数据库原理与应用

钟秋燕 黄灿辉 解正梅 编著

清华大学出版社

数据库原理与应用

钟秋燕 黄灿辉 解正梅 编著

清华大学出版社
北京

内 容 简 介

本书强化知识脉络，内容循序渐进，环环相扣；从培养应用型人才的目标出发，以数据库设计过程和数据库操作为主线，将数据库的原理与实际应用开发有机结合，增强学生的实际动手能力，培养真正满足社会需求的数据库技术人才。

本书共分为9章，第1章主要讲述数据库系统的基本概念以及数据库系统的组成和体系结构，第2章讲述数据库的设计过程；第3章～第5章主要讲述数据库的定义与操作；第6章讲述关系数据库的规范化；第7章讲述数据库系统管理；第8章和第9章讲述数据库的编程。

本书既可作为大中专院校学生学习数据库系统的教材，也可供数据库爱好者参考。

图书在版编目(CIP)数据

数据库原理与应用/钟秋燕，黄灿辉，解正梅编著. --北京：清华大学出版社，2016
21世纪高等学校规划教材·计算机科学与技术
ISBN 978-7-302-45000-9

Ⅰ. ①数… Ⅱ. ①钟… ②黄… ③解… Ⅲ. ①关系数据库系统－高等学校－教材 Ⅳ. ①TP311.138

中国版本图书馆CIP数据核字(2016)第213426号

责任编辑：闫红梅　李　晔
封面设计：傅瑞学
责任校对：李建庄
责任印制：刘海龙

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 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：14　　**字　　数**：339千字
版　　次：2016年10月第1版　　**印　　次**：2016年10月第1次印刷
印　　数：1～2000
定　　价：29.00元

产品编号：063385-01

出版说明

随着我国改革开放的进一步深化，高等教育也得到了快速发展，各地高校紧密结合地方经济建设发展需要，科学运用市场调节机制，加大了使用信息科学等现代科学技术提升、改造传统学科专业的投入力度，通过教育改革合理调整和配置了教育资源，优化了传统学科专业，积极为地方经济建设输送人才，为我国经济社会的快速、健康和可持续发展以及高等教育自身的改革发展做出了巨大贡献。但是，高等教育质量还需要进一步提高以适应经济社会发展的需要，不少高校的专业设置和结构不尽合理，教师队伍整体素质亟待提高，人才培养模式、教学内容和方法需要进一步转变，学生的实践能力和创新精神亟待加强。

教育部一直十分重视高等教育质量工作。2007 年 1 月，教育部下发了《关于实施高等学校本科教学质量与教学改革工程的意见》，计划实施“高等学校本科教学质量与教学改革工程”（简称“质量工程”），通过专业结构调整、课程教材建设、实践教学改革、教学团队建设等多项内容，进一步深化高等学校教学改革，提高人才培养的能力和水平，更好地满足经济社会发展对高素质人才的需要。在贯彻和落实教育部“质量工程”的过程中，各地高校发挥师资力量强、办学经验丰富、教学资源充裕等优势，对其特色专业及特色课程（群）加以规划、整理和总结，更新教学内容、改革课程体系，建设了一大批内容新、体系新、方法新、手段新的特色课程。在此基础上，经教育部相关教学指导委员会专家的指导和建议，清华大学出版社在多个领域精选各高校的特色课程，分别规划出版系列教材，以配合“质量工程”的实施，满足各高校教学质量和教学改革的需要。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上。精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学科体系有实质性的改革和发展，顺应并符合 21 世纪教学发展的规律，代表并引领课程发展的趋势和方向）、示范性（教材所体现的课程体系具有较广泛的辐射性和示范性）和一定的前瞻性。教材由个人申报或各校推荐（通过所在高校的“编委会”成员推荐），经“编委会”认真评审，最后由清华大学出版

社审定出版。

目前，针对计算机类和电子信息类相关专业成立了两个“编委会”，即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。推出的特色精品教材包括：

