面向CS2013计算机专业规划教材

HZ BOOKS
华章教育

分布式数据库系统

大数据时代新型数据库技术

第2版

于戈 申德荣 等编著

Distributed Database Systems

Second Edition

机械工业出版社
China Machine Press

需要全本请访问 www.ertongbook.com

面向CS2013计算机专业规划教材

分布式数据库系统

大数据时代新型数据库技术

第2版

于戈 申德荣 等编著

Distributed Database Systems

Second Edition

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

分布式数据库系统：大数据时代新型数据库技术 / 于戈等编著．—2 版．—北京：机械工业出版社，2015.10
（面向 CS2013 计算机专业规划教材）

ISBN 978-7-111-51831-0

I. 分… II. 于… III. 分布式数据库－数据库系统－高等学校－教材 IV. TP311.133.1

中国版本图书馆 CIP 数据核字（2015）第 245283 号

本书主要介绍分布式数据库系统和大数据库系统的基本理论与实现技术。全书共分 12 章，第 1 章和第 2 章介绍分布式数据库系统和大数据库系统的基础和背景，主要包括系统的基本概念、体系结构、发展历史、系统分类和主要研究问题；第 3 ～ 9 章为全书的重点，介绍分布式数据库系统和大数据库系统的核心技术，包括分布式数据库设计、分布式查询处理与优化、分布式查询的存取优化、分布式事务管理、分布式恢复管理、分布式并发控制、数据复制与一致性，并给出了 Oracle 应用示例；第 10 章和第 11 章介绍两个分布式的数据管理系统案例，分别为 P2P 数据管理系统和 Web 数据库集成系统；第 12 章介绍大数据库系统研究进展及发展趋势。

本书内容新颖，理论与实践相结合，可作为计算机专业高年级本科生和研究生的教材，也可作为大数据管理和应用的研究和开发人员的参考书。

出版发行：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　刘立卿　　责任校对：殷　虹
印　　刷：三河市宏图印务有限公司　　版　　次：2016 年 1 月第 2 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：26.5
书　　号：ISBN 978-7-111-51831-0　　定　　价：55.00 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　　投稿热线：(010) 88379604
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所　韩光 / 邹晓东

前　言

数据库系统的发展起始于20世纪60年代，从IBM的层次模型IMS、网状模型、关系模型，发展到多数据模型共存。随着科学技术的发展，各个行业、领域对数据库技术提出了更多的需求，推动了数据库技术同诸多新技术如分布式处理技术、并行计算技术、人工智能技术、多媒体技术、模糊计算技术等相结合，由此衍生出了多种新的数据库技术。分布式数据库系统是其中的一种新数据库技术。分布式数据库系统兴起于20世纪70年代中期。推动分布式数据库系统发展的动力来自于两方面：一是应用需求，二是硬件环境的发展。在应用需求上，全国甚至全球范围内的航空及铁路订票系统、银行通存通兑系统、水陆空联运系统、跨国公司管理系统、连锁配送管理系统等，都涉及地理上分布的企业或机构的局部业务管理和与整个系统有关的全局管理，采用传统的集中式数据库管理系统已无法满足这种分布式应用需求。在硬件环境上，提供了功能强大的计算机和成熟的广域公用数据网及快速增长的局域网。在上述两方面的推动下，人们期望符合现实需要的、能处理分散地域的、具备数据库系统特点的新数据库系统的出现。

从20世纪70年代中期开始，各发达国家纷纷投巨资支持分布式数据库系统的研究和开发计划。历时十年，呈现出了许多研究成果。典型的原型系统有美国国防部委托CCA公司设计和研制的SDD-1分布式数据库系统、美国加利福尼亚大学伯克利分校研制的分布式INGRES系统、IBM圣何塞实验室研制的R*分布式数据库系统、德国斯图加特大学研制的Porel分布式数据库系统、法国Sirius资助计划产生的若干原型系统(如Sirius-Delta、Polypheme等)。随后，商品化的数据库系统Oracle、Sybase、DB2、Informix、INGRES等都从分布式数据库系统研究中吸取了许多重要的概念、方法和技术，实现了相当程度上的分布式数据管理功能，并宣称它们都是分布式数据库系统产品。在分布式数据库系统的商品化进程中，随着研究的深入和应用的普及，更由于分布式数据库管理系统本身的高复杂性，研究者提出了更简洁、更灵活的实现技术来满足分布式数据处理的要求。目前，商品化数据库产品如Oracle、Sybase、DB2、SQL Server、Informix都支持异构数据库系统的访问和集成功能。它们都采用基于组件和中间件的松散耦合型事务管理机制来实现分布式数据的管理，具有高灵活性和可扩展性，并且具有替代传统分布式数据库管理系统中的紧耦合型事务管理机制的趋势。

随着Internet和Web的蓬勃发展，Web环境下的分布式系统已成为当前应用的主流，如电子商务系统、网格系统、P2P共享系统等。近来，云计算、物联网等新型分布式应用的提出，更凸显了分布式数据管理的重要地位。分布式数据处理是分布式系统中必不可少的重要组成部分，涉及数据的分布式存储管理、分布式数据的查询优化、分布式事务管理与故障恢复，以及并发控制处理机制等。分布式数据库系统的概念、基本理论、算法及其

相应的技术都将对分布式数据处理以及分布式系统的研究起到重要的指导作用。并且，随着分布式计算技术和应用的发展，分布式数据管理系统的基本理论和技术将发挥越来越重要的作用。

