“十三五”普通高等教育应用型规划教材

数据库应用技术

张延松　编著

DATABASE
APPLICATION TECHNOLOGY

APPLICATION ORIENTED

中国人民大学出版社

“十三五”普通高等教育应用型规划教材

数据库应用技术

张延松　编著

DATABASE
APPLICATION TECHNOLOGY

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

数据库应用技术/张延松编著. —北京：中国人民大学出版社，2016.4
“十三五”普通高等教育应用型规划教材
ISBN 978-7-300-22818-1

Ⅰ.①数… Ⅱ.①张… Ⅲ.①数据库系统-高等学校-教材 Ⅳ.①TP311.13

中国版本图书馆 CIP 数据核字（2016）第 083334 号

“十三五”普通高等教育应用型规划教材

数据库应用技术

张延松 编著

Shujuku Yingyong Jishu

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511770（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京密兴印刷有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2016 年 4 月第 1 版
印　　张	21.25　插页 1	**印　　次**	2016 年 4 月第 1 次印刷
字　　数	508 000	**定　　价**	38.00 元

前　言

数据库主要面向结构化数据管理，是计算机系统重要的系统软件之一，也是现代信息社会重要的支撑技术之一，是现代信息社会运行的基础性技术，是数据管理和数据分析处理的平台。数据库是一门系统科学，有独立的理论基础和成熟的应用技术，掌握数据库系统理论知识和实践技能是从事企业数据分析重要的基础。随着数据库技术在企业中的广泛应用，数据库分析处理成为企业数据分析的最重要功能，也为其他数据分析处理工具提供了良好的数据支持与服务。近年来，随着大数据分析处理需求的不断增长，企业级数据分析处理技术越来越成为数据分析处理的主要任务，这意味着具有不同知识背景的数据分析人员需要直接面对企业级数据平台，需要掌握足够的数据库知识来实现企业级数据分析处理任务。数据仓库是面向分析型应用的数据库应用技术，它以数据库系统所积累的大量业务数据为基础，通过数据仓库特有的存储体系结构对数据按分析主题进行整合，面向分析处理进行存储、查询优化设计，在企业级海量数据的基础上为决策分析提供联机分析、数据挖掘等功能，从企业海量数据中分析出有价值的信息，支持企业决策制定，提供商业智能支持。

本书面向数据库分析处理应用技术，以案例教学的方式系统地介绍数据库基本理论、数据仓库基本理论，以及基于 SQL Server 2012 数据库平台的数据库分析处理应用案例，实现数据库基本理论与数据库分析处理实践技术相结合，引导读者学习使用数据库平台的各种工具完成完整的数据分析处理任务，提高读者使用数据库工具进行企业级数据分析处理的能力。

本书主要介绍数据库系统的基本原理、SQL 命令操作实践、数据库应用技术、数据仓库和联机分析处理（On-Line Analytical Processing，OLAP）以及 OLAP 实践等基础性的理论知识及操作技能，实现数据库基本理论与数据库分析处理实践技术相结合，引导读者学习使用数据库平台的各种工具完成完整的数据分析处理任务，提高读者使用数据库工具进行企业级数据分析处理的能力。书中 FoodMart 数据文件下载地址为 https://sourceforge.net/projects/mondrian/files/mondrian/mondrian-3.7.0/mondrian-3.7.0.0-752.zip/download。具体见所下载的压缩文件的 demo\access 目录。SSB 数据生成器

dbgen 下载地址为 http://www.cs.umb.edu/~xuedchen/research/publications/。TPC-H 数据生成工具下载地址为 http://www.tpc.org/tpc_documents_current_versions/current_specifications.asp。

目　录

第1章

数据库基础知识

本章要点与学习目标

数据库是数据管理技术与计算机技术相结合的研究领域，主要面向海量数据的管理及处理技术，是现代信息系统的核心和基础性技术。当前主流的数据库是关系数据库，即采用关系模型存储数据，使用关系操作执行查询处理任务，主要面向结构化的海量数据存储与管理。随着信息技术应用领域的不断拓展，关系数据库不仅支持传统的结构化数据处理，还逐渐扩展了对半结构化 XML 数据和非结构化数据的处理能力，在大数据时代与 NoSQL（非关系型数据库）技术相结合，不断扩展其处理能力。因此，数据库技术不仅是当前企业级大数据最重要的平台，同样在新兴的大数据处理平台上发挥着重要的作用，数据库技术与 Hadoop 平台的结合成为数据库技术发展的新趋势，掌握现代数据库技术能够为大数据分析处理技术打下坚实的理论基础，有助于更深入地理解当前大数据分析技术的技术路线及未来发展趋势。