(1) 21 世纪高等学校规划教材·计算机应用——高等学校各类专业，特别是非计算机专业的计算机应用类教材。

(2) 21 世纪高等学校规划教材·计算机科学与技术——高等学校计算机相关专业的教材。

(3) 21 世纪高等学校规划教材·电子信息——高等学校电子信息相关专业的教材。

(4) 21 世纪高等学校规划教材·软件工程——高等学校软件工程相关专业的教材。

(5) 21 世纪高等学校规划教材·信息管理与信息系统。

(6) 21 世纪高等学校规划教材·财经管理与应用。

(7) 21 世纪高等学校规划教材·电子商务。

(8) 21 世纪高等学校规划教材·物联网。

清华大学出版社经过三十多年的努力，在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌，为我国的高等教育事业做出了重要贡献。清华版教材形成了技术准确、内容严谨的独特风格，这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

联系人：魏江江

E-mail：weijj@tup.tsinghua.edu.cn

前言

数据库技术是计算机数据处理与信息管理系统的核心，也是应用最广的技术之一。作为计算机专业的大学生甚至非计算机专业的学生，掌握数据库技术是非常必要的。

本书作者都是从事数据库教学多年并致力于数据库技术及应用和研究的一线教师，在多年教学经验的基础上，理顺知识脉络，精简知识内容，从培养应用型人才的目标出发，以数据库设计过程和数据库操作为主线，将数据库的原理与实际应用开发有机结合，增强学生的实际动手能力，培养真正满足社会需求的数据库技术人才。本书既可作为大中专院校学生学习数据库系统的教材，也可供数据库爱好者参考。

本书分为9章。第1章介绍数据库及其相关的概念；第2章介绍数据库的设计，基于数据库设计；第3章介绍利用SQL对数据库和表结构定义；在建好数据库、表的基础上，第4章和第5章介绍利用SQL语言对数据库的操作，第6章讲述关系数据库的规范化，第7章关系数据库系统管理，第8章和第9章介绍数据库编程技术，实现了学生选课系统的实例；形成从无到有，从理论到实践的体系结构。

本书的第1章、第2章和第9章由钟秋燕编写，第3章、第4章和第8章由解正梅编写，第5章、第6章和第7章由黄灿辉编写。清华大学出版社的编辑详细审阅了书稿，并提出了许多宝贵意见，在此表示衷心的感谢。

本书在编写过程中参考了国内外的同类教材，具体书目见书末参考文献，在此，我们谨向这些教材的编者表示衷心的感谢。

由于编者水平所限，缺点和疏漏之处在所难免，恳请同行专家和广大读者批评指正。

编　者

2016年6月

目 录

第 1 章　数据库系统概述 …… 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　文件系统管理阶段 …… 2

1.1.3　数据库系统管理阶段 …… 4

1.1.4　高级数据库阶段 …… 4

1.2　数据库系统 …… 5

1.2.1　数据库系统的组成 …… 5

1.2.2　数据库系统的特点 …… 7

1.3　数据库管理系统 …… 9

1.3.1　SQL Server 2008 简介 …… 10

1.3.2　SQL Server 2008 的组件与功能 …… 10

1.3.3　SQL Server Management Studio …… 11

1.3.4　配置 SQL Server 服务 …… 12

1.3.5　数据库的基本操作 …… 13

1.4　数据库系统结构 …… 18

1.4.1　三级模式结构 …… 18

1.4.2　二级映像功能 …… 20

本章小结 …… 21

习题 1 …… 21

第 2 章　关系数据库的设计 …… 23

2.1　数据库设计概述 …… 23

2.2　概念模型的设计 …… 24

2.2.1　E-R 模型的基本概念 …… 24

2.2.2　子类的设计 …… 28

2.2.3　E-R 图设计实例 …… 28

2.3　逻辑模型的设计 …… 31

2.3.1　数据结构——关系 …… 31

2.3.2　关系的操作和完整性约束 …… 35

2.3.3　E-R 图向关系模型的转换 …… 35

2.4　物理模型的设计 …… 38

2.4.1 物理结构设计的任务 …… 38
2.4.2 物理结构设计方法 …… 38
2.4.3 学生选课管理数据库的物理设计 …… 39
2.5 数据库的实施与维护 …… 40
2.5.1 数据库实施 …… 40
2.5.2 数据库运行和维护阶段 …… 40
2.6 使用 Management Studio 创建数据表 …… 40
本章小结 …… 44
习题 2 …… 45