随着技术的发展，大数据广泛存在，如 Web 数据、移动数据、社交网络数据、电子商务数据、企业数据、科学数据等，并且各行各业都期望得益于大数据中蕴含的有价值的知识。为此，呈现出了支持大数据管理和分析的技术，如大数据存储模型、键值模型、MapReduce 分布式处理架构、改进的支持分布式的事务协议、副本管理等，并推出了许多关系云系统和多存储结构的大数据库系统等。支持大数据库管理的基础理论和技术，典型代表是以经典的分布式数据库理论和技术为基础的扩展研究，满足大数据处理的实时性、高性能和可扩展性需求等。

多年来，作者在国家自然科学基金、国家 973 计划、国家 863 计划等课题的支持下，以大数据管理、Web 数据库集成、联盟企业数据集成为应用背景，针对分布式环境下的数据管理进行了深入研究。同时，作者一直承担东北大学计算机专业硕士研究生的分布式数据库系统课程以及计算机专业本科生的数据库系统概论和数据库系统实现课程的教学工作。本书正是基于以上工作而撰写的。

本书首先重点介绍经典的分布式数据库系统的基本理论和关键技术，介绍当前流行的商品化分布式数据管理机制，并进行特点分析和对比。同时，以经典的分布式数据库基本理论和技术为基础，介绍大数据库管理的关键技术和流行的大数据库系统。

本书共分为 12 章，内容包括分布式数据库系统概述、分布式数据库系统的结构、分布式数据库设计、分布式查询处理与优化、分布式查询的存取优化、分布式事务管理、分布式恢复管理、分布式并发控制、数据复制与一致性、典型的分布式数据库系统案例(P2P 数据管理系统、Web 数据库集成系统)和大数据库系统研究进展。

第 1 章主要介绍数据库基本知识、分布式数据库概念及其特性，以及分布式数据库系统的作用和特点。之后，概述大数据管理并介绍大数据库概念，主要包括大数据类型、特点、处理过程和大数据库关键技术。

第 2 章主要介绍分布式数据库系统的结构，包括分布式数据库系统的物理结构、逻辑结构、模式结构和组件结构，阐述典型的分布式数据集成系统的异同点，给出分布式数据库系统的分类。之后，介绍大数据库系统的分类、典型的体系结构和大数据库系统案例。

第 3 章主要介绍分布式数据库设计方法，包括全局关系模式的逻辑划分和实际物理分配，主要包括分片定义、分片设计和分配设计，具体包括水平分片、垂直分片和混合分片的设计。之后，介绍支持大数据库管理的存储模型、数据分布式存储策略以及大数据库存储案例。

第 4 章主要介绍分布式查询处理技术，包括查询优化的基本概念、查询处理与优化过程、查询分解、数据局部化和片段查询优化方法。之后，介绍大数据库的查询 API、查询处理和优化策略。

第 5 章主要介绍分布式查询的存取优化技术，包括存取优化的基本概念、存取优化的代价模型、典型的半连接优化技术、枚举法优化技术，以及几种典型的集中式查询优化算

需要全本请访问 www.ertongbook.com

法和分布式查询优化算法。之后，介绍大数据库管理的索引技术、缓存技术、并行处理技术。

第6章主要介绍分布式事务管理技术，包括分布式事务概念、分布式事务的实现模型、分布式事务执行的控制模型、分布式事务管理的实现模型以及分布式事务提交协议。之后，介绍大数据库的事务管理，包括大数据库管理理论、扩展的事务模型和实现方法。

第7章主要介绍分布式恢复管理技术，包括分布式数据库系统中的故障类型、集中式数据库的故障恢复方法、分布式数据库的恢复方法以及分布式数据库的可靠性协议。之后，介绍大数据库系统中的恢复管理问题、故障类型、故障检测技术和容错技术。

第8章主要介绍分布式并发控制技术，包括分布式并发控制概念及其理论基础、基于锁的并发控制方法、基于时间戳的并发控制方法、乐观的并发控制方法以及分布式死锁管理。之后，介绍支持大数据库并发控制的扩展技术。

第9章主要介绍分布式数据库的数据复制和一致性技术，包括复制策略、复制协议和一致性协议。之后，结合大数据库一致性协议介绍大数据库系统所采用的副本一致性实现策略。

第10章介绍一个典型的分布式数据库系统案例——P2P数据管理系统，包括几种典型的P2P系统的体系结构、数据管理机制以及查询处理与优化策略。

第11章介绍另一个典型的分布式数据库系统案例——Web数据库集成系统，包括典型的Web数据库集成系统的组成结构以及集成系统中的三个核心模块(搜索子系统、查询子系统和集成子系统)。

第12章介绍大数据库系统研究进展及展望，包括数据模型、基于MapReduce框架的查询处理与优化策略、事务管理技术、动态负载均衡策略、副本管理技术以及多存储模式的数据库系统。

本书由东北大学计算机科学与工程学院于戈、申德荣、赵志滨、李芳芳、聂铁铮、寇月、冯时、鲍玉斌撰写。其中，于戈负责本书前言部分，申德荣负责教学建议部分，于戈、申德荣负责第1章，赵志滨、申德荣负责第2章，申德荣、聂铁铮负责第3章，李芳芳、于戈负责第4章、第8章、第9章，聂铁铮负责第5章，寇月负责第6章和第7章，赵志滨负责第10章，申德荣、聂铁铮负责第11章，申德荣、于戈、鲍玉斌负责第12章，冯时负责各章中有关Oracle数据库的案例部分。参加本书撰写的还有博士研究生朱命冬、王习特等。全书由于戈和申德荣统稿。