本章学习目标是掌握数据库的基本概念，理解关系模型和关系操作，学习使用关系数据库标准语言 SQL 实现数据管理和查询处理。

第 1 节　数据库基本概念

数据库中最常用的术语和基本概念包括数据、数据库、数据库管理系统、数据库系统等，这些基本概念从不同的角度描述了数据管理与处理的不同层面。

一、数据、数据库、数据库管理系统、数据库系统

1. 数据

数据（data）是数据库存储和数据处理的基本对象。在维基百科中数据的定义为[1]：

数据是构成信息的一组定性或定量的描述客观事实的值的集合。数据在计算时表现为多种形式，如表格（由行和列组成）、树形结构（tree）或图（graph），数据是以图形、声音、文字、数、字符和符号等形式对事实描述的结果，通常可以通过表格、图或图像等形式展示给用户。

狭义地讲，数据是计算机对现实世界事实或实体的描述方式，包括数据形式、数据结

构和数据语义。数据形式指计算机支持的数据类型，如整型（int）、浮点型（float，double）、日期型（2014-03-19）、字符型（‘中国’）、逻辑型（True/False）、扩展数据类型（xml，binary，image等存储半结构化和非结构化文件的数据类型）。数据结构是将不同类型的数据按一定的结构组织起来表示实体或事务，如（1，110，High Roller Savings，Product Attachment，14435，1996-01-03 00：00：00.000，1996-01-06 00：00：00.000）表示一个促销记录在（promotion _ id，promotion _ district _ id，promotion _ name，media _ type，cost，start _ date，end _ date）结构上各个分量的数据。数据语义则是对数据含义的说明，如数据结构的 promotion _ name 部分表示促销名称，start _ date 表示促销开始日期，end _ date 表示促销结束日期等，数据语义定义了实体描述信息与计算机存储数据形式和结构之间的映射关系。

在数据库中，数据通常表示为由一定格式的数据项所组成的结构化的数据形式，通常称为记录。

2. 数据库

数据库（DataBase，DB）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、描述和存储，具有较小的冗余度、较高的数据独立性和易扩展性，并可为各种用户共享。整个数据库在建立、运用和维护时由数据库管理系统（DBMS）统一管理、统一控制。用户能方便地定义数据和操纵数据，并保证数据的安全性、完整性和多用户对数据的并发使用及发生故障后的数据库恢复。数据库是数据库系统的一个重要组成部分。

数据库按数据模型分，可分为层次数据库、网状数据库、关系数据库、面向对象数据库以及近年来出现的面向非结构化数据的以 key/value 存储为特点的 NoSQL 数据库。

数据库技术与其他学科的技术内容结合，出现了各种新型数据库：

- 数据库技术与分布处理技术结合➔分布式数据库；
- 数据库技术与并行处理技术结合➔并行数据库；
- 数据库技术与人工智能结合➔演绎数据库和知识库；
- 数据库技术与多媒体技术结合➔多媒体数据库；
- 数据库技术与特定的应用领域相结合，出现了工程数据库、地理数据库、统计数据库、空间数据库等特定领域数据库。

数据库中存储的数据不仅包括表示实体信息的数据，还包括表示实体之间的联系的数据，如（369，2，5，1191，1.52，2）表示在日期 ID 为 369 时，ID 为 2 的买家从 ID 为 5 的供应商购买了 ID 为 1191 的产品，其单价为 1.52 元，数量为 2，这种统一的数据结构描述了现实世界买家、卖家、产品、日期实体之间的销售联系数据。总体来说，数据库中的数据包括：数据本身、元数据（即对数据的描述）、数据之间的联系和数据的存取路径。数据库中的数据是整体结构化的，数据不再面向某一程序而组织，能够被不同的用户及应用程序通过统一的接口访问，从而大大降低数据冗余存储的代价，减少了数据之间的不一致性问题。

数据库是现代信息技术的数据基础，随着近年对大数据的关注不断升温，以数据为中心的新的应用模式不断拓展数据库应用的广度和深度。随着数据库技术应用领域的不断扩展，数据库中数据的类型由传统意义的数字、字符发展到文本、声音、图形、图像等多种类型，从结构化数据处理扩展到半结构化、非结构化数据处理领域，从传统的数据库平台扩展到新兴的数据库平台，应用领域从传统的面向商业与事务处理扩展到科学计算、经济、社会、移

