重点大学计算机教材

HZ BOOKS
华章教育

分布式数据库系统
原理与应用

申德荣　于戈　等编著
东北大学

机械工业出版社
China Machine Press

重点 机教材

分布式数据库系统原理与应用

申德荣 于戈 等编著
东北大学

机械工业出版社
China Machine Press

本书主要介绍分布式数据库系统的理论与实现机制方面的有关原理和方法。全书共分十章，第1章和第2章介绍分布式数据库系统的基础和背景，主要包括分布式数据库系统的基本概念、体系结构、发展历史和主要研究的问题；第3～8章为全书的重点，介绍分布式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、分布式恢复管理和分布式并发控制；第9章和第10章分别介绍P2P数据管理系统和Web数据库集成系统这两个分布式的数据管理系统案例。

本书是在作者长期的教学和科研基础上，结合分布式数据库基本原理及实际应用技术编写而成的。本书不仅介绍经典的分布式数据库理论和技术，还以流行的商用数据库Oracle为例介绍相关实现技术，以及特定领域的分布式数据管理系统案例。

本书内容新颖，理论与实践相结合，适合作为计算机专业以及相关专业的研究生或高年级本科生的教材，也适合作为数据库开发人员的参考书。

图书在版编目(CIP)数据

分布式数据库系统原理与应用/申德荣，于戈等编著. —北京：机械工业出版社，2011.6
(重点大学计算机教材)

ISBN 978-7-111-34524-4

Ⅰ. 分…　Ⅱ. ①申…　②于…　Ⅲ. 分布式数据库－数据库系统－研究生－教材
Ⅳ. TP311.133.1

中国版本图书馆CIP数据核字(2011)第082550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：李　荣
北京市荣盛彩色印刷有限公司印刷
2011年7月第1版第1次印刷
185mm×260mm · 17印张
标准书号：ISBN 978-7-111-34524-4
定价：35.00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

前 言
Preface

数据库系统的发展起始于19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的数据库技术。分布式数据库系统兴起于19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实现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的数据库系统的出现。

从19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。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，研究者们提出了更简洁、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。

作者多年来在国家自然科学基金、国家863等课题的支持下，以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软件专业硕士研究生的分布式数据库系统课程和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基础而撰写的。

本书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，同时也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。

本书共分为十章，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、分布式并发控制和典型的分布式数据库系统案例(P2P数据管理系统和Web数据库集成系统)。

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。

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。

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。

第4章主要介绍分布式查询处理与优化技术，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。

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。

第6章主要介绍分布式事务管理技术，包括分布式事务的概念、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。

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。

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的概念及其理论基础、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。

第9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10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两个核心模块(搜索子系统和查询子系统)。

本书由东北大学信息科学与工程学院计算机软件研究所于戈、申德荣、聂铁铮、寇月、李芳芳、赵志滨、冯时撰写。其中，于戈、申德荣负责本书前言部分、第1章及第2章，申德荣、赵

志滨负责第3章，李芳芳负责第4章和第8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9章，申德荣、聂铁铮负责第10章，各章Oracle案例均由冯时负责。研究生单菁、杨丹、朱命冬、王习特等也参与了本书的审校。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，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，本书不足之处在所难免，敬请专家和学者批评指正。

教学章节	教学要求	课　时
第 1 章 分布式数据库系统概述	掌握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	4
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 了解对等型(P2P)数据库系统(P2PDBS) 了解 DDBS 的分类 掌握元数据的管理 了解 Oracle 系统体系结构	4 ~ 6
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 数据分布式设计	4 ~ 6
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	4
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、数据库的特征参数 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	8 ~ 12
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分布式事务管理实现模型 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	4 ~ 6
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	4 ~ 6

（续）

教学章节	教学要求	课　时
第7章 分布式恢复管理	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	4～6
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 了解基于时间戳的分布式并发控制算法 了解乐观的并发控制算法 掌握分布式死锁等待图概念 掌握典型的集中死锁的检测、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	4～6
第9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	2～4
第10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	2～4
教学总课时建议		40～58

说明：建议总课时为40～58学时，各学校可以根据自己的教学要求和计划学时数对教学内容进行取舍。

目 录
Contents

第1章
Chapter 1

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。

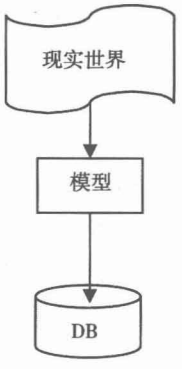

图1-1 数据库模型

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繁琐，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。