我们在撰写本书的过程中，努力使本书覆盖已有分布式数据库系统的经典理论和技术，尽力跟踪该学科的新发展和新技术，尤其是用大篇幅介绍了大数据库技术，力求使本书具有先进性和实用性，并突出本书自身的特色。但由于作者学识有限，一定存在许多不足之处，敬请专家和学者批评指正。

教学建议

章　号	教学要点及教学要求	课时安排	
		计算机专业	非计算机专业
第1章 分布式数据库 系统概述	了解数据库系统的基本概念 掌握分布式数据库系统的基本概念 了解分布式数据库系统的作用和特点 了解分布式数据库系统中的关键技术 掌握大数据的基本概念 了解大数据库系统和关键技术	2～4	2 （选讲）
第2章 分布式数据库 系统的结构	掌握 DDBS 的物理结构和逻辑结构 掌握 DDBS 的体系结构、模式结构和组件结构 了解多数据库集成系统、对等型(P2P)数据库系统(P2PDBS)和 DDBS 的分类 掌握元数据的管理 了解 Oracle 系统体系结构 了解分布式大数据库管理系统的分类 掌握分布式大数据库管理系统的体系结构 了解基于 HDFS 的分布式数据库和其他分布式数据库系统	4～6 （选讲）	2～4 （选讲）
第3章 分布式 数据库设计	了解分布式数据库的设计策略 掌握分布式数据库的分片定义 掌握分布式数据库的分片设计和分配设计 了解分布式数据库的数据复制技术 了解 Oracle 数据库分布式设计 了解分布式文件系统 HDFS 和基于 SSTable 的数据存储结构 掌握大数据存储模型和数据分区策略 了解大数据库分布式存储	6～8 （选讲）	4～6 （选讲）
第4章 分布式查询 处理与优化	了解查询处理的目标和意义 掌握查询优化的基本概念和查询优化过程 了解查询处理器的特点和查询处理层次 掌握查询分解、数据局部化和片段查询优化 了解 Oracle 的分布式查询处理与优化过程 了解大数据库系统的查询 API 掌握大数据库查询处理方法 了解基于 MapReduce 的查询处理 了解大数据库查询优化	4～6 （选讲）	2～4 （选讲）
第5章 分布式查询 的存取优化	了解分布式查询的执行与处理过程和存取优化的内容 掌握存取优化的查询代价模型、数据的特征参数 掌握半连接优化技术和枚举法优化技术 了解典型的集中式数据库的查询优化算法 了解典型的分布式数据库的查询优化算法 了解 Oracle 的分布式查询优化技术 掌握布隆过滤器索引和键值二级索引方法	8～10 （选讲）	6～8 （选讲）

（续）

章　号	教学要点及教学要求	课时安排	
		计算机专业	非计算机专业
第5章 分布式查询的存取优化	了解跳跃表的实现方法 掌握分布式缓存的体系结构 了解典型的分布式缓存系统及应用	8～10 （选讲）	6～8 （选讲）
第6章 分布式事务管理	掌握事务的概念 掌握分布式事务的两段提交协议 掌握分布式事务执行的控制模型和实现模型 掌握两段提交协议的实现方法 了解非阻塞的三段提交协议 了解 Oracle 数据库的分布式事务管理技术 掌握大数据库的事务管理问题 掌握大数据库系统设计的理论基础 了解弱事务型与强事务型大数据库系统 掌握大数据库中的事务特性 了解大数据库的事务实现方法	4～6 （选讲）	2～4 （选讲）
第7章 分布式恢复管理	掌握数据库的故障模型和恢复模型 掌握集中式数据库的故障恢复技术 掌握分布式事务的故障恢复 了解可靠性和可用性的含义 了解分布式可靠性协议的组成 了解两段提交协议的终结协议以及演变过程 了解三段提交协议的终结协议以及演变过程 了解 Oracle 数据库的故障恢复技术 掌握大数据库的恢复管理问题、大数据库系统中的故障类型和大数据库系统的故障检测技术 了解基于事务的大数据库容错技术和基于冗余的大数据库容错技术	4～6 （选讲）	2～4 （选讲）
第8章 分布式并发控制	掌握并发控制的基本概念和并发控制理论 掌握基于锁的并发控制方法和两段封锁协议 掌握基于锁的分布式并发控制方法 掌握基于时间戳的分布式并发控制方法 掌握乐观的并发控制方法 了解分布式死锁等待图概念 了解几种典型的死锁的检测、预防和避免死锁的实现方法 了解 Oracle 数据库的并发控制方法 掌握大数据库并发控制技术 了解事务读写模式扩展和封锁机制扩展 掌握基于多版本并发控制扩展 了解基于时间戳并发控制扩展	4～6 （选讲）	4～6 （选讲）
第9章 数据复制与一致性	了解数据复制的作用 了解数据复制一致性模型 掌握分布式数据库复制策略 掌握 Paxos 协议、反熵协议、NWR 协议 了解向量时钟技术 了解大数据库复制一致性管理	4～6 （选讲）	2～4 （选讲）