第 3 章 关系数据库的定义与完整性的实现 …… 47

3.1 SQL 语言 …… 47
3.1.1 SQL 的特点 …… 47
3.1.2 SQL 的主要功能 …… 48
3.1.3 SQL Server 提供的主要数据类型 …… 49
3.2 关系数据库的定义 …… 50
3.2.1 数据库的创建 …… 50
3.2.2 数据库的删除 …… 53
3.3 SQL 表结构的定义 …… 53
3.3.1 基本表的创建 …… 53
3.3.2 修改表结构 …… 54
3.3.3 删除表 …… 55
3.4 完整性约束 …… 55
3.4.1 实体完整性 …… 56
3.4.2 参照完整性 …… 57
3.4.3 用户定义完整性 …… 58
本章小结 …… 61
习题 3 …… 62

第 4 章 查询、视图与索引 …… 64

4.1 关系代数 …… 64
4.1.1 传统的集合运算 …… 65
4.1.2 专门的关系运算 …… 67
4.2 单表查询 …… 73
4.2.1 基本查询 …… 73
4.2.2 使用列表达式 …… 75
4.2.3 查询满足条件的元组 …… 76
4.2.4 对查询结果进行排序 …… 80
4.2.5 聚合函数 …… 80

4.2.6 GROUP BY 子句 …… 81
4.3 连接查询 …… 82
4.3.1 内连接查询 …… 82
4.3.2 自连接查询 …… 84
4.3.3 外连接查询 …… 86
4.4 子查询 …… 88
4.5 集合查询 …… 94
4.6 视图 …… 96
4.6.1 定义视图 …… 97
4.6.2 修改和删除视图 …… 99
4.6.3 查询视图 …… 99
4.6.4 更新视图数据 …… 101
4.6.5 视图的作用 …… 102
4.6.6 物化视图 …… 103
4.7 索引 …… 104
4.7.1 索引的建立 …… 104
4.7.2 索引的删除 …… 105
4.7.3 建立索引的原则 …… 106
本章小结 …… 106
习题 4 …… 107

第 5 章 数据操作 …… 109

5.1 数据的插入 …… 109
5.1.1 插入一个元组 …… 109
5.1.2 插入多个元组 …… 110
5.2 数据的更改 …… 110
5.2.1 无条件更改 …… 111
5.2.2 有条件更改 …… 111
5.3 数据的删除 …… 111
5.3.1 无条件删除 …… 112
5.3.2 有条件删除 …… 112
本章小结 …… 113
习题 5 …… 113

第 6 章 关系数据库的规范化 …… 114

6.1 函数依赖 …… 114
6.1.1 关系数据库中的问题 …… 114
6.1.2 函数依赖的基本概念 …… 115
6.1.3 一些术语和符号 …… 116

6.1.4 关系模式中的码 …… 117
6.1.5 函数依赖的推理规则 …… 118
6.2 关系模式的规范化 …… 120
6.2.1 第一范式 …… 120
6.2.2 第二范式 …… 121
6.2.3 第三范式 …… 122
6.2.4 BC 范式 …… 123
6.2.5 将关系规范到 BCNF …… 124
6.3 模式分解 …… 125
本章小结 …… 127
习题 6 …… 128

第 7 章 管理数据库 …… 129

7.1 数据库的安全管理 …… 129
7.1.1 数据库安全控制的目标 …… 130
7.1.2 数据库安全的威胁 …… 130
7.1.3 数据库安全问题的类型 …… 131
7.1.4 安全控制模型 …… 131
7.1.5 授权和认证 …… 131
7.1.6 自主存取控制方法 …… 132
7.1.7 强制存取控制(MAC)方法 …… 134
7.1.8 视图机制 …… 135
7.1.9 审计跟踪 …… 136
7.1.10 统计数据库安全性 …… 136
7.2 数据库的恢复技术 …… 137
7.2.1 事务的基本概念 …… 137
7.2.2 数据库恢复概述 …… 139
7.2.3 恢复的实现技术 …… 141
7.2.4 恢复策略 …… 144
7.2.5 具有检查点的恢复技术 …… 145
7.2.6 数据库镜像 …… 147
7.3 并发控制 …… 148
7.3.1 并发控制概述 …… 148
7.3.2 封锁 …… 153
7.3.3 并发调度可串行化的两个充分条件 …… 156
本章小结 …… 158
习题 7 …… 159

