HZ BOOKS
华章教育

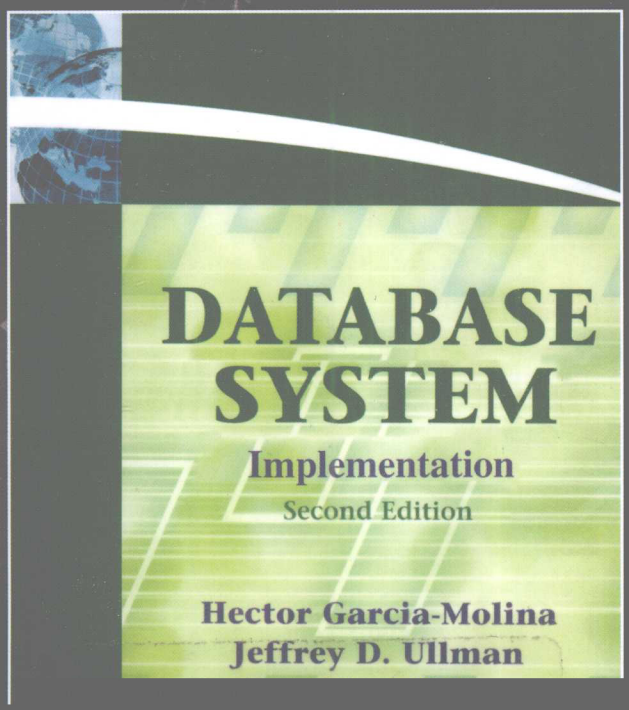
计 算 机 科 学 丛 书

第2版

PEARSON

数据库系统实现

(美) Hector Garcia-Molina Jeffrey D. Ullman Jennifer Widom 著
斯坦福大学
杨冬青 吴愈青 包小源 唐世渭 等译



Database System Implementation
Second Edition

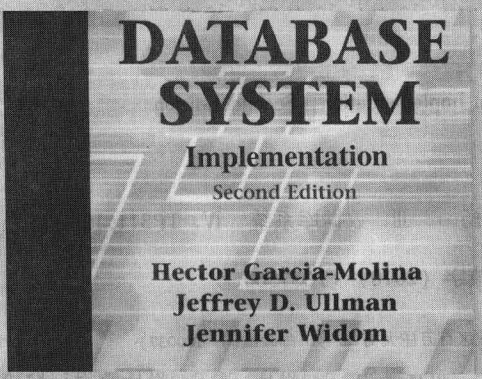
机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

第2版

数据库系统实现

(美) Hector Garcia-Molina Jeffrey D. Ullman Jennifer Widom 著
斯坦福大学
杨冬青 吴愈青 包小源 唐世渭 等译



Database System Implementation

Second Edition

机械工业出版社
China Machine Press

本书是斯坦福大学计算机科学专业数据库系列课程第二门课的教科书。书中对数据库系统实现原理进行了深入阐述，并具体讨论了数据库管理系统的三个主要成分——存储管理器、查询处理器和事务管理器的实现技术。此外，第2版充分反映了数据管理技术的新进展，对内容进行了扩充，除了在第1版中原有的“信息集成”一章（第10章）中加入了新的内容外，还增加了两个全新的章：“数据挖掘”（第11章）和“数据库系统与互联网”（第12章）。

本书适合作为高等院校计算机专业研究生的教材或本科生的教学参考书，也适合作为从事相关研究或开发工作的专业技术人员的高级参考资料。

本书版权登记号：图字：01-2009-1342

图书在版编目（CIP）数据

数据库系统实现　第2版/(美) 加西亚-莫利纳 (Garcia-Molina, H.) 等著；杨冬青等译. —北京：机械工业出版社，2010.5

(计算机科学丛书)

书名原文：Database System Implementation, Second Edition

ISBN 978-7-111-30287-2

Ⅰ. 数…　Ⅱ. ①加…　②杨…　Ⅲ. 数据库系统　Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字（2010）第057560号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：迟振春

北京京师印务有限公司印刷

2010年5月第2版第1次印刷

184mm×260mm · 25印张

标准书号：ISBN 978-7-111-30287-2

定价：59.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010) 88378991；88361066

