大学伯克利分校的Ingres等为这一时期的代表。

第二阶段从20世纪70年代后期开始，是关系数据库的应用阶段。这一时期从理论上解决了诸如查询优化、并发控制、完整性机制和故障恢复等一系列重大技术问题，从而使得关系数据库走向实用化、商业化。比较典型的商业关系数据库管理系统有Oracle、DB2和Informix等。

第三阶段从20世纪80年代开始，自那时以来，分布式关系数据库系统成为数据库研究的重点，并日趋成熟。目前，几乎所有主流的DBMS产品都支持分布式。这一时期的代表产品有Oracle、Sybase、Informix、DB2和SQL Server等。

虽然大多数商业DBMS已经开始提供面向对象的开发和应用，尤其是基于XML数据库的应用在不断增加，但它们仍然是基于关系模型的。

2. 关系数据库的基本概念

关系模型为人们提供了一种利用称为关系的二维表来描述数据的方法，关系模型的中心概念为关系。而一个关系由模式和模式的实例两部分构成。关系模型中常涉及以下概念。

- *关系实例*。关系实例就是指由行和列组成的表。通常人们仅用“关系”来代表关系实例。图1-3是一个学生的关系实例。

学生

学号	姓名	性别	年龄	所在系
001	张三	男	18	计算机
002	李四	男	20	计算机
003	王五	女	23	计算机

图1-3 关系实例

- *关系名*。每一个关系实例都有一个名称，称为关系名。图1-3的关系名为“学生”。
- *属性*。关系表中的列称为属性，其中第一行是属性名，其余各行的对应内容是属性值。图1-3中有学号、姓名、性别、年龄及所在系5个属性。属性由名称、类型和长度构成。不同的属性要赋予不同的属性名。属性的次序可以任意交换。

- 域。关系表中属性的取值范围称为域。在图1-3中，属性“性别”的域为“男”或“女”两个值。
- 元组。关系表中的行称为元组或记录。例如，图1-3中有3个元组。通常，任意两个元组不能完全相同。所有元组的集合就是关系表本身。
- 分量。元组中的每一个属性值称为元组的一个分量。例如，元组（001，张三，男，18，计算机）有5个分量，对应“年龄”属性的分量是18。同一属性的分量应是同一类型的数据，即来自同一个域，且每一个分量都必须是不可再分的数据项。
- 候选键。如果一个属性或若干属性的组合且该属性的组合中不包含多余的属性，能够唯一地标识一个关系的元组，则称该属性或若干属性的组合为候选键。一个关系可以有多个候选键。在最简单的情况下，候选键只包含一个属性。在极端的情况下，关系模式的所有属性构成关系模式的候选键，此时称为全键。
- 主键。当一个关系中有多个候选键时，可以从中选择一个候选键作为主键。一个关系上只能有一个主键。主键是能辨识记录的最小属性组。例如，图1-3中的学号可以作为主键。
- 主属性和非主属性。包含在任意一个候选键中的属性称为主属性，不包含在任意一个候选键中的属性称为非主属性。
- 关系模式。关系名和其属性集合的组合称为关系的模式。设关系名为R，其属性分别为A_1、A_2和A_3，则关系模式可以表示为R（A_1, A_2, A_3）。图1-3的关系模式为学生（学号，姓名，性别，年龄，所在系）。

关系模式仅仅是对数据特性的描述，因此可以将关系模式理解为一个数据类型，这样，关系实例就是一个具体值。

1.2.3 数据库系统的研究与发展

尽管到1990年，关系数据库管理系统已成为标准。但是，数据库领域仍在继续发展，关于数据管理的新课题、新方法不断涌现。

1. 并行数据库系统

随着数据量的大幅度增加，数据库的处理速度也要提高。加速的手段之一是为数据库增加索引结构；另一个就是采用分布/并行计算技术，与之对应的数据库系统主要有分布式数据库系统、并行数据库系统及网格数据库系统。其中，分布式数据库系统是数据库技术与网络技术结合的产物，目前已比较成熟。网格数据库系统是网格技术与数据库技术结合的产物，此概念刚刚提出，还没有大的进展。下面主要对并行数据库系统进行简单的介绍。

并行数据库系统的出现有其硬件和软件两方面的原因。硬件方面，主要是随着微处理器技术和磁盘阵列技术的进步，并行计算机得到了迅速发展，出现了商品化的并行计算机系统。软件方面，随着数据库规模的急剧膨胀，数据库服务器对大型数据库的各种复杂查询响应时间和联机事务处理（online transaction processing，OLTP）吞吐量的要求顾此失彼。数据库应用的发展对数据库性能和可用性提出了更高的要求，能否为用户提供低响应时间和高吞吐量已成为衡量DBMS性