动计算等各个领域，从事务处理走向分析处理，从数据库系统平台走向云计算平台。

3. 数据库管理系统

数据库管理系统（DataBase Management System，DBMS）是用于建立、使用和维护数据库的软件。它是位于用户和操作系统之间的数据管理软件，用于对数据库进行统一的管理和控制，保证数据库的安全性和完整性，提供给用户访问数据库、操纵数据、管理数据库和维护数据库的用户界面。数据库管理系统的主要功能包括以下几个方面：

（1）数据定义。

数据库管理系统提供数据定义语言（Data Definition Language，DDL），用户通过 DDL 对数据库中的对象进行定义，包括数据库中的表、视图、索引、约束等对象。

（2）数据组织、存储和管理。

数据组织和存储的目标是提高存储空间利用率，提供方便的存储接口，提供多种存储方法（如索引查找、哈希查找、顺序查找等）提高存取效率。数据的组织与存取提供数据在外部存储设备上（如磁盘、SSD 固态硬盘等）的物理组织与存取方法，涉及三个方面：1）提供与操作系统特别是与文件系统的接口，包括数据文件的物理存储组织（行存储、列存储或混合存储）及内外存数据交换方式等；2）提供数据库的存取路径及更新维护等功能；3）提供与数据库描述语言和数据库操纵语言的接口，包括对数据字典的管理等。

（3）数据操纵功能。

数据库管理系统通过数据操纵语言（Data Manipulation Language，DML）来操纵数据，支持交互式查询处理，如查询、插入、删除、修改等操作，并将查询结果返回用户或应用程序。

（4）数据库事务管理和运行管理。

数据库管理系统提供事务运行管理及运行日志，事务运行的安全性监控和数据完整性检查，事务的并发控制及系统恢复等功能，保证数据库系统的安全性和完整性、多用户对数据的并发访问控制及数据库发生故障后的系统恢复等机制。

（5）数据库维护。

数据库维护为数据库管理员提供数据加载、数据转换、数据库转储、数据库恢复、数据安全控制、完整性保障、数据库备份、数据库重组以及性能监控等维护工具。

（6）其他数据库功能。

数据库管理系统提供的功能还包括数据库与应用软件的通信接口、不同数据库系统之间的数据转换、异构数据库互访及互操作等功能。

基于关系模型的数据库管理系统已经成为数据库管理系统的主流技术。随着新型数据模型及数据管理实现技术的推进，DBMS 软件的性能还将进一步更新和完善，应用领域也将进一步拓展。

4. 数据库系统

数据库系统（DataBase System，DBS）是存储、管理、处理和维护数据的软件系统，是在计算机系统中引入数据库后的系统，包括数据库、数据库管理系统、数据库开发工具、应用系统、数据库管理员等。它由数据库、数据库管理员和有关软件组成。这些软件包括数据库管理系统（DBMS）、宿主语言、开发工具和应用程序。DBMS 用于建立、使用和维护数据库。宿主语言是可以嵌入数据库语言的程序设计语言。数据库是长期存储在计算机中有组织的、大量的和可共享的数据集合。数据库管

理员负责创建、监控和维护数据库。

数据库系统的发展主要以数据模型和 DBMS 的发展为标志。数据库诞生于 20 世纪 60 年代中期。第一代数据库系统以层次和网状数据模型的数据库系统为特征，代表性的数据库系统是 1969 年美国 IBM 公司研制的层次数据库系统 IMS 和美国数据系统语言会议（CODASYL）的数据库任务组（DataBase Task Group，DBTG）提出的 DBTG 报告所确定的网状模型数据库系统。第二代数据库系统是指关系数据库系统，其代表性事件是 1970 年 IBM 公司 San Jose 研究所的 E. F. Codd 发表的题为“大型共享数据库的关系模型”的论文，开创了关系数据库系统方法和理论的研究。20 世纪 90 年代随着面向对象、人工智能和网络等技术的发展，产生了面向对象数据库系统和演绎数据库系统。近年来随着数据库应用领域的拓展，在 WEB 数据管理和生物数据管理等应用的推动下，半结构化和非结构化 NoSQL 数据库成为主要的发展方向，在当前大数据应用背景下，数据库概念也逐渐从关系数据库平台扩展到大规模分布式计算平台，出现了以 NewSQL 为代表的各种新的可扩展/高性能数据库。这类数据库不仅具有 NoSQL 对海量数据的存储管理能力，还保持了传统数据库支持 ACID 和 SQL 等特性，如基于分布式集群的 Google Spanner，VoltDB 等系统，基于高扩展性 SQL 存储引擎的 MemSQL 等系统，基于分片的中间件层的数据库 ScaleBase 等系统。随着硬件技术的发展，高可扩展性、高性能的数据库是未来数据库发展的主要趋势。

