HZ BOOKS
华章教育

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

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

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

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

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。

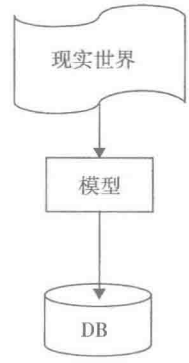

图 1-1 数据库模型

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,

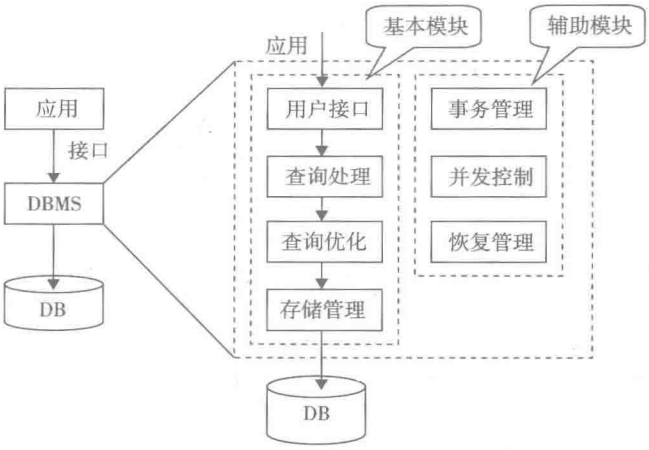

图 1-2　数据库管理系统模型

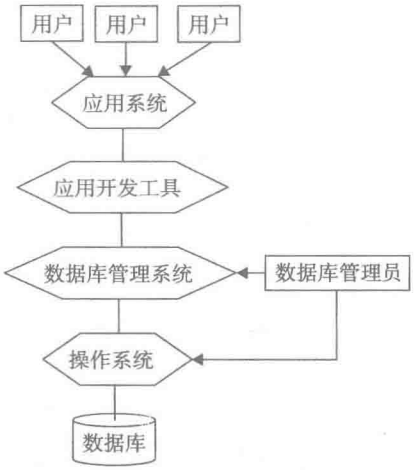

图 1-3　数据库系统组成

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。

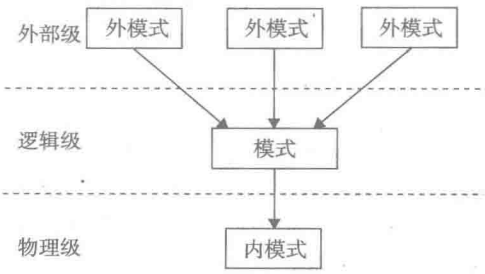

图 1-4　ANSI-SPARC 三层模式参考模型

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)
        }
}
```