高等学校通识教育系列教材

高级数据库基础教程

叶小平 李强 陈瑛 叶晟 编著

清华大学出版社

高等学校通识教育系列教材

高级数据库基础教程

叶小平 李强 陈瑛 叶晟 编著

清华大学出版社
北京

内 容 简 介

本书是“高级数据库”课程的基础教材。全书突出计算机数据管理技术的逻辑主线，在编写思路和材料组织上具有体现整体架构、注重相互关联、彰示关键细节和强化实例讲解等特点。书中有选择性地介绍从经典数据库到当今的若干新型数据库的基本原理和相关技术，力求实现由经典关系数据库到新一代数据库技术的有效对接，并有助于完成由数据库理论技术的学习到从事相关研究探讨的基本过渡。

全书共10章，内容包括绪论、关系数据库基础、面向对象数据库和对象关系数据库、空间数据库与时态数据库、XML数据库、移动对象数据库、大数据技术简述、时态数据索引技术。

本书可作为高等院校计算机科学与技术及其相关专业本科生选修课和研究生必修课教材，同时由于主线突出和内容简明，也适合于相关人员作为自学材料参考使用。

图书在版编目(CIP)数据

高级数据库基础教程/叶小平等编著. —北京：清华大学出版社，2018
(高等学校通识教育系列教材)
ISBN 978-7-302-50165-7

Ⅰ. ①高… Ⅱ. ①叶… Ⅲ. ①数据库系统－高等学校－教材 Ⅳ. ①TP311.13

中国版本图书馆CIP数据核字(2018)第112402号

责任编辑：刘向威　王冰飞
封面设计：文　静
责任校对：焦丽丽
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：22.5　　**字　　数**：546千字
版　　次：2018年10月第1版　　**印　　次**：2018年10月第1次印刷
印　　数：1～1500
定　　价：59.00元

产品编号：069648-01

前言

计算机处理的对象是数据，数据可以是以数字符号为基础的数值型数据和由字符、音频、视频乃至网页等组成的非数值型数据。作为一个学科，计算机科学技术主要应用于非数值型数据。在众多的非数值型数据的应用中，有一类称为数据密集型的计算机应用领域，其基本特征是涉及数据体量巨大、数据计算结果需要长久驻留计算机及数据保持大范围共享等。这是迄今为止最大的计算机应用领域，因为任何规模化的计算机信息管理系统都需要以其为底层技术支撑，这就是数据库技术。

计算机应用从技术实现角度来看，可以分为数据计算和数据管理。数据库技术属于数据管理的技术范畴。自 20 世纪 50 年代中期开始至今，数据管理经历了人工管理（基于应用程序）、文件系统管理（基于操作系统）和数据库管理（基于 DBMS）三段历史进程；而自 20 世纪 60 年代中后期开始至今，数据库自身也经历了由第一代数据库（层次和网状数据库）到第二代数据库（关系数据库）再到第三代数据库（以对象数据库为典型代表的各类新型数据库）的 3 个发展阶段。如果将关系数据库等看作是经典数据库，则通常可以将包括对象数据库在内及其之后出现的各类数据库通称为高级数据库或现代数据库。这些数据库或者基于数据模型的创新，或者出于应用维度的扩展，或者与计算机各类新鲜的主流技术密切结合，它们共同构成当今兴旺发达的整个数据库家族。这个数据库家族相当庞大，以致需要从各个不同角度进行审视和不同层面展开讨论才有可能理解与把握。本书主要是从数据模型及建立在其上的数据操作原理视角进行内容组织和展开叙述的。