二、数据库系统的特点

1. 数据结构化

数据库的主要特征是整体数据结构化，不仅数据内部是结构化的，而且数据之间也要遵循一定的结构要求，即数据之间具有逻辑联系。

数据内部结构化是指数据库的数据文件由记录构成，每个记录由若干属性组成，如图 1—1 中的 supplier 表中的记录由 s_suppkey，s_name，s_address，s_city，s_nation，s_region，s_phone 属性组成，每个属性具有不同的数据类型、格式和语义，构成了描述 supplier 实体的数据分量。同理，表 part，lineorder，date，customer 中的记录都是结构化的数据。

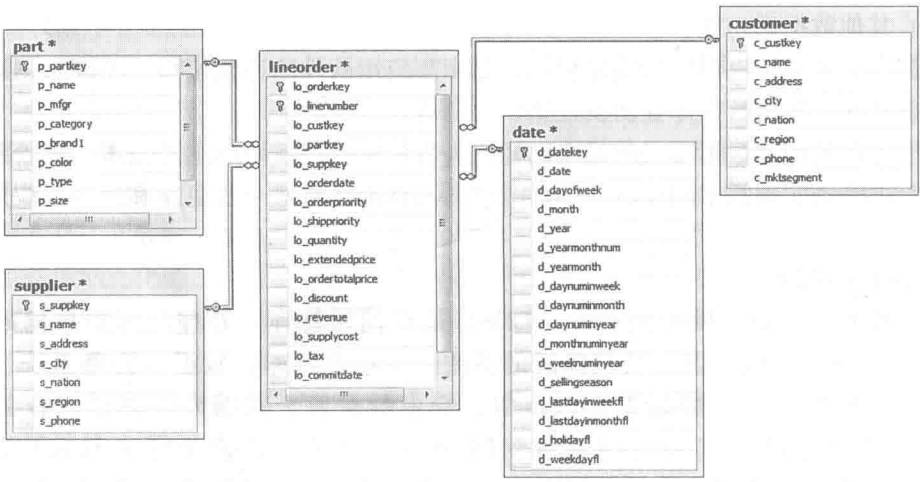


图 1—1　结构化数据

数据之间的结构要求体现在不同表的记录之间的逻辑联系。例如，表 lineorder 代表订单信息，其中 lo_custkey 代表订单的 customer ID，订单必须满足订单的 lo_custkey 在 customer 表的 c_custkey 中存在且唯一存在的约束条件才能保证订单数据的合法性和正确性。因此，数据库需要通过参照完整性来保证 lineorder 数据与 customer 数据之间的联系，支持整体结构化。

数据库不是一个将数据堆积在一起的仓库，而是要通过结构化过程按业务和分析需求抽取出现实世界实体的共性属性，通过结构化数据设计抽取出实体的共性属性用于描述整体数据特征。通过定义数据之间的关系，一方面保证了在事务处理时数据的正确性和合法性，另一方面也定义了数据分析处理时数据的相关性，定义了表间操作的类型（如参照完整性约束定义了表之间使用等值连接操作）。

2. 数据共享性高、冗余度低，系统易扩充

数据库中的数据不是面向某个应用定制，而是面向整个系统应用，可以被多个用户、多个应用共享使用，能够从应用软件中独立出来，成为基础的数据库平台。数据库中的数据面向共享访问，减少不同应用访问数据库时的数据复本，避免了复本不一致的问题。在数据库的结构设计上采用的模式优化技术能够优化数据库结构，减少冗余数据存储代价，提高存储效率。

数据库的数据在设计上面向整个系统，可以被多个应用共享使用，容易增加新的应用。结构化的设计方法通过增加新的数据，扩展数据之间的联系，在保持整体结构不变的前提下扩充新的数据和新的应用。