购书热线：(010) 68326294；88379649；68995259

投稿热线：(010) 88379604

读者信箱：hzjsj@hzbook.com

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www. hzbook. com

电子邮件：hzjsj@ hzbook. com

联系电话：（010）88379604

联系地址：北京市西城区百万庄南街 1 号

邮政编码：100037

华章教育

华章科技图书出版中心

译者序

随着计算机硬件、软件技术的飞速发展和计算机系统在各行各业的广泛应用，数据已经成为各种机构的宝贵资源，数据库系统对于当今科研部门、政府机关、企事业单位等来说都是至关重要的。而数据库系统中的核心软件是数据库管理系统（DBMS）。DBMS 用于高效地创建和存储大量的数据，并对数据进行有效的管理、处理和维护，是数据库专家和技术人员数十年研究开发的结果，是当前最复杂的系统软件之一。要深入掌握数据库系统的原理和技术，进而从事数据库管理软件和工具的开发，必须学习和研究数据库管理系统实现技术。要深入了解数据库系统的内部结构，以开发出高效的数据库应用系统，也需要学习和研究数据库管理系统实现技术。

Hector Garcia-Molina、Jeffrey D. Ullman 和 Jennifer Widom 是斯坦福大学著名的计算机科学家，多年来他们在数据库系统领域中做了大量的开创性工作，由他们撰写的《数据库系统实现》一书是关于数据库系统实现方面内容最为全面的著述之一。我们于 2000 年将《数据库系统实现》的第 1 版译成中文，国内许多大学采用它作为研究生数据库课程的教材或主要教学参考书，收到了良好的效果。

现在我们又翻译了《数据库系统实现》第 2 版。第 2 版保持了第 1 版的总体风格，首先对数据库系统实现原理进行了深入阐述，并具体讨论了数据库管理系统的三个主要成分——存储管理器、查询处理器和事务管理器的实现技术。与第 1 版相比，第 2 版对于数据存储和索引结构的阐述进行了适当的压缩，分别将原来的两章合并为一章；另外，增加了一章“并行与分布式数据库”（第 9 章），其中包括了第 1 版中分散在查询处理和事务管理的相关章节中的内容，并增加了有关分布式查询执行的一些新内容，例如，map-reduce 并行架构、P2P 数据库以及分布式散列的实现等。同时，第 2 版充分反映了数据管理技术的新进展，对内容进行了扩充，除了在第 1 版中原有的“信息集成”一章（第 10 章）中加入了新的内容外，还增加了两个全新的章：“数据挖掘”（第 11 章）和“数据库系统与互联网”（第 12 章）。“数据挖掘”一章中包含了关联规则与频繁项集挖掘技术，从一个非常大的数据库或 Web 页面集合中发现“相似”的项的“最小散列”和“局部敏感散列”等关键技术，以及高维空间中大规模数据的聚簇问题等。“数据库系统与互联网”一章中重点阐述了与互联网相关的两个方面的数据库技术：Web 搜索引擎及其 PageRank 算法，流数据模型以及管理数据流形式的大量数据所需的技术。

我们认为这本书既适合作为高等院校计算机专业研究生的教材或本科生的教学参考书，又适合作为从事相关研究或开发工作的专业技术人员的高级参考资料。

杨冬青全面组织了本书的翻译，吴愈青、包小源、唐世渭在本书的翻译和审校中做了大量的工作。除此之外，参加翻译的还有闫秋玲、郑丽丽、蔡慧慧、马煜、张棋、陈巍、郭思祺、夏海峰、翁学天、郭少松、李树节。

限于译者水平，译文中难免有疏漏和错误，欢迎批评指正。

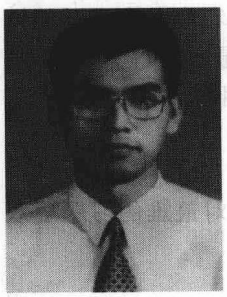
译　者

于北京大学

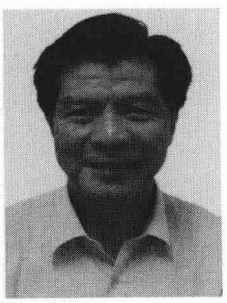
译者简介

杨冬青 1969年毕业于北京大学数学力学系数学专业，现任北京大学信息科学技术学院教授，博士生导师，计算机科学技术系主任，中国计算机学会数据库专委会委员。多年来承担并完成973、863、国家科技攻关、国家自然科学基金等多项国家重点科研项目，曾获国家科技进步二等奖、三等奖和多项省部级奖励，在国内外杂志及会议上发表论文百余篇，著译作十余部。目前主要研究方向为数据库系统实现技术、Web环境下的信息集成与共享、数据仓库和数据挖掘等。