本书共分为 10 章。第 1 章是绪论，简述数据管理的产生背景、技术路线，以及数据库技术在整个计算机科学技术领域中的地位和意义。第 2 章是关系数据库基础，由于高级数据库中许多概念、原理和技术都与关系数据库有着密切关联，深刻地认识和把握关系数据库技术对于其他数据库技术的学习与研究是不可或缺的。第 3 章和第 4 章讨论对象数据库技术，其中第 3 章的面向对象数据库是基于全新数据模型的数据库，可以看作是 C++基于数据库机制的扩充；第 4 章的对象关系数据库的数据模型基础仍然是关系模型，可以看作是 SQL 关于面向对象原理方法的扩充。第 5 章和第 6 章分别是空间数据库和时态数据库，可以看作是由于数据库应用领域扩大和应用层面深化而驱动数据库在空间和时间应用维度方面的扩展，这种扩展可能被局限于关系数据模型（如时态关系数据模型），也可能需要建立新的数据模型（如空间数据的镶嵌数据模型和矢量数据模型）。第 7 章是 XML 数据库，这是一种半结构化数据的管理技术，有别于结构化的关系数据与对象数据，通常需要建立反映出“数据与结构融合”自身特点的更为复杂的数据模型。第 8 章是移动对象数据库，这是为了适应由于网络技术发展和移动通信终端设备普及而带来的新的数据管理需求。第 9 章是大数据技术简述，从数据管理角度考虑，可以看作是一种范围更为广阔和内容更为新颖的分布

式数据管理技术，其中 NoSQL 实际上是将常规的数据库技术推向了一个新的阶段。第 10 章是时态数据索引技术。各类新型数据库通常都没有成熟 DBMS 支持，数据存取的有效途径多是基于研究和开发相应的数据索引，因此，数据索引也就成为高级数据库技术的基本内容之一。本章应用时态数据库相关知识建立了时态数据索引框架，并将其应用到 XML 数据索引和移动对象数据索引，这实际上可看作是本书主要内容的一个综合应用。

本书的第 1 章、第 3～7 章和第 10 章主要由叶小平编写，第 2 章主要由李强编写，第 8 章主要由叶晟编写，第 9 章主要由陈瑛编写，同时，陈瑛参与了第 5 章和第 10 章的编写，李强参与了第 9 章的编写，叶晟参与了第 7 章的编写。全书由叶小平负责统筹。在本书编写过程中得到汤庸教授的热情鼓励和大力支持，其中不少观点的提出和材料的选择都得到了汤庸教授的启示和帮助，在此谨致以衷心感谢！同时，书中参考和借鉴了较多的数据库方面相关专著、经典教材和科研论文，书中每章之后附有其中的主要参考文献，然而难以一一列举，作者对此表示歉意。由于本书涉及的许多内容已经成为经典，不少专著和教材中对其都有专项论述，加之本书性质定位所限，因而大多没有列举原始的文献资料，这里谨对本书涉及的相关书目和文献的专家学者们表示诚挚的谢意！

此外，林衍崇和陈钊滢等研究生也参与了本书部分内容的讨论和完成。

由于高级数据库领域范畴广而深邃，相关技术日新月异，即使本书所论及部分也难免挂一漏万，失之偏颇。加上编著者学识和经验所限，疏漏与不当之处在所难免，恭待专家学者和教学同行批评指正。

2018 年 5 月于

中山大学海滨红楼及广东东软学院 Lisp 园

目 录

第1章 绪 论

数据库可以看作是基于数据管理的计算机技术系统，一般而言，它由一组相互关联的数据集合和一组用于访问及操纵数据的计算机程序组成。数据库技术是计算机学科中发展最快和应用最广的重要领域之一。由于任何信息管理系统都需要有数据库的后台支持，因此数据库应用现在已经成为人们社会经济活动中不可或缺的核心技术支撑。同时数据库还通过互联网融入人们日常生活的方方面面，如 ATM、网上购物、浏览信息和社交网络等。正是由于数据库技术在各类计算机应用中占有很大比重，专注数据库技术研制开发的 Oracle 早已成为当今最大的计算机软件公司之一，而其他具有重大影响力的计算机巨头，如微软和 IBM 也都以相应的数据库管理系统为其主打的支柱产品。本章简要介绍数据库技术的发展及其在整个计算机学科领域中的地位和意义。

1.1 数据及其特性

在计算机信息时代，“数据”是广泛使用的一个术语，但越是使用广泛的通常也是越难以明确定义的，因为它可能是所有相关概念的“源头”，或者说是元概念，抑或根本就无法进行严格定义和准确描述。“不幸”的是，“数据”与“信息”一样，就是这样的元概念。没有明确定义而又应用极其广泛，这一有趣现象事实上是普遍存在的，如“生命”“人”“智慧”“能量”和“质量”等都是如此。这类“伞形”概念实际上需要从其最常见、最有用特征和与其他相关概念最基本联系等方面进行适当描述和把握，从而以其为基础建立起庞大适用的体系。对于“数据”而言，人们只能如此处理。