3. 数据独立性强

数据库的独立性包括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存储在磁盘中的数据库的数据相互独立，数据库系统负责数据的存储和访问，应用程序通过数据库访问接口访问数据，并不直接操纵物理存储的数据。当数据的物理存储结构变化时并不影响应用程序的数据访问，从而保证了应用程序良好的平台适应性。

数据的逻辑独立性是指应用程序与数据库的逻辑结构相独立，数据库提供给应用程序数据访问视图，数据库维护数据访问视图与数据库内部逻辑结构的映射关系，当数据库的逻辑结构改变时，通过更新数据访问视图与数据库内部逻辑结构映射关系来保证数据访问视图的稳定性。

在数据库的优化技术中，数据的物理存储结构和逻辑存储结构可能会发生改变，如从行存储转换为列存储，选择不同的存储引擎或者修改数据库的模式结构，数据库通过二级映像功能来保证当数据库的物理存储结构和逻辑存储结构发生变化时应用程序保持不变。

随着近年来硬件技术的突飞猛进，数据库系统与硬件技术相结合是当前数据库系统发展的主要趋势，数据库技术与最新的多核处理技术、众核处理技术、内存存储技术、flash（闪存）存储技术、GPU（图形处理器）处理技术、高速网络技术、云计算等新兴技术相结合，在大数据分析处理领域发挥着越来越重要的作用。

第 2 节　关系数据模型

数据模型（data model）是对现实世界数据特征的抽象，用于描述数据、组织数据和对数据进行操作。数据模型可以分为概念模型和逻辑模型两大类：在数据库中广泛使用的概念模型是实体联系模型，用于描述现实世界的数据结构；数据库的逻辑模型包括层次模型、网状模型、关系模型、面向对象模型以及对象关系模型等，当前应用最为广泛的是关系模型。

一、实体-联系模型

实体-联系模型（entity-relationship model）是通过实体型及实体之间的联系型来反映现实世界的一种数据模型，又称 E-R 模型。实体-联系模型是由 Peter P. S. Chen 于 1976 年提出的，广泛适用于软件系统设计过程中的概念设计阶段。

实体-联系模型的基本语义单位是实体和联系。

实体（entity）是代表现实世界中客观存在的并可以相互区别的事物。实体可以是具体的人或事物，如客户、供应商、产品，也可以是抽象的概念或度量，如日期等。

属性（attribute）是实体的某一种可以数据化的特征。一个实体由若干属性来表示，每个属性对于该实体有一个数据取值，这些取值用于区分该实体与其他实体。如客户实体的属性包括客户 ID、客户姓名、客户地址、客户电话、客户所在地区等，日期实体的属性包括年、月、日、季度、周等，实体的属性组合起来能够表示一个实体的特征，属性也定义了未来用于分析和处理的数据结构。

实体型（entity type）由实体名及相应的属性名集合构成。属性是描述实体共同特征和性质的数据，实体型则定义了描述相同类型实体的公共数据结构。例如，客户（客户 ID、客户姓名、客户地址、客户电话、客户所在地区）、日期（年、月、日、季度、周）分别是表示客户实体和日期实体的实体型。

实体集（entity set）指同一类型实体的集合。如客户、日期等都是实体集。

联系（relationship）指实体内部属性之间或者实体之间的联系。实体内部属性之间的联系包含属性之间的函数依赖关系，实体之间的联系包含实体集之间一对一联系（1∶1）、一对多联系（1∶n）、多对多联系（m∶n），定义了实体集 A 中的每一个实体与实体集 B 中若干个实体之间的对应关系。

实体-联系模型可以形象地用图形表示，称为实体-联系图，其中：矩形表示实体型，内部为实体名；椭圆形为属性，内部为属性名，用无向边与实体型连接；菱形表示联系，内部为联系名，用无向边与实体型连接，同时在无向边旁边标注联系的类型，联系的属性也要用无向边与联系连接起来。图 1—2 为图 1—1 结构化数据的实体-联系图，四个实体：CUSTOMER，SUPPLIER，PART，DATE 之间通过订单（Ordering）构成联系，订单联系中包含 quantity，price，discount，revenue 等属性。