（续）

章　号	教学要点及教学要求	课时安排	
		计算机专业	非计算机专业
第10章（选讲） P2P数据管理系统	了解P2P系统的基本概念 了解P2P系统的几种典型的体系结构 了解P2P系统中的资源定位和路由策略 了解P2P系统中的查询处理与优化策略 了解P2P系统的数据管理策略与分布式数据库管理策略的异同	0～1 （选讲）	1 （选讲）
第11章（选讲） Web数据库 集成系统	了解一个Web数据库集成系统案例 了解分布式数据库理论和技术在Web数据库集成系统中的实际应用	0～1 （选讲）	1 （选讲）
第12章(选讲) 大数据库系统 研究进展	了解有关大数据的研究进展，包括数据模型、查询处理与优化技术、事务管理、负载均衡技术、副本管理技术、典型的多存储模式的数据库系统等	0～2 （选讲）	2 （选讲）
教学总学时建议		48～72	36～54

说明：① 计算机专业本科教学使用本教材时，建议课堂授课学时数为48。不同学校可以根据各自的教学要求和计划学时数酌情对教材内容进行取舍。

② 计算机专业研究生教学使用本教材时，建议课堂授课学时数为48～72。不同学校可以根据各自的教学要求和计划学时数酌情对教材内容进行取舍。

③ 非计算机专业的师生使用本教材时可适当降低教学要求，建议课堂授课学时数为36～54。不同学校可以根据各自的教学要求和计划学时数酌情对教材内容进行取舍。

目　录

前言

教学建议

第1章　分布式数据库系统概述 …… 1

1.1　引言及准备知识 …… 1

1.1.1　相关基本概念 …… 1

1.1.2　相关基础知识 …… 3

1.2　分布式数据库系统的基本概念 …… 4

1.2.1　节点/场地 …… 4

1.2.2　分布式数据库 …… 4

1.2.3　分布式数据库管理系统 …… 5

1.2.4　分布式数据库系统应用举例 …… 5

1.2.5　分布式数据库的特性 …… 5

1.3　分布式数据库系统的作用和特点 …… 7

1.3.1　分布式数据库系统的作用 …… 7

1.3.2　分布式数据库系统的特点 …… 7

1.4　分布式数据库系统中的关键技术 …… 8

1.4.1　关键技术 …… 8

1.4.2　典型的分布式数据库原型系统简介 …… 9

1.5　大数据应用与分布式数据库技术 …… 9

1.5.1　大数据类型和应用 …… 10

1.5.2　大数据特点 …… 12

1.5.3　大数据处理过程 …… 12

1.5.4　大数据管理新模式 …… 13

1.5.5　分布式大数据库系统及关键技术 …… 14

1.6　本章小结 …… 15

习题 …… 16

主要参考文献 …… 16

第2章　分布式数据库系统的结构 …… 18

2.1　DDBS的物理结构和逻辑结构 …… 18

2.2　DDBS的体系结构 …… 19

2.2.1　基于客户端/服务器结构的体系结构 …… 19

2.2.2　基于中间件的客户端/服务器结构 …… 20

2.3　DDBS的模式结构 …… 22

2.4　DDBS的组件结构 …… 23

2.4.1　应用处理器功能 …… 23

2.4.2　数据处理器功能 …… 23

2.5　多数据库集成系统 …… 24

2.5.1　数据库集成 …… 24

2.5.2　多数据库系统 …… 26

2.6　对等型数据库系统 …… 28

2.6.1　P2PDBS的数据集成体系结构 …… 28

2.6.2　P2PDBS的体系结构 …… 29

2.6.3　P2PDBS与DDBS的典型区别 …… 29

2.7　DDBS的分类 …… 30

2.7.1　非集中式数据库系统及P2PDBS的特性 …… 30

2.7.2　DDBS的分类图 …… 31

2.8　元数据的管理 …… 32

2.8.1　数据字典的主要内容 …… 32

2.8.2　数据字典的主要用途 …… 33

2.8.3　数据字典的组织 …… 33

2.9　Oracle系统体系结构 …… 34

2.9.1 Oracle 系统体系结构简介 …… 34
2.9.2 Oracle 中实现分布式功能的关键组件 …… 34
2.9.3 Oracle 分布式数据库架构 …… 36
2.10 分布式大数据库系统 …… 37
2.10.1 分布式大数据库系统的分类 …… 37
2.10.2 分布式大数据库系统的体系结构 …… 38
2.10.3 基于 HDFS 的分布式数据库 …… 40
2.10.4 其他分布式数据库系统 …… 50
2.11 本章小结 …… 57
习题 …… 57
主要参考文献 …… 57
第 3 章 分布式数据库设计 …… 59
3.1 设计策略 …… 59
3.1.1 Top-Down 设计过程 …… 59
3.1.2 Bottom-Up 设计过程 …… 60
3.2 分片的定义及作用 …… 60
3.2.1 分片的定义 …… 60
3.2.2 分片的作用 …… 61
3.2.3 分片设计过程 …… 61
3.2.4 分片的原则 …… 62
3.2.5 分片的种类 …… 62
3.2.6 分布透明性 …… 62
3.3 水平分片 …… 62
3.3.1 水平分片的定义 …… 62
3.3.2 水平分片的操作 …… 64
3.3.3 水平分片的设计 …… 65
3.3.4 水平分片的正确性判断 …… 66
3.4 垂直分片 …… 67
3.4.1 垂直分片的定义 …… 67
3.4.2 垂直分片的操作 …… 68
3.4.3 垂直分片的设计 …… 68