1.1.1 数据概念

数据(data)一词来自拉丁文“to give”，意为“给”或“供给”。由此引申，数据可以看作是确定的事实，并且能从中推断出新的事实。

为什么会有数据？人之所以能够从一般灵长类动物中脱颖而出，就是因为逐步进化出能够描述、认识和利用客观事物和现象的基本能力。随着人类文明的不断进步，人们意识到仅仅使用一般的语言文字和图形图像描述他们所处的这个世界是不够精确的，这种描述对于发展科学技术乃至推动人类社会不断前行更是远远不够的。为了准确描述客观世界(如科学技术所必需的各种测量等)，也为了有效地展开社会经济活动(如货币使用和贸易交换等)，更为了充分改造和利用自然(如按照科学规律设计、建造机器和建筑等)，人们还需要“数据”这种特定的信息表述形式，并进行彼此间的交互。从本质上来看，人类的一切生产、交换等社会活动都可以说是以“数据”为基础而有效展开的，数据的出现和使用，应当是人类

文明的重大进步之一。

鉴于“数据”概念的重要性和基础性，通常需要从下述不同的角度来理解和掌握。

1. 从数据表现形式上考虑

从本源上考虑，数据是客观事物某种特征在人们意识中的反映，因此具有特定的表示形式。

(1) 广义数据：描述客观实体特征的各种实体或符号记录。例如，远古人类的小棍计数、结绳刻痕记事等以具体实物形式表示的数据；文明社会中以语言文字、声音图形和各类数字等具有不同抽象层级的符号形式对事物特征或数量上进行的描述等。

(2) 狭义数据：能够通过数字化编码进入计算机并由计算机进行处理的抽象符号集合。在当今的信息时代，人们通常从这种狭义角度理解和界定“数据”概念。

2. 从数据基本来源上考虑

按照数据的来源区分，数据可以有下述几种形式。

(1) 测量型数据：如上所述，数据首先源于人们认识和改造客观世界所必需的“直接”测量。作为“有根据的数字”，数据指的就是对客观世界测量结果的表述。测量是人类进行各种活动中不可或缺的基本手段，更是科学技术的必备基础。没有测量，就不会有数据；离开了数据，任何科学技术都会成为无本之木和无源之水。

(2) 计算型数据：数据可以作为测量结果直接使用，还可以将已有数据通过数据处理后得到新的数据，这是数据本身含义的体现，也是人们使用数据进程中的一个重大进步，因为有些数据根本不能通过直接测量获得，而只能通过对已有数据进行计算处理后而得到，如到太阳的距离(约 1.5 亿千米)和太阳内部的温度(2000 万摄氏度)等。这样就有了“原始数据”和“非原始数据”之分。

(3) 记录型数据：测量只是涉及客观世界中的事物，是数据最早的来源。随着人类科学文化技术的向前发展，记录极大地扩展了人类社会活动的深度和广度，为了丰富社会文化生活和保障文明传承，需要通过文字、图形图像、音频、视频和多媒体等记录人们自身的各类活动。在信息时代，这些记录大多需要借助于计算机系统进行存储、处理和管理，都需要转化为计算机意义下的数据。这样，数据就有了“测量”“计算处理”之外的第三个来源：由文本文字、图形图像、音频视频和多媒体等组成的“记录”数据。

3. 从数据、信息和知识关系上考虑

数据与信息一样都是元概念，难以进行严格逻辑意义上的定义。但从计算机应用角度来说，可对“数据”“信息”和“知识”三者关系进行描述，这种描述有益于对数据概念的理解和把握，在实际应用中也是行之有效的。

(1) 数据：通常可以描述为事实或观察的结果，作为对客观事物或其特征的某种形式上的归纳，主要用作未经加工的原始素材。数据的一个基本特征是在使用时需置于具体场景之中表明其语义。例如，“37”这个数据并没有表示任何意义(即语义)。但将它置于人体温度语境中，就表明了一个人的体温是 37℃；而将其置于人的年岁语境中，就说明一个人的年龄是 37 岁等。也就是说，数据需要解释语义，不能解释或没有语义的数据就没有使用价值。