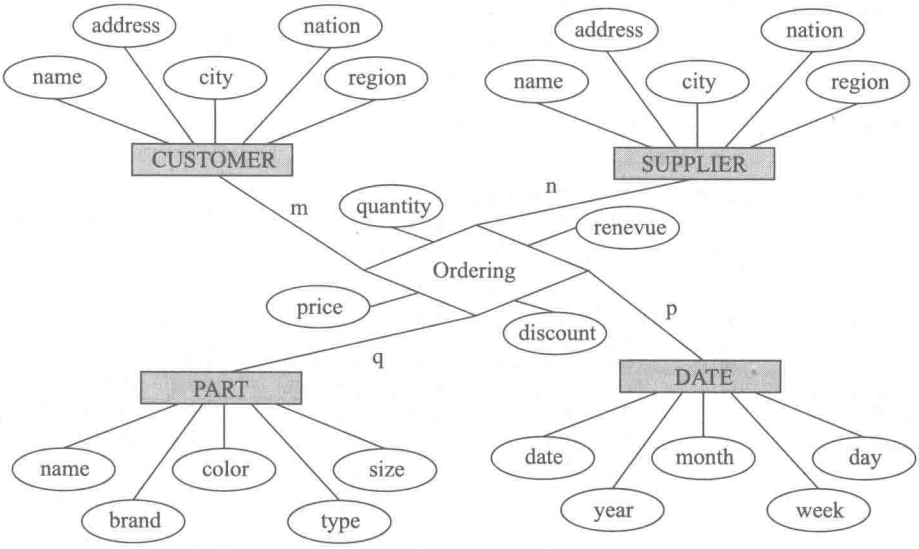


图 1—2　订单业务实体-联系图

二、关系模型

关系模型是最重要的一种数据模型，当前主流数据库采用的都是关系模型。关系模型由关系数据结构、关系操作集合和关系完整性约束三部分组成。

关系模型以关系作为唯一的数据结构。关系用二维表来表示实体以及实体之间的联系，二维表形象地看由行和列组成，列又称为字段（field）、属性（attribute），定义了实体的一个描述数据分量，关系中的属性必须是不可分的数据项；行又称为元组（tuple）、记录（record），是具有相同属性结构的数据集合。

在关系的定义中，关系的名字（表名）必须唯一，表中各列的名字必须唯一，称为属性名。在关系中，能够唯一确定一个元组的属性或属性组称为候选码。当关系中存在多个候选码时，可以选择其中的一个候选码作为关系的主码，主码用于定义表内或表间的约束关系。作为候选码的属性称为主属性（primary attribute），不包含在任何候选码中的属性称为非主属性（non-primary attribute）或非码属性（non-key attribute）。当所有属性构成的属性组是码时称为全码（all-key）。

关系具有以下五条基本性质：

（1）关系中的列是同质的。列对关系的一个分量，由相同类型的数据组成，数据来自相同的值域。如图 1—3 所示的关系 promotion 中，列 promotion_id 由整型数据构成，值域为正整数，表示每一个促销活动的 ID；列 start_date 由日期型数据构成，格式为 yyyy/mm/dd，值域为有效日期范围。

（2）不同的列值域可以相同。列 start_date，end_date 的值域都是日期范围，但具有不同的语义，使用不同的属性名。

（3）行列的顺序不重要。关系是一种集合，行是属性的集合，关系是行的集合，行与列的顺序对关系操作不重要，相同的关系可以有不同的行、列顺序。在关系数据库中，增加列或者增加行时对位置没有要求。

（4）关系中任意两个元组的候选码不能相同。候选码起到唯一标识关系中元组的作用，通常情况下关系中至少需要一个候选码，当所有的属性组成候选码时称为全码。候选码不能相同决定了关系中不存在重复的行，保证了任何记录可以被唯一标识和访问。例如图 1—3 所示的关系中，promotion_id 和 promotion_name 都是候选码。