第 8 章 T-SQL 程序设计与开发 …… 161

8.1 T-SQL 程序设计基础 …… 161
8.1.1 变量 …… 161
8.1.2 运算符 …… 163
8.1.3 函数 …… 165
8.2 流程控制语句 …… 169
8.2.1 语句块：BEGIN…END …… 170
8.2.2 条件执行：IF…ELSE 语句 …… 170
8.2.3 多分支 CASE 表达式 …… 171
8.2.4 循环：WHILE 语句 …… 172
8.2.5 非条件执行：GOTO 语句 …… 174
8.2.6 调度执行：WAIT FOR …… 174
8.3 游标 …… 175
8.3.1 游标的原理及使用方法 …… 175
8.3.2 游标应用举例 …… 178
8.4 存储过程 …… 180
8.4.1 存储过程的创建与执行 …… 180
8.4.2 存储过程的管理与维护 …… 182
8.4.3 用户自定义函数 …… 184
8.5 触发器 …… 187
8.5.1 触发器的基本概念 …… 188
8.5.2 创建触发器 …… 188
8.5.3 管理触发器 …… 191
本章小结 …… 192
习题 8 …… 192

第 9 章 SQL Server 2008 编程应用实例 …… 194

9.1 数据库应用结构 …… 194
9.1.1 客户/服务器结构 …… 194
9.1.2 浏览器/服务器结构 …… 195
9.2 数据访问接口 …… 195
9.2.1 ODBC …… 195
9.2.2 ADO …… 196
9.2.3 JDBC …… 197
9.3 数据库应用系统的开发 …… 198
9.4 数据库设计 …… 199
9.4.1 数据的需求分析 …… 199
9.4.2 概念模式设计 …… 199

9.4.3 逻辑模式设计…… 200
9.4.4 物理模型的设计…… 200
9.4.5 数据库的实施…… 201
9.5 系统实现 …… 203
本章小结…… 208
习题 9 …… 208

参考文献…… 209

第1章 数据库系统概述

随着信息管理水平的不断提高，信息资源已成为企业的重要财富和资源，用于信息管理的数据库技术也得到很大的发展，其应用领域也越来越广泛。数据库的应用形式日益多样，从小型事务处理到大型信息系统，从联机事务处理到联机分析处理，从一般企业管理到计算机辅助设计与制造(CAD/CAM)，乃至全地理信息系统等都应用了数据库技术。数据库技术已经渗透到人们日常生活的方方面面，比如用信用卡购物，飞机、火车订票系统，图书馆对书籍从借阅的管理等无一不使用了数据库技术。数据库的建设规模、数据库中信息量的大小以及使用的程度已经成为衡量企业乃至国家的信息化程度的重要标志。

简单地说，数据库技术就是研究如何科学地管理数据，以便为人们提供可共享的、安全的、可靠的数据的技术。数据库技术一般包括数据管理和数据处理两部分内容。

数据库系统实质上是一个用计算机存储数据的系统，可以将数据库看作一个电子文件柜，也就是说，数据库是收集数据文件的仓库或容器。

1.1 数据管理技术的发展

数据管理是指对数据进行分类、组织、编码、存储、检索和维护，它是数据处理的核心。而数据处理是指对各种数据进行收集、存储、加工和传播的一系列活动的总称。

在计算机产生之前，对数据的管理只能是手工和机械的方式。在计算机问世以后，在应用需求的驱动下，在计算机硬件、软件发展的支撑基础上，数据管理技术经历了人工管理、文件系统管理和数据库管理3个阶段。

1.1.1 人工管理阶段

在人工管理阶段(20世纪50年代中期以前)，计算机主要用于科学计算。硬件方面的状况是，外部存储器只有磁带、卡片和纸带等，还没有磁盘等直接存取存储设备，所以数据不能联机保存。软件方面还没有出现操作系统，尚无数据管理软件，应用程序(用户)负责数据管理，可以对数据进行批处理，如图1.1所示。由于数据管理是由用户自己完成，因此称为人工管理。人工管理数据具有如下特点。