(2) 信息：通常可以看作是具有明确语义的数据或数据整合体，信息会“明确”告知人们一定的含义，但不能保证该含义是否合适与正确。

(3) 知识：通常可看作是经过人类的归纳、整理和加工，最终呈现某种规律的正确性信息。

数据、信息和知识在递进的链条上可以看作：在内涵上一个比一个明确有力，在表现上一个比一个丰富多彩，但归根结底，数据是这一切的基础。

4. 计算机程序和数据

经过多年的探讨和实践，人们认识到计算机科学与技术的主体是其中的软件原理研究、方法设计与技术开发。对于计算机软件而言，程序和数据是两个最重要的组成部分。因此，从某种考量出发可以认为，计算机软件正是由于其中的程序和数据才得以构成了真正意义上的计算机运行实体。

实际上，对于计算机软件来说，程序和数据通常是相互关联与密切整合的，但在实际应用中却有孰重孰轻和谁主谁次的考虑。为了讨论此项问题，需要先从不同角度对计算机数据进行适当的分类。

(1) 数值型数据和非数值型数据。如前所述的整数、实数等基于测量和计算的数据就是数值型数据，其特点是可以通过转化为二进制数而“直接”进入到计算机并为计算机程序所处理。主要用于记录的字符、图形图像、声频视频及多媒体等数据都是非数值型数据，其特点是需要经过适当的编码方可进入计算机并为应用程序所处理。如今，非数值型数据已经成为所有计算机数据的主体组成。

(2) 挥发性数据和持久性数据。从是否长期驻留计算机来看，可以将数据分为挥发性(transient)数据和持久性(persistent)数据。显然，存在于内存中且当相应程序结束就被“析构”的数据是挥发性的，而相应程序结束后会被“自动”建构存储在外存中的数据就是持久性的。

(3) 私有性数据和共享性数据。从是否为多个程序共享同用来看，可以将数据划分为私有性(private)数据和共享性(share)数据。只能在个别特定程序中使用和处理的数据是私有性数据，能够被多个不同程序共同使用的数据则是共享性数据。显然，使用同一数据的应用程序越多，相应数据的共享程度就越高。

① 在直接使用程序设计语言解决实际问题的计算机应用过程中，程序是主体，数据是从属于特定应用程序的，此时的数据多是数值型数据，通常具有挥发性和私有性。

② 在各类涉及信息存储和管理的软件系统中，数据是主体，程序是围绕和服从于相关数据的，此时的数据大多是非数值型数据，通常具有持久性和共享性。

实际上，根据软件系统中程序是主体还是数据是主体，可以认为各类众多的计算机应用由“数据处理”和“数据管理”两大部分组成。

1.1.2 数据处理和数据管理

计算机的英文为“computer”，其原始含义是“计算工匠”。在最初时期，计算机应用的对象是“数”，此时“数据”就是“以数字形式表现出来的客观事物的特征证据”。这很自然，因为任何数字都可以“直接”转换为二进制数字，而数字计算机就是基于二进制数字的存储处理装置。此时如同工具是手的延伸一样，“computer”是人类大脑“计数”智力的延伸。ASCII码标准出现是数字处理技术中的划时代事件，它使得起源于数字“运算”的计算机技术能够应用到字符文本的处理。从此，就有了“数值型”数据和“非数值型”数据的技术之分。人类

思维需要借助语言来实现，而字符就是语言的载体，计算机应用进入由文字字符为代表的非数值型数据领域，为计算机具有真正意义上的“人脑智能”提供了可能，打开了计算机实现真正意义上的人脑“延伸”通道，此时，计算机才可以名副其实地称为“电脑”。

人类大脑的功能实际可以分为两个方面：一个是智慧，即处理问题的能力；另一个是记忆，即传承知识的能力。从数据角度考虑，计算机作为“电脑”，其智力也突出表现在数据处理（即数据计算）能力和数据管理（即数据存储（记忆数据）检索（记忆数据的使用））能力上。

1. 数据处理

数据处理的操作通常可以看作是通过对已有数据进行“计算”或“运算”以获取新的有用数据。这些运算可以是加减乘除等算术运算和“或”“与”等逻辑运算，也可以是更为复杂的计算机意义上的算法运算，如排序、查找和索引等。这方面内容集中体现在“数据结构与算法”课程当中，同时也普遍分布在计算机的各个领域与技术实现当中。数据处理计算具有下述特点。