吴愈青 分别于1995年和1998年在北京大学计算机系获得学士学位和硕士学位；2004年于EECS Department, University of Michigan获得博士学位。现就职于美国Indiana University，任Assistant Professor。主要研究方向为数据库系统及实现，包括数据库查询语言、查询优化、索引技术等，及其在半结构化数据上的应用与实现。在国际会议及期刊上发表论文三十余篇。



包小源 博士，副教授。兰州大学计算数学专业硕士、北京大学计算机应用专业博士。主要研究方向为数据库实现技术、XML数据管理、对等计算、服务计算等。



唐世渭 1964年毕业于北京大学数学力学系计算数学专业，毕业后留校任教至今，现为北京大学信息科学技术学院教授，博士生导师，中国计算机学会数据库专委会委员，中国软件行业协会数据库及应用软件分会理事长。多年来承担并完成973、863、国家科技攻关、国家自然科学基金等多项国家重点科研项目，曾获国家科技进步二等奖、三等奖各1项，省部级科技进步奖多项，在国内外杂志及会议上发表论文百余篇，著译作多部。目前主要研究方向为数据库系统、数据仓库和数据挖掘、Web环境下的信息集成与共享、典型应用领域的信息系统等。

出版前言

在斯坦福大学，因为实行的是一年四学期制，所以数据库引论课被分为两门课程。第一门课程是 CS145，该课程只要求学生学会使用数据库系统，而不要求知道 DBMS 实现的内容。CS145 是 CS245 的预修课，CS245 介绍 DBMS 实现。学生若想进一步学习数据库方面的课程，可以学习 CS345（此课是理论课）、CS346（此课是 DBMS 实现实验课）以及 CS347 课程（此课介绍事务处理及分布式数据库）。

从 1997 年开始，我们已经出版了两本配套教材：《数据库系统基础教程》[⊖]是为 CS145 课程编写的，《数据库系统实现》是为 CS245 课程以及部分 CS346 课程编写的。本书就是《数据库系统实现》的最新版——第 2 版。

第 2 版保持了第 1 版的总体风格，但对于数据存储和索引结构的阐述进行了适当的压缩，分别将原来的两章合并为一章；另外，增加了一章“并行与分布式数据库”（第 9 章），其中包括了第 1 版中分散在查询处理和事务管理的相关章节中的内容，并增加了有关分布式查询执行的一些新内容。同时，第 2 版充分反映了数据管理技术的新进展，对内容进行了扩充，除了在第 1 版中原有的“信息集成”一章（第 10 章）中加入了新的内容外，还增加了两个全新的章：“数据挖掘”（第 11 章）和“数据库系统与互联网”（第 12 章）。

预备知识

学生一般不会在大学的第一学年选修数据库系统实现课程，所以本书读者应具有广泛的计算机科学背景知识。我们假定读者已经学过数据库语言，特别是 SQL。读者最好了解关系代数，并且熟悉基本的数据结构。同样，关于文件系统和操作系统的知识也是很有用的。

习题

本书几乎在每一节都包括大量的练习，我们用感叹号对难题做了标记，对最难的习题用双感叹号做了标记。

网上支持

本书的网址是：

http：//infolab. stanford. edu/ ~ ullman/fcdb. html

该网站包括勘误表及支持材料。

⊖ 本书中文版（ISBN 978-7-111-26828-4）和影印版（ISBN 978-7-111-24733-3）已由机械工业出版社出版。——编辑注

目 录

第 1 章　DBMS 系统概述

数据库对于当今的任何部门都是至关重要的。无论何时访问一个主要的网站——Yahoo!、Amazon. com 或者数以千计的提供信息的小网站，总有一个数据库在幕后服务，提供你所需要的信息。公司用数据库来维护其所有重要的记录。在许多科学研究的核心中也同样需要数据库。天文学家、人类学研究人员、探索蛋白质的医药特性的生物化学家以及许多其他的科学研究人员收集到的数据也是用数据库表示的。

数据库的能力来源于一个知识和技术的结合体，这是数十年研究开发的结果，并且已经嵌入到专门的软件中，这个软件称作*数据库管理系统*或 DBMS，或更通俗地称为“数据库系统”。DBMS 是一个强有力的工具，用于高效地创建和管理大量的数据，并使得数据能够安全地长期保存。DBMS 是当前最复杂的软件系统之一。在本书中，我们主要学习数据库管理系统实现技术，还将探讨当前数据库系统研究的一些新课题。