3.4.4 垂直分片的正确性判断 …… 68
3.5 混合分片 …… 69
3.6 分片的表示方法 …… 69
3.6.1 图形表示法 …… 70
3.6.2 分片树表示法 …… 70
3.7 分配设计 …… 71
3.7.1 分配类型 …… 71
3.7.2 分配设计原则 …… 72
3.7.3 分配模型 …… 73
3.8 数据复制技术 …… 74
3.8.1 数据复制的优势 …… 74
3.8.2 数据复制的分类 …… 74
3.8.3 数据复制的常用方法 …… 75
3.9 Oracle 数据分布式设计案例 …… 75
3.9.1 Oracle 分布式数据库的水平分片 …… 76
3.9.2 Oracle 分布式数据库的垂直分片 …… 78
3.9.3 Oracle 集中式数据库的数据分区技术 …… 78
3.10 大数据库的分布存储策略 …… 80
3.10.1 分布式文件系统 HDFS …… 80
3.10.2 基于 SSTable 的数据存储结构 …… 86
3.10.3 大数据存储模型 …… 92
3.10.4 数据分区策略 …… 96
3.11 大数据库分布式存储案例 …… 99
3.11.1 Bigtable …… 99
3.11.2 Cassandra …… 103
3.11.3 Spanner …… 105
3.12 本章小结 …… 107
习题 …… 108
主要参考文献 …… 109
第 4 章 分布式查询处理与优化 …… 111
4.1 查询处理基础 …… 111
4.1.1 查询处理目标 …… 111
4.1.2 查询优化的意义 …… 112
4.1.3 查询优化的基本概念 …… 115
4.1.4 查询优化的过程 …… 116
4.2 查询处理器 …… 118
4.2.1 查询处理器的特性 …… 118
4.2.2 查询处理层次 …… 120
4.3 查询分解 …… 122

需要全本请访问 www.ertongbook.com

4.3.1 查询规范化 …… 122
4.3.2 查询分析 …… 123
4.3.3 查询约简 …… 124
4.3.4 查询重写 …… 125
4.4 数据局部化 …… 127
4.5 片段查询的优化 …… 128
4.6 Oracle 分布式查询处理与优化案例 …… 131
4.7 大数据库系统的查询 API …… 134
4.7.1 基于类 SQL 的查询语言 …… 134
4.7.2 基于编程接口的查询语言 …… 137
4.8 大数据库的查询处理及优化 …… 139
4.8.1 大数据库查询处理方法 …… 139
4.8.2 基于 MapReduce 的查询处理 …… 141
4.8.3 大数据库查询优化 …… 144
4.9 本章小结 …… 146
习题 …… 147
主要参考文献 …… 148
第 5 章 分布式查询的存取优化 …… 149
5.1 分布式查询的基本概念 …… 150
5.1.1 分布式查询的执行与处理 …… 150
5.1.2 查询存取优化的内容 …… 151
5.2 存取优化的理论基础 …… 152
5.2.1 查询代价模型 …… 152
5.2.2 数据库的特征参数 …… 154
5.2.3 关系运算的特征参数 …… 154
5.3 基于半连接的优化方法 …… 161
5.3.1 半连接操作及相关规则 …… 162
5.3.2 半连接运算的作用 …… 162
5.3.3 使用半连接算法的通信代价估计 …… 163
5.3.4 半连接算法优化原理 …… 164
5.4 基于枚举法的优化技术 …… 165
5.4.1 嵌套循环连接算法 …… 165
5.4.2 基于排序的连接算法 …… 167
5.4.3 散列连接算法 …… 169
5.4.4 连接关系的传输方法 …… 169
5.5 集中式系统中的查询优化算法 …… 170
5.5.1 INGRES …… 170
5.5.2 System R 方法 …… 173
5.5.3 考虑代价的动态规划方法 …… 174
5.5.4 PostgreSQL 的遗传算法 …… 177
5.6 分布式系统中的查询优化算法 …… 178
5.6.1 Distributed INGRES 方法 …… 178
5.6.2 System R* 方法 …… 183
5.6.3 SDD-1 方法 …… 184
5.7 Oracle 分布式查询优化案例 …… 192
5.8 大数据库的索引查询优化方法 …… 195
5.8.1 布隆过滤器 …… 195
5.8.2 键值二级索引 …… 197
5.8.3 跳跃表 …… 200
5.9 大数据库的查询处理与优化 …… 201
5.9.1 并行查询处理 …… 202
5.9.2 基于分析引擎的大数据库查询优化 …… 206
5.10 分布式缓存 …… 216
5.10.1 分布式缓存概述 …… 216
5.10.2 分布式缓存的体系结构 …… 217
5.10.3 典型分布式缓存系统 …… 218
5.10.4 分布式缓存与存储引擎的结合使用 …… 223
5.11 本章小结 …… 225
习题 …… 226
主要参考文献 …… 229
第 6 章 分布式事务管理 …… 231
6.1 事务的基本概念 …… 231
6.1.1 事务的定义 …… 231
6.1.2 事务的基本性质 …… 233
6.1.3 事务的种类 …… 234
6.2 分布式事务 …… 235
6.2.1 分布式事务的定义 …… 235
6.2.2 分布式事务的实现模型 …… 235