（1）算法复杂性。算法内容复杂深入，算法设计灵活多变，但计算涉及的应用范围都有相对窄小的边界。

（2）基于程序设计语言。通常都需要借助某一种高级程序设计语言实现相应的数据处理。

（3）数据量相对较小。计算数据多是基于键盘输入，因此计算过程中涉及数据量相对较小。

（4）数据的挥发性和私有性。数据没有长时间存留和大范围多程序共享的一般需求。

2. 数据管理

数据处理计算是计算机最重要的应用之一，可以看作是一种“CPU 密集型”（CPU intensive）应用，另一类更为广阔的被称为“数据密集型”（data intensive）的应用领域就是数据管理。数据管理着眼于数据的持久存储、高效查询和大范围共享互用等，因此具有下述突出特征。

（1）数据量巨大。巨大的数据量需要存储在外存储器当中，在计算机运行过程中内存只能装载其中很小的一部分数据。

（2）数据持久性。与数据计算处理不同，管理过程中涉及的数据需要长期驻留计算机系统。

（3）数据共享性。系统管理的数据为众多应用程序或应用单位等大范围共享。

数据管理具体涉及数据收集整理、组织存储、维护传送和查询检索等数据操作，包括管理信息系统、办公室自动化系统、人事管理系统、酒店预订管理系统和金融信息系统等方面，实际上已经形成迄今为止最大的计算机应用系统。自从计算机由主要从事数值型数据的科学计算转变到从事更为广泛的非数值型数据应用以来，数据管理就已在计算机科学技术领域占据重要的核心地位。

1.1.3 数据管理和数据库

现在，整个计算机科学技术实际上几乎都以非数值型数据为基本应用对象，而其中非数值型数据管理已经成为最大的一类计算机应用领域。当一个计算机软件系统具有了数据共

享、数据独立乃至最重要的数据模型时，就可以看作是具有数据管理系统的基本特征。由此通常认为，自计算机科学技术诞生发展以来，数据管理技术经历了人工管理（应用程序管理）、文件系统管理（操作系统管理）和数据库管理（专用 DBMS 管理）三段历史进程。

1. 人工管理

人工管理实际上就是人们通过编写应用程序进行数据管理，其基本特点是一组数据对应一个特定应用程序，当多个不同程序使用同一数据集时，需分别设计数据结构，无法自动关联和相互参照，需要人工进行干预处置，因此也称为基于程序的数据管理。由此会导致下述问题。

（1）数据共享性不足。同一数据在不同程序中需要各自设计逻辑与物理结构及相应存取方式，难以进行有效的数据共享。

（2）管理工作重复进行。数据使用过程中出现大量冗余，从而导致需要对冗余数据进行重复管理，共享性品质较差。

（3）数据独立性差。数据的逻辑结构和物理结构交叠影响，使得数据与程序关系密切，当数据本身发生改变时，相应管理程序必须改变，数据缺乏基本的独立性。

基于程序的数据管理主要出现在 20 世纪 50 年代中期之前，当时没有磁盘等可直接存取的必要设备和操作系统支持等技术条件，因此应用程序（即人工方式）也只能是当时对数据进行管理的唯一可行办法。

人工管理方式如图 1-1 所示。

2. 文件系统管理

基于文件的数据管理主要出现在 20 世纪 50 年代末期到 20 世纪 60 年代中期，实际上就是使用操作系统中专门的文件系统完成相关工作。文件系统管理具有下述特点。

（1）数据长期驻留。计算机磁盘和磁鼓等提供了长期保存数据的硬件条件，文件系统提供了数据在外部存储器多次进行查询和更新的软件环境。

（2）一定程度数据独立性。应用程序与数据之间由文件系统提供的存取方法进行转换，数据与程序之间具有一定独立性。

（3）一定程度的数据共享性。数据按照内容、结构和用途组成文件，而文件面向应用，可以为一组使用同一数据的应用程序所共同使用。但当不同应用只具有部分相同数据时则需要建立不同的数据文件，此时又回到了程序管理的情形。