1.1　数据库系统的发展

什么是数据库？本质上讲，数据库就是信息的集合，它可以存在很长时间，往往是很多年。一般来讲，“数据库”这个词指的是由数据库管理系统管理的数据的集合。数据库管理系统将满足：

1. 允许用户使用专门的*数据定义语言*来创建新的数据库并指定其*模式*(数据的逻辑结构)。

2. 给予用户使用适当的语言来*查询*数据(“查询”是数据库术语，指关于数据的问题)和修改数据的能力，这种语言通常称为*查询语言*(query language)或*数据操纵语言*(data-manipulation language)。

3. 支持对非常大量的数据(许多 TB 或者更多)长期地进行存储，允许高效地存取数据以进行查询和数据库修改。

4. 使数据具有*持久性*(durability)，即能够从故障、多种类型的错误或者故意滥用中进行恢复。

5. 控制多个用户同时对数据进行访问，不允许用户间有不恰当的相互影响(称作*孤立性*(isolation))，并且不会发生在数据上进行了部分的而不是完整的操作的情况(称作*原子性*(atomicity))。

1.1.1　早期的数据库管理系统

第一个商用数据库管理系统产生于 20 世纪 60 年代末。这些系统是由文件系统演变而来的，提供了对上述第(3)条的部分支持：文件系统可以长时间地存储数据，并且允许存储大量的数据。但是，文件系统并不能保证数据不丢失，如果没有备份的话；而且它们也不支持对数据项的高效存取，如果不知道它在特定的文件里的存储位置的话。

文件系统不直接支持第(2)条，即针对文件中数据的查询语言。它们对第(1)条(数据模式)的支持仅限于创建文件的目录结构。第(4)条并不总被文件系统支持，你可能会丢失并没有备份的数据。最后，文件系统也不满足第(5)条。虽然它们允许几个用户或进程并发地访问文件，但文件系统一般并不阻止两个用户同时修改同一个文件，从而导致一个用户的修改不能出现在文件中。

在 DBMS 最初的重要的应用中，数据由许多小的数据项组成，对数据有许多查询或修改。这些应用的实例如下：

1. 银行系统，维护账户和确保系统故障时并不引起钱的丢失。

2. 航空公司订票系统，这些系统就像银行系统一样，要求确保数据不会丢失，并且必须能接受顾客大量的小操作。

3. 企业记录系统，雇员和税务记录、财产清单、销售记录和大量的其他类型的信息，大部分是很关键的。

早期的 DBMS 要求程序员直接面对数据的存储格式。这些数据库系统用一些不同的数据模型来描述信息在数据库中的结构，其中主要有“层次型的”或基于树的模型和基于图的“网状”模型。后者在 20 世纪 60 年代末通过了 CODASYL(数据系统和语言委员会)报告[⊖]，从而实现标准化。

这些早期的模型和系统有一个问题，就是它们并不支持高级查询语言。例如，CODASYL 查询语言具有允许用户通过数据元素间的指针从一个数据元素跳到另一个数据元素的语句。即使对于非常简单的查询，也需要相当大的工作量来写这样的程序。

1.1.2　关系数据库系统

在 1970 年 Ted Codd 发表了一篇著名的论文[⊜]后，数据库系统发生了显著的变化。Codd 认为数据库系统应该呈现给用户组织成叫做“关系”的表的数据。在幕后，应该有一个复杂的数据结构，允许对各式各样的查询进行快速响应。但是，与早期数据库系统的程序员不同的是，关系数据库的程序员并不需要关心存储结构。查询可以用很高级的语言来表达，这样可以极大地提高数据库程序员的效率。我们将在这本书的大部分章节涵盖数据库系统的关系模型，并将广泛涵盖 SQL(结构化查询语言)这一基于关系模型的最重要的查询语言。

到 1990 年，关系数据库系统已经成为标准。然而数据库领域一直在发展，而且针对数据管理的新的问题和方法也不断浮出水面。面向对象的特性渗入了关系模型。一些大型数据库已不再使用关系方法组织。后续部分将讨论数据库系统的一些新趋势。

1.1.3　越来越小的系统

最初，DBMS 是庞大的、昂贵的、在大型计算机上运行的软件系统。因为保存吉字节(GB)的数据需要大的计算机系统，所以大容量是必需的。今天，几百 GB 的数据都可以放在单个磁盘上，在个人计算机上运行 DBMS 已无任何问题。所以，基于关系模型的数据库系统甚至可以在非常小的机器上运行，而且它们也开始作为一种计算机应用的常见工具出现，如同电子表格和文字处理器一样。