能的重要指标。并行数据库系统就是在此背景下出现的。所谓并行数据库系统是指在并行机上运行的具有并行处理能力的数据库系统。它具有如下3个特点：

- 高性能。由于并行处理机通常有数十、数百甚至上千个处理器，因此它可以为用户提供一个具有高性能的数据库管理系统。
- 高可靠性。由于采用多处理器，当一个处理器的磁盘坝坏时，保存在其他磁盘上的数据副本仍可使用。
- 高扩展性。可以通过增加处理和存储能力使整个数据库系统的性能得到扩展。

2. 多媒体数据库系统

媒体是指信息的载体。多媒体是指多种媒体（如数字、字符、文本、图形、图像、语音和视频）的有机集成，而不是简单的组合。其中数字、字符等称为格式化数据，文本、图形、图像、语音、视频等称为非格式化数据，非格式化数据具有大数据量、处理复杂等特点。

多媒体数据的上述特性使DBMS向多个方向扩展，例如：

- 为了允许用户创建和使用复杂的数据操作，如图像处理，DBMS必须提供允许用户自己选择功能的能力。这类扩展通常要用到面向对象技术。
- 多媒体数据的大数据量迫使DBMS修改存储管理策略，以便能处理吉字节（gigabyte，1 GB是10^9B）、太字节（terabyte，1 TB是10^{12}B）甚至拍字节（petabyte，1 PB是10^{15}B）的数据对象，并解决查询结果在网上传输的问题。
- 提供更强的适合非格式化数据的查询功能，以便可以对图像、语音及视频等非格式化数据进行整体和部分搜索。

总之，多媒体数据库应实现对格式化和非格式化多媒体数据的存储、管理和查询功能。

3. 对象关系数据库系统

在关系数据库设计和应用领域，数据及其操作、表的划分以及用户权限等都符合面向对象的思想，因此许多生产关系数据库产品的厂商结合面向对象技术对其产品进行了改进。人们提出了一种折中方案，即基于扩展的关系数据模型的对象关系数据库。对象关系数据库系统兼有关系数据库和面向对象数据库两方面的特征，即除了具有一般关系数据库的各种特点外，还具有如下特点。

- 允许用户扩充基本数据类型。允许用户根据应用需求自己定义数据类型、函数和操作符；而且一经定义，这些新的数据类型、函数和操作符将存放在数据库管理系统的核心中，供所有用户共享。
- 能够在SQL中支持复杂对象。SQL语言应能支持由多种基本类型或用户定义的类型构成的对象。
- 能够支持多种继承。应支持子类对超类的各种特性的继承，支持数据继承和函数继承，支持多重继承和函数重载等。
- 能够提供功能强大的通用规则系统。规则系统应与其他对象-关系系统所提供的功能集成为一体。例如，规则中的事件和动作可以是任意的SQL语句，可以使用用户自定义的函数，规则能够被继承，等等。

4．数据仓库及数据挖掘技术

数据仓库（data warehouse，DW）是近年来信息领域中迅速发展起来的数据库新技术。目前，数据仓库一词尚没有一个统一的定义。著名的数据仓库专家W. H. Inmon在其著作*Building the Data Warehouse*（中译本《数据仓库》，机械工业出版社）中给予了如下描述：数据仓库是一个面向主题、集成、相对稳定且反映历史变化的数据集合，用于支持管理决策。

对于这个定义，可以从两个层次理解。首先，数据仓库用于支持决策，面向分析型数据处理，因此，不同于操作型数据库；其次，数据仓库是对多个异构数据源的有效集成，集成后按照主题进行重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

根据上述对数据仓库概念的描述，可以看出数据仓库有如下4个特点。

（1）数据仓库是面向主题的。操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织的。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面。一个主题通常与多个操作型信息系统相关。比如，一个保险公司的数据仓库所组织的主题可能为：客户、政策、保险金和索赔。而按应用来组织，则可能是：汽车保险、生命保险、健康保险和伤亡保险。可以看出，基于主题组织的数据被划分为各自独立的领域，每个领域有自己的逻辑内涵而互不交叉。而基于应用的数据组织则完全不同，它的数据只是为处理具体应用而组织在一起。应用是客观世界既定的，它对于数据内容的划分未必适用于分析所需。

（2）数据仓库是集成的。面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的。在将数据集成进数据仓库之前，需要对原有分散在各个数据库中的数据进行系统加工、汇总和整理。不但要统一原始数据中的不一致之处，如字段的同名异义、异名同义、单位不统一及字长不一致等，还要将原始数据结构从面向应用转变为面向主题。

（3）数据仓库是稳定的。操作型数据库中的数据会经常更新。数据仓库中的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，很少进行修改和更新操作。因此，一旦某个数据进入数据仓库，一般情况下将被长期保留。