基于文件系统的数据管理如图 1-2 所示。

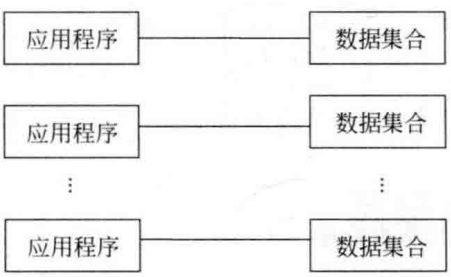

图 1-1　应用程序（人工）数据管理

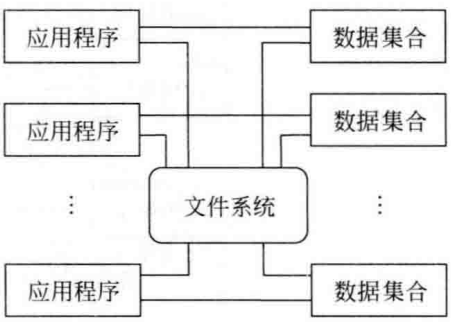

图 1-2　基于文件系统的数据管理

操作系统以专门的文件系统软件对数据进行操作，提供较人工阶段更为有效的数据管理模式。由于只具有部分的数据独立性，文件系统中数据冗余仍然较大，数据共享性也不够理想。随着对数据管理性能要求的提高，如更高的共享性、更好的独立性和更有效的数据查询与数据更新等实际需求，推动着数据管理的方法和技术朝新的方向不断提升和突破，数据库技术应运而生。

3. 数据库管理

通过应用程序和文件系统进行数据管理的实践进程，人们逐步认识到数据的有效管理实际上就是数据的结构化管理，这是因为在计算机系统内，数据和文件的简单堆积将缺乏使用价值。具体而言，由于实际应用中的数据结构复杂、数据量巨大，简单依靠应用程序乃至操作系统中的文件系统对数据进行管理使用存在着很大缺陷，需要有建立在操作系统之上的专门软件系统，这就是以统一管理和共享数据为设计目标的数据库管理系统(Database Management System，DBMS)。

1) 数据库数据管理特征

数据库系统出现于20世纪60年代末直至现在仍在使用，它具有下述基本特征。

(1) 数据共享性。数据作为整体应用单位的共享资源由DBMS统一管理。这种管理不依赖任何个别应用程序和个别用户，能够在系统级别上确实保证和真正实现数据的通用共享。

(2) 数据独立性。数据由DBMS统一调配使用，用户与数据管理在真正的逻辑和技术层面上实现了数据独立。由于其独立性导致数据存储和组织等细节透明，从而使得用户可以在更高的抽象层面上审视和访问数据库中的存储数据，为共享性提供了必需的技术支撑。

(3) 数据规范性。统一管理数据之后，系统能够立足于全局结构更加合理地组织和更为有效地调配数据，能够最小程度地减少数据冗余，更合理地设计和实现数据的标准化与规范性，从而更加有利于数据的转移传输和更大范围内的共用共享。

(4) 管理完备性。由于面对整个应用单位而非个别用户，因此DBMS能够研制得更为复杂庞大，从而具有更加多样和更为有效的功能。事实上，现有DBMS功能已经不仅仅限于一般的数据存储和查询，还具有查询优化、数据库保护(完整性与安全性)和事务管理(并发控制和故障恢复)等一整套完备机制，DBMS已经成为在层级和规模上都不逊于操作系统(OS)和办公自动化系统(OA)的大型系统软件。

数据库系统管理数据情形如图1-3所示。

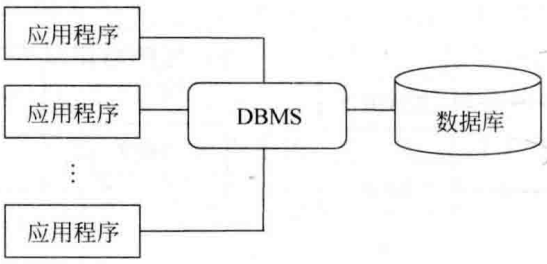

图1-3　基于数据库的数据管理

数据库技术是计算机学科中发展最快的应用领域之一，也是应用广泛的计算机关键技术之一。自20世纪60年代中期至今，四十多年的发展经历了三代演变过程，如今已经成为