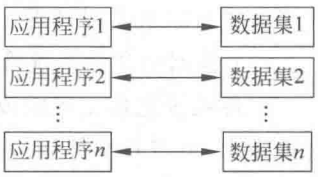

图1.1 人工管理阶段应用程序与数据之间的对应关系

1. 数据不保存

计算机主要用于计算，并不对数据进行其他操作，也没有磁盘等直接存取设备，数据不保存在计算机系统中，程序中的数据，随着程序的运行完成，其所占用的内存空间同指令所占用的内存空间一起释放，退出计算机系统。

2. 数据的管理者是应用程序

数据需要由应用程序自己设计、说明(定义)和管理，程序员在编写程序时，要规定数据的存储结构、存取方法和输入方式等。

3. 数据的共享程度：无共享、冗余度极大

数据完全面向特定的应用程序，数据的产生和存储依赖于定义和使用数据的程序。多个程序使用相同数据时，也必须各自定义，数据不能共享，造成数据的重复存储，产生数据冗余。

4. 数据的独立性：不独立，完全依赖于程序

数据独立性包括数据的物理独立性和数据的逻辑独立性。物理独立性是指用户的应用程序与数据的存储结构是相互独立的，当数据的存储位置或者存储结构改变了，应用程序不需要发生改变。逻辑独立性是指用户的应用程序与数据的逻辑结构是相互独立的，即当数据的逻辑结构改变时，比如增加列或者删除列，用户程序也可以不变。

在人工管理阶段，没有专门的软件对数据进行管理，程序直接面向存储结构，当数据的存储结构发生变化时，应用程序必须做相应的修改，对数据进行重新定义。因此程序员的负担很重。

1.1.2 文件系统管理阶段

文件系统管理阶段是指 20 世纪 50 年代后期到 20 世纪 60 年代中期这一阶段。从那时起，计算机不仅大量用于科学计算，也开始大量用于信息管理。在计算机硬件方面，有了磁盘、磁鼓等直接存取设备；在计算机软件方面，已经有了操作系统和高级语言，操作系统中有了专门管理数据的软件，即文件管理系统；在数据处理方式上，不仅可以进行批处理，而且还能进行联机实时处理。文件系统管理数据有如下特点。

(1) 数据由文件系统管理。

文件系统把数据组织成相互独立的数据文件，利用“按文件名访问，按记录进行存取”的管理技术，可以对文件进行插入、删除和修改操作。文件系统实现了记录内的有结构，但整体无结构。程序和数据之间由文件系统提供存取方法进行转换，是应用程序与数据之间有了一定的独立性，程序员可以不必过多地考虑物理细节。

(2) 数据可以长期保存。

数据可以以“文件”的形式长期保存在磁盘等外部存储器上，应用程序可通过文件系统对磁盘上的文件中的数据进行管理。

现在看一下文件管理方式下的数据的操作模式。假设现在用系统来实现对学生进行管理的程序，在此系统中，要对学生的基本信息和选课情况进行管理；在管理学生况信息包括学生的基本信息、课程的基本信息和学生的选课信息。假设用 F2 和 F3 两个文件分别存储课程基本信息和学生选课信息。学生选课情况管理中涉及的学生基本信息可以使用学生基本信息管理系统中的 F1 文件。假设实现学生基本信息管理功能的应用程序叫 A1，实现学生选课管理功能的应用程序叫 A2，则学生的基本信息和选课情况可用图 1.2 表示。