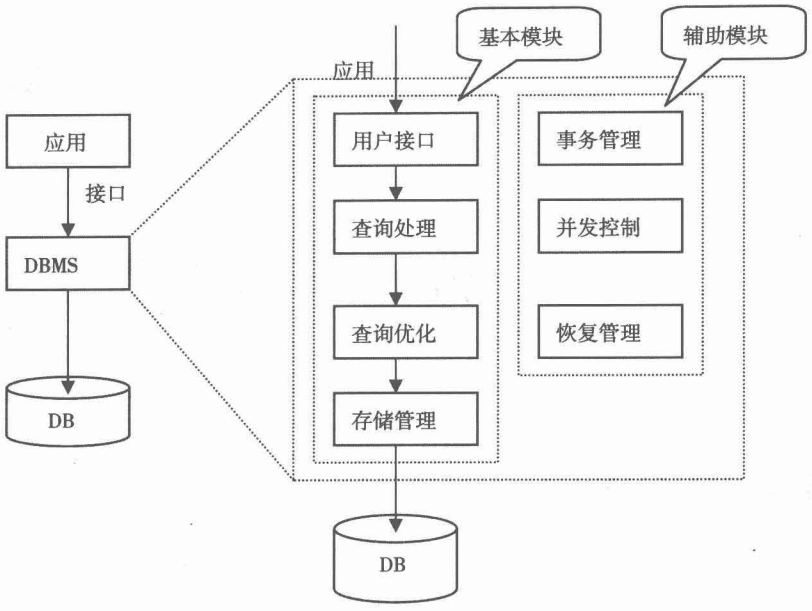

图 1-2 数据库管理系统模型

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, 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。

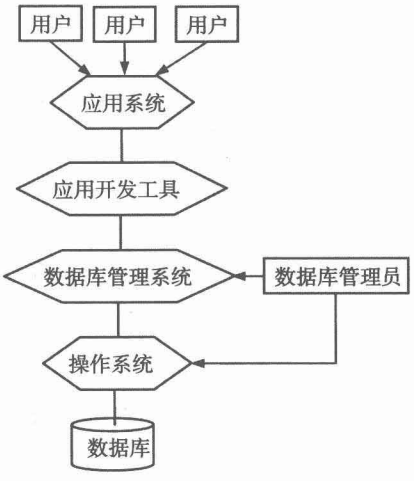

图 1-3 数据库系统组成

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。

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
```

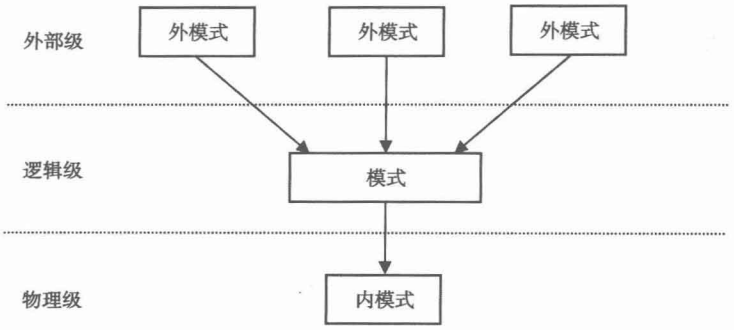

图 1-4　ANSI-SPARC 三层模式参考模型

```
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)
        }
}

RELATION PAY{
    KEY = {TITLE}
ATTRIBUTE = {
        TITLE:CHAR(10)
        SAL:NUMBER(5)
        }
}
```

内模式是关于数据物理结构和存储方式的描述，是数据在数据库内部的表示方式。一个数据库只有一个内模式，如关系表的存储方式是按照堆存储还是按照属性值聚簇存储，索引是采用B+树索引，还是采用散列索引等。内模式由 DBMS 中提供的内模式描述语言定义。

例如，某内模式定义如下：