（4）数据仓库是随时间变化的。操作型数据库主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含历史信息，系统地记录了从过去某一时点到目前各个阶段的信息。通过这些信息，可以对以往的数据和未来趋势做出定量分析和预测。

数据仓库的建设，是以现有的业务系统和大量业务数据的积累为基础的。数据仓库不是静态的概念，只有把数据仓库中记录的信息及时交给需要这些信息的使用者，供他们做出改善其业务经营的决策，信息才能发挥作用，信息才有意义。而把信息加以整理归纳和重组，并及时提供给相应的管理决策人员，是数据仓库的根本任务。因此，从产业界的角度看，数据仓库建设是一个工程，是一个过程。

数据仓库系统通常具有4层体系结构，如图1-4所示。

（1）数据源。数据源是数据仓库系统的数据源泉，通常包括企业内部信息和外部信息。内部信息包括存放于RDBMS中的各种业务处理数据和各类文件数据。外部信息包括各类法律法规、市场信息和竞争对手的信息等。

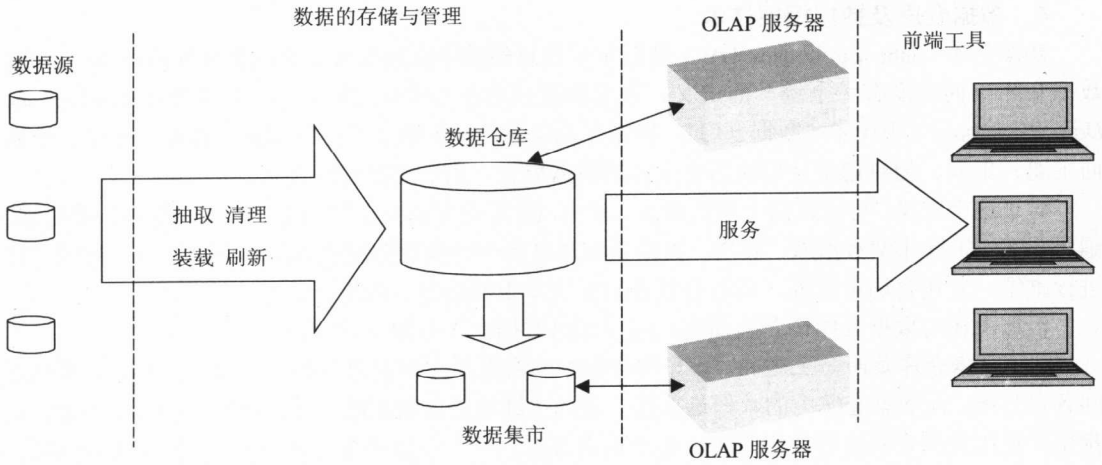


图1-4 数据仓库系统体系结构

（2）*数据的存储与管理*。数据的存储与管理是整个数据仓库系统的核心。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外的表现形式。按照数据的覆盖范围，数据仓库可以分为企业级数据仓库和部门级数据仓库。

（3）OLAP（online analytical processing，*联机分析处理*）*服务器*。OLAP服务器对需要分析的数据进行有效集成，按多维模型予以组织，并进行多角度、多层次的分析，以便发现潜在的知识。

（4）*前端工具*。前端工具主要包括各种报表处理、查询和数据分析（包括数据挖掘）等工具，以及各种基于数据仓库或数据集市的应用开发工具。其中数据分析主要针对OLAP服务器，报表处理和数据挖掘主要针对数据仓库。常用的分析技术有OLAP和数据挖掘。

- OLAP。OLAP技术发展迅速，产品越来越丰富。它们具有灵活的分析功能、直观的数据操作和可视化的分析结果表示等优点，从而使用户对基于大数据量的复杂分析变得轻松而高效。目前，OLAP工具可分为两大类，一类是基于多维数据库的，另一类是基于关系数据库的。两者相同之处是基本数据源仍是数据库和数据仓库，是基于关系数据模型的，向用户呈现的也都是多维数据视图。不同之处是前者把分析所需的数据从数据仓库中抽取出来，并物理地组织成多维数据库；后者则利用关系表来模拟多维数据，并不物理地生成多维数据库。
- 数据挖掘（data mining，DM）。数据挖掘是从大型数据库或数据仓库中发现并提取隐藏在内的信息的一种新技术。目的是帮助决策者寻找数据间潜在的关联，发现被忽略的要素，这些信息对预测趋势和决策行为也许是十分有用的。

可以看出，在OLAP中，用户发出请求的目的是获得特殊的信息，而数据挖掘是试图从存储于数据仓库的数据中提取新的信息。要描述数据挖掘中的查询条件，通常需要相关的数学、人工智能及机器学习等领域内的知识。

此为试读，需要完整PDF请访问：www.ertongbook.com