6.2.3 分布式事务管理的目标 …… 238
6.3 分布式事务的提交协议 …… 238
6.3.1 协调者和参与者 …… 239
6.3.2 两段提交协议的基本思想 …… 239
6.3.3 两段提交协议的基本流程 …… 240
6.4 分布式事务管理的实现 …… 241
6.4.1 LTM 与 DTM …… 241
6.4.2 分布式事务执行的控制模型 …… 242
6.4.3 分布式事务管理的实现模型 …… 244
6.5 两段提交协议的实现方法 …… 245
6.5.1 集中式方法 …… 245
6.5.2 分布式的 2PC …… 245
6.5.3 分层式方法 …… 246
6.5.4 线性方法 …… 247
6.6 非阻塞分布式事务提交协议 …… 248
6.6.1 三段提交协议的基本思想 …… 248
6.6.2 三段提交协议的基本流程 …… 250
6.7 Oracle 分布式事务管理案例 …… 251
6.8 大数据库的事务管理 …… 254
6.8.1 大数据库的事务管理问题 …… 254
6.8.2 大数据库系统设计的理论基础 …… 255
6.8.3 弱事务型与强事务型大数据库 …… 256
6.8.4 大数据库中的事务特性 …… 258
6.8.5 大数据库的事务实现方法 …… 260
6.9 本章小结 …… 268
习题 …… 268
主要参考文献 …… 269
第 7 章 分布式恢复管理 …… 271
7.1 分布式恢复概述 …… 271
7.1.1 故障类型 …… 271
7.1.2 恢复模型 …… 273
7.2 集中式数据库的故障恢复 …… 276
7.2.1 局部恢复系统的体系结构 …… 276
7.2.2 数据更新策略 …… 276
7.2.3 针对不同更新事务的恢复方法 …… 277
7.3 分布式事务的故障恢复 …… 278
7.3.1 两段提交协议对故障的恢复 …… 278
7.3.2 三段提交协议对故障的恢复 …… 281
7.4 分布式可靠性协议 …… 283
7.4.1 可靠性和可用性 …… 284
7.4.2 分布式可靠性协议的组成 …… 285
7.4.3 两段提交协议的终结协议 …… 286
7.4.4 两段提交协议的演变 …… 288
7.4.5 三段提交协议的终结协议 …… 288
7.4.6 三段提交协议的演变 …… 290
7.5 Oracle 分布式数据库系统故障恢复案例 …… 291
7.6 大数据库的恢复管理 …… 294
7.6.1 大数据库的恢复管理问题 …… 294
7.6.2 大数据库系统中的故障类型 …… 294
7.6.3 大数据库系统的故障检测技术 …… 295
7.6.4 基于事务的大数据库容错技术 …… 297
7.6.5 基于冗余的大数据库容错技术 …… 300
7.7 本章小结 …… 303
习题 …… 303
主要参考文献 …… 304
第 8 章 分布式并发控制 …… 306
8.1 分布式并发控制的基本概念 …… 306

8.1.1 并发控制问题 …… 306
8.1.2 并发控制定义 …… 307
8.2 并发控制理论基础 …… 308
8.2.1 事务执行过程的形式化描述 …… 308
8.2.2 集中式数据库的可串行化问题 …… 308
8.2.3 分布式事务的可串行化问题 …… 310
8.3 基于锁的并发控制方法 …… 310
8.3.1 锁的类型和相容性 …… 310
8.3.2 封锁规则 …… 310
8.3.3 锁的粒度 …… 311
8.4 两段封锁协议 …… 311
8.4.1 基本的两段封锁协议 …… 311
8.4.2 严格的两段封锁协议 …… 313
8.4.3 可串行化证明 …… 313
8.5 分布式数据库并发控制方法 …… 314
8.5.1 基于锁的并发控制方法的实现 …… 314
8.5.2 基于时间戳的并发控制算法 …… 316
8.5.3 乐观的并发控制算法 …… 318
8.6 分布式死锁管理 …… 320
8.6.1 死锁等待图 …… 320
8.6.2 死锁的检测 …… 320
8.6.3 死锁的预防和避免 …… 322
8.7 Oracle 分布式数据库系统并发控制案例 …… 323
8.7.1 Oracle 中的锁机制 …… 323
8.7.2 Oracle 中的并发控制 …… 323
8.8 大数据库并发控制技术 …… 324
8.8.1 事务读写模式扩展 …… 325
8.8.2 封锁机制扩展 …… 326
8.8.3 基于多版本并发控制扩展 …… 328
8.8.4 基于时间戳并发控制扩展 …… 331
8.9 本章小结 …… 333
习题 …… 333
主要参考文献 …… 335
第 9 章 数据复制与一致性 …… 336
9.1 数据复制的作用 …… 336
9.2 数据复制一致性模型 …… 337
9.3 分布式数据库复制策略 …… 338
9.3.1 数据复制的执行方式 …… 338
9.3.2 数据复制的实现方法 …… 339
9.3.3 数据复制的体系结构 …… 339
9.4 数据复制协议 …… 340
9.4.1 主从复制协议 …… 340
9.4.2 对等复制协议 …… 343
9.5 大数据库一致性协议 …… 345
9.5.1 Paxos 协议 …… 345
9.5.2 反熵协议 …… 346
9.5.3 NWR 协议 …… 347
9.5.4 向量时钟技术 …… 348
9.6 大数据库复制一致性管理 …… 349
9.6.1 基于 Paxos 的复制管理技术 …… 349
9.6.2 基于反熵的复制管理技术 …… 353
9.6.3 基于 NWR 的复制管理技术 …… 354
9.6.4 基于向量时钟的复制管理技术 …… 355
9.7 本章小结 …… 356
习题 …… 356
主要参考文献 …… 356
第 10 章 P2P 数据管理系统 …… 358
10.1 P2P 系统概述 …… 358
10.2 P2P 系统的体系结构 …… 359
10.2.1 集中式 P2P 网络 …… 359
10.2.2 全分布式 P2P 网络 …… 360
10.2.3 混合型 P2P 网络 …… 361
10.3 P2P 系统中的数据管理 …… 361
10.4 资源的定位和路由 …… 362
10.4.1 面向非结构化 P2P 网络的资源定位方法 …… 362
10.4.2 面向结构化 P2P 网络的资源定位方法 …… 363