```
INTERNAL - RELA {
    INDEX ON E# CALL EMINX
    FIELD = {
        E#:BYTE(9)
        ENAME:BYTE(15)
        TITLE:BYTE(10)
        }
}
```

1.1.2 相关基础知识

在后面章节的学习中，将涉及关系模型、关系代数和 SQL 语言知识。下面将简单介绍这些知识。

1. 关系模型

关系模型是数据库数据模型的三种模型（层次数据模型、网状数据模型和关系数据模型）之一。关系是二维表，也称为表。表中的一行称为关系的一个元组，表中的一列称为关系的一个属性。

2. 关系代数

关系是一个集合，关系的元组是集合的元素。常见的关系代数包括 5 个集合运算和 3 个关系运算。

5 个集合运算为：

- **并（union）运算**：设有两个关系 R 和 S，具有相同的关系模式，R 和 S 的并运算的结果是由两个关系中所有元组组成的一个新关系，记为 $R \cup S$ 或 $R+S$。
- **交（intersect）运算**：设有两个关系 R 和 S，具有相同的关系模式，R 和 S 的交运算的结果是由两个关系中所有公共元组组成的一个新关系，记为 $R \cap S$。
- **差（difference）运算**：设有两个关系 R 和 S，具有相同的关系模式，R 和 S 的差运算结果是由属于关系 R 但不属于关系 S 的元组组成的一个新关系，记为 $R-S$。
- **乘（product）运算**：设 R 有 m 个属性，S 有 n 个属性，R 有 i 个元组，S 有 j 个元组，R 和 S 的乘（笛卡儿积）的运算结果是由 $(m+n)$ 个属性、$i \times j$ 个元组组成的一个新关系，每个元组的前 m 个分量（属性值）来自 R 的一个元组，后 n 个分量来自 S 的一个元组，记为 $R \times S$。
- **除（divied）运算**：设有关系 $R(X, Y)$ 和 $S(Y, Z)$，其中 X，Y，Z 为属性组，R 中的 Y 和 S 中的 Y 可有不同的属性名，但必须出自相同的值域。R 和 S 的除运算得到一个新关系 $P(X)$，P 是 R 中满足下列条件的元组在 X 属性上的投影：元组在 X 上分量值 x 的象集 Y_x 包含 S 在 Y 上投影的集合，记为 $R \div S$。

3 个关系运算为：

- **选择（select）运算**：选择运算是从指定的关系 R 中选择满足条件（条件表达式）的元组组成的一个新关系，记为 $\sigma_{<条件表达式>}(R)$。
- **投影（project）运算**：投影运算是从指定的关系 R 中选择属性集 A 的所有值组成的一个新关系，记为 $\prod_A(R)$。
- **连接（join）运算**：连接运算包括 θ 连接、等值连接和自然连接三种运算，θ 是算术比较符。设有关系 $R(A, B)$ 和 $S(C, D)$，A 和 C 出自于同一值域，R 和 S 的 θ 连接运算是由两个关系 R 和 S 中满足 $A\theta C$ 连接条件的元组连接在一起组成的一个新关系，记为 $R \underset{A\theta C}{\infty} S$。若 θ 是等号“=”，该连接操作称为等值连接，记为 $R \underset{A=C}{\infty} S$。若 A 和 C 具有相同的属性名，R 和 S 的连接运算默认按 $A=C$ 连接条件进行连接，并去除重复列属性，则为自然连接运算，记为 $R \infty S$。

3. SQL 语言

SQL(Structured Query Language)是一种非过程性语言，提供了数据定义(建立数据库和表结构)、数据操纵(输入、修改、删除、查询)、数据控制(授予、回收权限)等数据库操作命令，较好地满足了数据库语言的要求。美国国家标准局(ANSI)与国际标准化组织(ISO)制定了 SQL 标准，相继推出了 SQL/86、SQL/92、SQL/99、SQL/2003、SQL/2006 等。SQL 提供了灵活而强大的查询功能，具有可移植性。SQL 已为广大用户所采用，成为用户访问数据库系统的标准接口语言。

1.2　分布式数据库系统的基本概念

1.2.1　节点/场地

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。管理分布式数据库的软件称为分布式数据库管理系统。分布式系统通常是由计算机网络将地理上分散的各逻辑单位连接起来而组成的。被连接的逻辑单位称为**节点**(node)或**场地**(site)。节点或场地是指物理上或逻辑上的一台计算机(如集群系统)。节点强调的是计算机和处理能力，场地强调的是地理位置和通信代价，二者只是看问题的角度不同，本质上没有区别。

1.2.2　分布式数据库

分布式数据库(Distributed DataBase，DDB)是分布在一个计算机网络上的多个逻辑相关的数据库的集合。也就是说，分布式数据库是一组结构化的数据集合，逻辑上属于同一系统，物理上分布在计算机网络的各个不同的场地上，如图 1-5 所示。

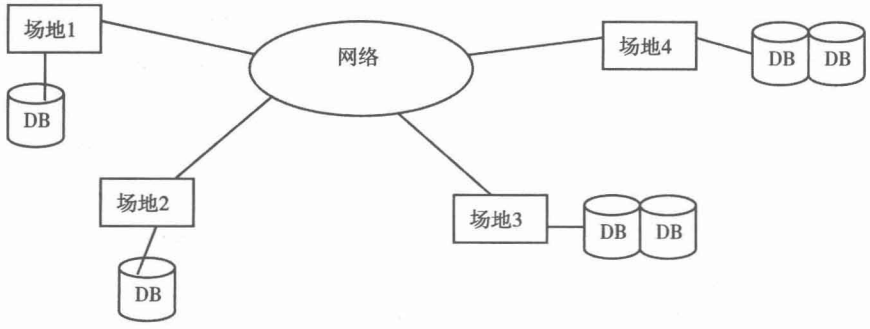

图 1-5　分布式数据库系统

为了与分布式数据库相区别，将传统的单场地上的数据库称为集中式数据库。

1.2.3　分布式数据库管理系统

分布式数据库系统由分布式数据库和分布式数据库管理系统(Distributed DataBase Management System，DDBMS)组成。分布式数据库管理系统是分布式数据库系统的一组软件，负责对分布式数据库中的数据进行管理和操作。由于分布式数据库管理系统基于分布式环境实现，因此必须保证