promotion_id	promotion_district_id	promotion_name	media_type	cost	start_date	end_date
1	110	High Roller Savings	Product Attachment	14435	1996/1/3	1996/1/6
2	110	Green Light Special	Product Attachment	8907	1996/1/18	1996/1/20
3	110	Wallet Savers	Radio	12512	1996/2/2	1996/2/5
4	110	Weekend Markdown	In-Store Coupon	11256	1996/2/13	1996/2/15
5	110	Bag Stuffers	Sunday Paper, Radio	12275	1996/2/28	1996/3/1
6	110	Save-It Sale	Daily Paper	9472	1996/3/14	1996/3/16
7	110	Fantastic Discounts	Sunday Paper, Radio, TV	14278	1996/3/29	1996/4/2
8	110	Price Winners	Sunday Paper, Radio	14731	1996/4/10	1996/4/13
9	110	Dimes Off	Daily Paper	14065	1996/4/26	1996/4/29
10	110	Green Light Special	Sunday Paper, Radio	9298	1996/5/8	1996/5/9
11	110	Dollar Cutters	Daily Paper, Radio, TV	5306	1996/5/24	1996/5/25
12	110	Three for One	TV	14812	1996/6/6	1996/6/9
13	110	Price Winners	Cash Register Handout	11674	1996/6/19	1996/6/22
14	110	Big Promo	Street Handout	14945	1996/7/3	1996/7/7
15	110	Save-It Sale	Sunday Paper, Radio	6842	1996/7/18	1996/7/22
16	110	Sale Winners	Sunday Paper, Radio	14615	1996/8/2	1996/8/3
17	110	Savings Galore	Cash Register Handout	13694	1996/8/16	1996/8/17
18	110	Super Savers	Daily Paper, Radio	12346	1996/8/27	1996/8/30
19	110	Tip Top Savings	Daily Paper, Radio	8099	1996/9/10	1996/9/13
20	110	Sale Winners	Street Handout	11024	1996/9/26	1996/9/28

图 1—3　关系示例

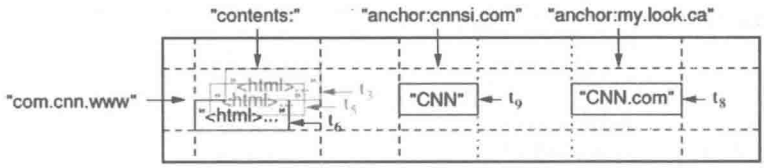
（5）关系中的属性必须是不可分的数据项。关系模型最基础的约束条件是关系的每一个分量不可分，属性中不可嵌套属性，不允许“表中嵌套表”的递归结构。

图 1—4 是一种属性嵌套结构，一个属性可以分解为多个属性，这是一种报表中常见的格式，但不属于关系模型。可以通过修改表中属性名字来消除属性嵌套，将其转换为关系模型。

Region	Unemployed Persons (10 000 persons)						Unemployment Rate (%)					
	1990	2005	2009	2010	2011	2012	1990	2005	2009	2010	2011	2012
Beijing	1.7	10.6	8.2	7.7	8.1	8.1	0.4	2.1	1.4	1.4	1.4	1.3
Tianjin	8.1	11.7	15.0	16.1	20.1	20.4	2.7	3.7	3.6	3.6	3.6	3.6
Hebei	7.7	27.8	34.5	35.1	36.0	36.8	1.1	3.9	3.9	3.9	3.8	3.7
Shanxi	5.5	14.3	21.6	20.4	21.1	21.0	1.2	3.0	3.9	3.6	3.5	3.3
Inner Mongolia	15.2	17.7	20.1	20.8	21.8	23.1	3.8	4.3	4.0	3.9	3.8	3.7

图 1—4　嵌套表

当前大数据应用中广泛使用的 key/value 存储模型中通常使用 column family 结构来描述复杂的数据结构，如图 1—5（A）所示的 key/value 存储结构中，“com. cnn. www”是唯一标识非结构化记录的键值，column family “contents:” 中存储了网页时间戳为 t_3，t_5 和 t_6 的三个版本，column family “anchor:” 中存储了时间戳为 t_8 和 t_9 的 anchor 信息“CNN”和“CNN. com”。这种 key/value 数据结构不符合关系模型的定义，属性中包含属性，不能被关系数据库所存储。但这种嵌套结构能够分解为如图 1—5（B）所示的关系存储，即将每一个 column family 分解为一个独立的关系，关系 anchor 和关系 contents 通过主码（row key，time stamp）建立元组之间的联系，从而表示完整的信息。



（A）key/value存储

Row Key	Time Stamp	Column Family "anchor"
"com. cnn. www"	t9	anchor:cnnsi. com = "CNN"
"com. cnn. www"	t8	anchor:my. look. ca = "CNN. com"

Row Key	Time Stamp	Column Family "contents:"
"com. cnn. www"	t6	contents:html = "<html>..."
"com. cnn. www"	t5	contents:html = "<html>..."
"com. cnn. www"	t3	contents:html = "<html>..."