另一个重要的趋势是文档的使用，经常使用 XML(可扩展的建模语言)来标记。大量小文档的集合可以作为一个数据库，而查询和操作它们的方法与关系数据库系统不同。

1.1.4　越来越大的系统

另一方面，吉字节也不再是大的数据了。公司数据库系统经常存储太字节(TB，10^{12}B)数据。还有许多数据库的数据量达到拍字节(PB，10^{15}B)，并把它们都提供给用户。一些重要的例子如下：

1. 卫星向下发送 PB 的信息，存储在特殊的系统中。

⊖ *CODASYL Data Base Task Group April* 1971 *Report*, ACM, New York.

⊜ Codd, E. F., “A relational model for large shared data banks”, *Comm.* ACM, 13:6, pp. 377-387, 1970.

2. 一张图片比 1000 个字的存储空间大。只用 5kB 或 6kB 可以存储 1000 个字，而存储一张图片需要更大的空间。一些像 Flick 这样的信息库存储了数以百万计的图片，并且提供对这些图片的搜索。甚至像 Amazon 这样的数据库也存有数以百万计的产品图片。

3. 如果静态的图片耗费空间，电影则耗费更多。一小时的视频需要至少 1GB。像 YouTube 这样的网站拥有数十万或是上百万的电影，并且能轻易地提供它们。

4. P2P 文件共享系统使用普通计算机构成的大的网络来存储和发布各种各样的数据。虽然网络中每个节点可能只存储几百 GB 的数据，但它们所存储的数据合起来非常巨大。

1.1.5 信息集成

在很大的程度上，建立和维护数据这个老问题已经变成了信息集成的问题：把许多相关的数据库所包含的信息连接成一个整体。例如，一个大公司有许多分部。每个分部可能都建有各自独立的产品记录或员工记录的数据库。这些分部中可能有一些曾经是独立的公司，它们有自己做事情的方式。这些分部可能会用不同的 DBMS 和不同的结构来存储信息，可能会用不同的术语来表示同一件事或用相同的术语来表示不同的事。更糟糕的是，因为存在使用这些不同的数据库的遗留应用程序，使得几乎不可能废弃这些数据库。

结果是，在现存数据库上建立结构体变得越来越必要，其目的是把分布在这些数据库中的信息集成起来。一种常用的方法是建立*数据仓库*(data warehouse)，将众多的遗留数据库中的信息进行适当的翻译，周期性地拷贝到一个中心数据库中。另一种方法是实现一个 Mediator 或“中间件”，它的功能是支持各个不同数据库中数据的一个集成的模型，并在这个模型和每个数据库所使用的实际模型之间进行翻译。

1.2 数据库管理系统概述

在图 1-1 中，我们可以看到一个完整的 DBMS 的轮廓。单线框表示系统成分，双线框表示内存中的数据结构。实线指明控制和数据流，虚线指明仅是数据流。由于该图很复杂，我们分几个步骤来考虑它的细节。首先，我们说对于 DBMS 有两个不同的命令来源：

1. 普通用户和应用程序，他们要求对数据进行访问或修改。

2. *数据库管理员*(database administrator，DBA)，负责建立数据库的结构或模式的一个人或一组人。

1.2.1 数据定义语言命令

第二类命令处理起来相对比较简单，我们以图 1-1 的右上部分为始点给出它的踪迹。例如，一个大学注册管理数据库的数据库管理员(或 DBA)可能决定需要一个表或关系，它的列是学生、学生选的课程和学生修该课程的成绩。DBA 可能还决定所允许的成绩只能是 A、B、C、D 和 F。此结构和约束信息都是数据库模式的一个部分。如图 1-1 所示，由 DBA 输入这些信息，DBA 需要特殊的权限才能执行模式修改命令，因为这些命令对于数据库会发生深刻的影响。这些进行模式修改的数据定义语言(DDL)命令由 DDL 处理程序进行分析，然后传给执行引擎，由执行引擎经过索引/文件/记录管理器去改变*元数据*(metadata)，即数据库的模式信息。

1.2.2 查询处理概述

与 DBMS 的大部分交互都沿着图 1-1 左侧的路径进行。用户或应用程序使用数据操纵语言(DML)启动某个活动，该活动不影响数据库模式，但是可能影响数据库内容(如果该活动是一个修改命令)，或者会从数据库中抽取数据(如果该活动是一个查询)。DML 语句被如下两个分离的