10.5 处理语义异构性 …… 366
10.6 查询处理与优化 …… 367
10.6.1 查询处理 …… 367
10.6.2 查询优化 …… 368
10.7 本章小结 …… 369
习题 …… 369
主要参考文献 …… 369
第 11 章 Web 数据库集成系统 …… 371
11.1 Web 数据库集成系统概述 …… 371
11.2 三种体系结构介绍 …… 372
11.2.1 数据供应模式 …… 372
11.2.2 数据收集模式 …… 372
11.2.3 元搜索模式 …… 373
11.3 基于元搜索模式的 Web 数据库集成系统 WDBIntegrator …… 374
11.3.1 系统总体结构 …… 374
11.3.2 Web 数据库资源搜索子系统 …… 376
11.3.3 资源查询子系统 …… 377
11.4 本章小结 …… 380
习题 …… 380
主要参考文献 …… 380
第 12 章 大数据库系统研究进展 … 382
12.1 数据模型的研究 …… 382
12.1.1 支持大数据库管理的数据模型研究 …… 382
12.1.2 读写方式 …… 384
12.1.3 支持大数据库管理的分布式索引技术 …… 385
12.1.4 支持的查询 …… 387
12.2 基于 MapReduce 框架的查询处理与优化技术研究 …… 387
12.2.1 基于 MapReduce 的支持大数据处理的优化框架研究 …… 387
12.2.2 基于 MapReduce 的支持大数据计算的优化策略研究 …… 388
12.2.3 基于 MapReduce 的支持多数据集的连接查询研究 …… 389
12.2.4 MapReduce 与 NoSQL 数据库相结合的研究 …… 390
12.3 支持事务的研究 …… 391
12.3.1 应用层保证事务一致性 … 392
12.3.2 本地事务支持 …… 392
12.3.3 有限范围内的事务支持 … 392
12.3.4 弹性的事务支持 …… 392
12.3.5 面向分区数据支持分布式事务的研究 …… 393
12.3.6 异构多存储的可扩展的事务 …… 393
12.4 动态负载均衡技术的研究 …… 394
12.4.1 面向多租户的动态迁移技术 …… 394
12.4.2 面向查询处理的负载均衡技术 …… 395
12.4.3 基于中间件的面向负载的动态均衡技术 …… 395
12.5 副本管理研究 …… 396
12.5.1 自适应副本策略研究 …… 396
12.5.2 数据一致性维护策略研究 …… 396
12.5.3 多数据中心的副本一致性维护策略研究 …… 397
12.6 支持多存储模式的数据库系统 …… 398
12.6.1 支持访问多数据模式的大数据库系统 …… 398
12.6.2 自适应的多数据模式的大数据库系统 …… 399
12.6.3 支持分析型数据的分布式数据库系统 …… 399
12.7 其他研究 …… 401
12.8 总结及研究展望 …… 402
12.8.1 关键技术问题 …… 402
12.8.2 研究挑战 …… 404
习题 …… 405
主要参考文献 …… 405

CHAPTER 1

第1章 分布式数据库系统概述

1.1 引言及准备知识

分布式数据库系统(Distributed DataBase System, DDBS)是随着计算技术的发展和应用需求的推动而提出的新型软件系统。简单地说，分布式数据库系统是地理上分散而逻辑上集中的数据库系统，即通过计算机网络将地理上分散的各局域节点连接起来共同组成一个逻辑上统一的数据库系统。因此，分布式数据库系统是数据库技术和计算机网络技术相结合的产物。

分布式数据库系统与集中式数据库系统一样，包含两个重要部分：数据库和数据库管理系统。在介绍分布式数据库系统之前，先重温一下有关数据库和数据库管理系统的基本概念。

1.1.1 相关基本概念

1. 数据库

数据库(DataBase, DB)的定义有很多，从用户使用数据库的角度出发，可定义如下：数据库是长期存储在计算机内、有组织、可共享的数据集合。数据库中的数据按一定的数据模型组织、描述、存储，具有较小的冗余度、较高的数据独立性并易于扩展，同时可为各种用户共享。数据库设计就是对一个给定的应用环境(现实世界)设计出最优的数据模型，然后按模型建立数据库，见图1-1。典型的数据模型是E-R概念模型和关系数据模型。