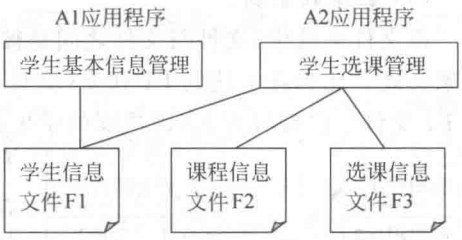

图 1.2　文件管理系统实例

假设 F1、F2 和 F3 文件分别包含如下信息：

F1 包含学号、姓名、性别、出生日期、所在系、专业、所在班、特长、家庭住址。

F2 包含课程号、课程名、授课学期、学分、课程性质。

F3 包含学号、姓名、所在系、课程号、课程名、修课类型、修课时间、考试成绩。

我们将文件中所包含的每一个子项称为文件结构中的字段或列，将每一行数据称为一个记录。

“学生选课管理”系统的处理过程大致为：在学生选课管理系统中，若有学生选课，则先查 F1 文件，判断有无此学生。若有此学生，则再访问 F2 文件，判断其所选的课程是否存在。若课程也存在，就将学生选课信息写到 F3 文件中。

这看起来似乎很好。但仔细分析一下，就会发现使用文件管理系统管理数据有如下一些缺点。

(1) 数据冗余大。

假设 A2 需要用到 F3 文件中包含的学生的所有或大部分信息，比如，除了学号之外，还需要姓名、性别、专业、所在系等信息，而 F1 个也包含了这些信息，因此 F3 和 F1 文件中有重复的信息，但这些重复的信息只是不同文件的部分内容，因此很难在两个文件中共用这些公共信息，从而造成数据的重复，即数据的冗余。

(2) 数据不一致性。

数据冗余不仅会造成存储空间的浪费；其实，随着计算机硬件技术的飞速发展，存储容量不断扩大，空间问题已经不是解决问题时需要关心的主要问题，更为严重的是造成了数据的不一致。例如，假设某个学生所学的专业发生了变化，我们一般只会想到在 F1 文件中进行修改，而往往忘记在 F3 中要进行同样的修改。这样就会造成同一名学生在 F1 文件和 F3 文件中的“专业”不一样，也就是数据不一致。

(3) 程序和数据之间的独立性差。

文件和记录的结构通常是应用程序代码的一部分，如 C 语言的结构(struct)。文件结

构每进行一次修改，比如添加字段、删除字段甚至是修改字段的长度（如电话号码从 7 位扩到 8 位），都要对应用程序进行相应的修改，因为我们在打开文件读取数据时，必须要将文件记录中的不同字段的值对应到程序变量中。随着应用环境和需求的变化，修改文件的结构是不可避免的事情，这样就需要在应用程序中进行相应的修改，也就是说，程序和数据之间的独立性差。频繁修改应用程序是很麻烦的。

（4）数据联系弱。

在文件系统中，文件与文件之间是彼此独立、毫不相干的，文件之间的联系必须通过程序来实现。比如在上述的 F1 和 F3 文件中，F3 文件中的学号、姓名等学生的基本信息必须是 F1 文件中已经存在的（即选课的学生必须是已经存在的学生）。同样，F3 中的课程号等与课程有关的基本信息也必须是 F2 文件中已经存在的。这些数据之间的联系是客观需求当中所要求的很自然的联系。但文件系统本身不具备自动实现这些联系的功能，所以必须通过应用程序来保证这些联系，也就是说，必须编写代码来手工地保证这些联系。

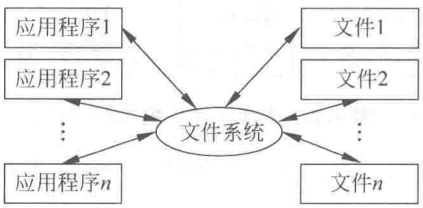

图 1.3　文件系统阶段应用程序与数据之间的对应关系

图 1.3 描述了文件系统阶段应用程序与数据之间的对应关系。

1.1.3　数据库系统管理阶段

20 世纪 60 年代后期以来，计算机应用范围越来越广泛，数据量急剧增加，计算机管理数据的规模越来越大，同时多种应用同时共享数据集合的要求也越来越强烈。随着大容量磁盘的出现，硬件价格不断下降，软件价格不断攀升；数据处理方式是，联机实时处理要求更多，并开始提出和开始考虑分布式处理。在这种背景下，以文件方式管理数据已经不能满足应用的需求，于是出现了新的数据管理技术，即数据库技术；同时出现了专门管理数据的软件：数据库管理系统（DataBase Manager System，DBMS）。数据库的数据不再面向某个应用程序，而是面向整个企业或者整个应用，它克服了人工管理阶段和文件管理阶段的缺陷，图 1.4 示意了这种系统的特点，1.2 节将详细阐述数据库相关知识。