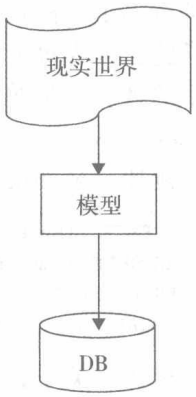


图1-1 数据库模型

2. 数据库管理系统

数据库管理系统(DataBase Management System, DBMS)是人们用于管理和操作数据库的软件，介于应用程序和操作系统之间。实际的数据库很复杂，对数据库的操作也相当烦琐，因此，为有效地管理和操作数据库，需要有数据库管理系统，使用户不必涉及数据的具体结构描述及实际存储，从而方便、最优地操作数据库。DBMS不仅具有最基本的数据管理功能，还提供多用户的并发控制、事务管理和访问控制，可保证数据的完整性和安全性，当数据库出现故障时能对系统进行恢复。数据库管理系统可描述为用户接口、查询处理、查询优化、存储管理四个基本模块和事务管理、并发控制、恢复管理三个辅助模块。其模型见图1-2。

3. 数据库系统

数据库系统(DataBase System, DBS)是指与数据库相关的整个系统。数据库系统一般由数据库、数据库管理系统、应用开发工具、应用系统和数据库管理员构成，如图1-3所示。

4. 模式

从现实世界的信息抽象到数据库存储的数据是一个逐步抽象的过程。美国国家标准协会的计算机与信息处理委员会中的标准计划与需求委员会(American National Standards Institute,

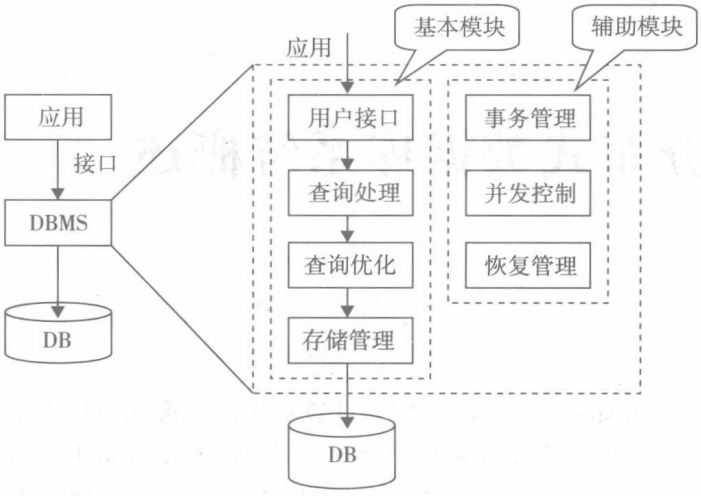


图 1-2　数据库管理系统模型

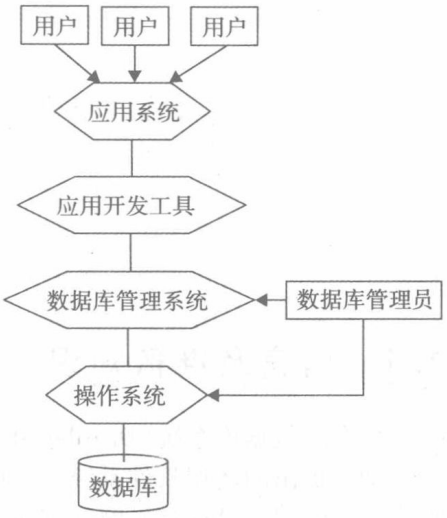


图 1-3　数据库系统组成

Standards Planning And Requirements Committee，ANSI-SPARC）根据数据的抽象级别为数据库定义了三层模式参考模型，如图 1-4 所示。

外模式是数据库用户和数据库系统的接口，是数据库用户的数据视图（view），是数据库用户可以看见和使用的局部数据的逻辑结构和特征的描述，是同应用有关的数据的逻辑表示。一个数据库系统通常有多个外模式。外模式是保证数据库安全的重要措施，因为每个用户只能看见和访问特定的外模式中的数据。通常，由 DBMS 中的视图定义命令（create view）定义数据库的外模式。

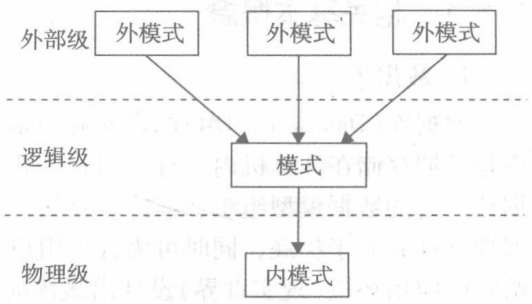


图 1-4　ANSI-SPARC 三层模式参考模型

例如，某外模式定义如下：

```
CREATE VIEW PAYROLL(EMP_ENO,EMP_NAME,SAL)
AS SELECT EMP.ENO,EMP.NAME,PAY.SAL
FROM EMP,PAY
WHERE EMP.TITLE = PAY.TITLE
```

模式是关于数据库中全体数据的逻辑结构和特征的描述，是所有用户的公共数据视图。模式是数据库中数据在逻辑级上的视图。一个数据库只有一个模式。模式以某种数据模型为基础，综合考虑了所有用户的需求，并将这些需求有机地结合成一个逻辑整体。定义模式时不仅要定义数据的逻辑结构，如组成关系模式的属性名、属性的类型、取值范围，还要定义属性间的关联关系、完整性约束等。模式由 DBMS 中提供的模式描述语言定义。

例如，某模式定义如下：

```
RELATION EMP{
    KEY = {ENO}
ATTRIBUTE = {
        ENO:CHAR(9)
        ENAME:CHAR(15)
        TITLE:CHAR(10)
        }
}
```

需要全本请访问 www.ertongbook.com