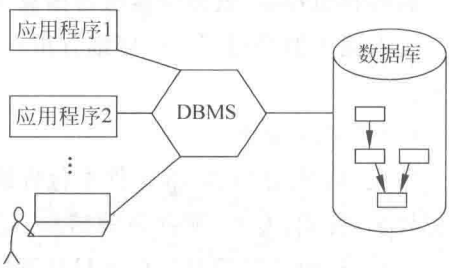

图 1.4　数据库管理阶段应用程序与数据之间的对应关系

1.1.4　高级数据库阶段

从 20 世纪 80 年代开始，数据库系统的技术也在不断地完善和发展，有关数据库的新的研究课题不断取得进展，如分布式数据库、Web 数据库、XML 数据库的应用日益成熟。传统数据库并未专为数据分析而设计，数据仓库专用设备的兴起（Data Warehouse Appliance），如 Teradata、Netezza、Greeplum、Sybase IQ 等等，正表明面向事务性处理的传统数据库和面向分析的分析型数据库走向分离，泾渭分明。数据仓库专用设备，一般都会采

用软硬一体，以提供最佳性能。这类数据库会采用更适于数据查询的技术，以列式存储或MPP(大规模并行处理)两大成熟技术为代表。另外，新兴的互联网企业也在尝试一些新技术，比如MapReduce技术(这要感谢Google公司将它发扬光大)，Yahoo的开源小组开发出Hadoop，就是一种基于MapReduce技术的并行计算框架。在2008年之前，Facebook就在Hadoop基础上开发出类似数据仓库的Hive，用来分析点击流和日志文件。几年下来，基于Hadoop的整套数据仓库解决方案已日臻成熟。目前在国内也有不少应用，尤其在互联网行业的数据分析，很多就是基于这个开源方案，比如淘宝的数据魔方。而在一些商业性的产品中，也已经融入MapReduce技术，如AsterData。

随着大数据时代的到来，数据类型非常丰富，比如文本、语音、图像、社交网络、地理位置。用关系型数据库存储这类数据，再深入去分析挖掘这些数据，开始让人感到有些麻烦。

于是，越来越多的NoSQL数据库涌现出来，其中很大一部分是用于分析用途。比如西班牙有个小厂商，叫illumnate，他们拥有一个叫Correlation DBMS的数据库产品。它不像关系数据库那样按照表、字段存储，那样冗余很大。CDBMS的做法是，针对每个不同的值，只有一个地方存储，而所有对这个值的引用，都在索引中记录。比如有个客户的姓名叫“张三”，而还有一个公司名字也叫“张三”，那么在CDBMS里面，只存有一个“张三”的值，但在索引里面记录了有两个地方引用它。这种数据库是专门为分析而设计的。因为不存储冗余数据，所以它对于海量数据，非常节省空间。如果说这还不够吸引人的话，另一个突出的优点就是做ad-hoc查询非常快捷。

随着计算机技术的发展和各种应用的普及，数据库技术还会朝着支持更大规模、更快的速度、更广泛的应用等方向发展。

1.2 数据库系统

1.2.1 数据库系统的组成

数据库系统(DataBase System，DBS)是指在计算机系统中引入数据库后的系统。一般包括4个主要部分：数据库、数据库管理系统、应用程序和数据库管理员。如图1.5所示，其中数据库管理系统是核心，应用程序对数据库的所有操作都由数据库管理系统来完成。

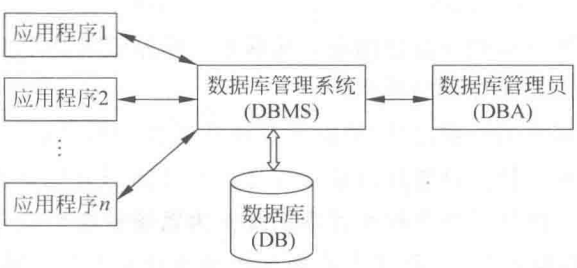

图1.5 数据库系统简图