（B）关系存储

图 1—5　column family 存储

关系模型具有良好的适应性，能够较好地表示现实世界的各种数据模型，如层次模型、网状模型以及部分非结构化数据模型。但在一些特殊的应用领域，尤其是互联网应用的非结构化数据处理领域，关系模型并不能够完全胜任。当前的大数据技术和 NoSQL 技术一方面通过新兴的 Map/Reduce，Hadoop 等技术扩展了在传统的关系数据库不能适用的大数据分析领域的应用，另一方面也促进了关系数据库技术在一些大数据分析领域的处理能力，使关系数据库与新兴的非结构化数据处理技术相结合，扩展关系数据库的应用领域。

三、关系操作

关系模型中常用的关系操作分为两大类：查询（query）操作和数据更新操作。

数据更新操作包括元组的插入（insert）、删除（delete）和修改（update）操作，负责数据库的元组管理功能。查询操作中最重要的操作是选择（select）、投影（project）和连接（join）操作。

关系操作是一种集合操作，操作的对象和操作的输出结果都是集合。也就是说，操作的对象是一个或多个关系，操作的结果也是一个关系，是操作对象关系的一个子关系或新生成的关系。

下面介绍选择、投影和连接关系操作。

1. 选择

选择（selection）操作是在关系 R 中选择满足给定条件的元组集合的操作，记作

$$\delta_F(R)=\{t \mid t \in R \wedge F(t)='\text{True}'\}$$

式中，F 表示选择条件，使用逻辑表达式形式，结果为 True 或 False。选择操作是对 R 中的每一个元组 t 在选择条件 F 上进行逻辑表达式计算，结果为 True 的元组为选择操作结果。

表 1—1 为常用的比较运算符，选择条件通常为属性名与常量或变量之间的逻辑表达式，根据表达式的结果对关系 R 中的元组进行过滤，选择出满足条件的输出元组。

表 1—1　　逻辑运算

查询条件	运算符	意义	示例
比较	=，>，<，>=，<=，!=，<>，!>，!<	比较大小	Cost>9000
确定范围	BETWEEN … AND，NOT BETWEEN…AND	判断值是否在范围内	Cost between 9000 and 12000
确定集合	IN，　NOT IN	判断值是否为列表中的值	Promotion_name in（‘Big Promo’，‘Super Savers’）
字符匹配	LIKE，NOT LIKE	判断值是否与指定的字符通配格式相符	Promotion_name like ‘Big%’
空值	IS NULL，IS NOT NULL	判断值是否为空	Promotion_name IS NULL
非运算	¬	逻辑结果取反	¬ Cost>9000
与运算	∧	合取，需要同时满足两个条件	Cost>9000 ∧ Promotion_name IS NULL
或运算	∨	析取，满足一个条件即为 True	Cost>9000 ∨ Promotion_name IS NULL

【例 1】 查询 promotion 表中 cost 大于 1 000 的元组。

$\delta_{cost>10000}(promotion)$

查询结果如图 1—6（A）所示。

【例 2】 查询 promotion 表中 promotion_name 为 Big Promo 的元组。

$\delta_{promotion\_name='BigPromo'}(promotion)$

查询结果如图 1—6（B）所示。

promotion_id	promotion_district_id	promotion_name	media_type	cost	start_date	end_date
1	110	High Roller Savings	Product Attachment	14435	1996/1/3	1996/1/6
3	110	Wallet Savers	Radio	12512	1996/2/2	1996/2/5
4	110	Weekend Markdown	In-Store Coupon	11256	1996/2/13	1996/2/15
5	110	Bag Stuffers	Sunday Paper, Radio	12275	1996/2/28	1996/3/1
7	110	Fantastic Discounts	Sunday Paper, Radio, TV	14278	1996/3/29	1996/4/2
8	110	Price Winners	Sunday Paper, Radio	14731	1996/4/10	1996/4/13
9	110	Dimes Off	Daily Paper	14065	1996/4/26	1996/4/29
12	110	Three for One	TV	14812	1996/6/6	1996/6/9
13	110	Price Winners	Cash Register Handout	11674	1996/6/19	1996/6/22
14	110	Big Promo	Street Handout	14945	1996/7/3	1996/7/7
16	110	Sale Winners	Sunday Paper, Radio	14615	1996/8/2	1996/8/3
17	110	Savings Galore	Cash Register Handout	13694	1996/8/16	1996/8/17
18	110	Super Savers	Daily Paper, Radio	12346	1996/8/27	1996/8/30
20	110	Sale Winners	Street Handout	11024	1996/9/26	1996/9/28

（A）

promotion_id	promotion_district_id	promotion_name	media_type	cost	start_date	end_date
14	110	Big Promo	Street Handout	14945	1996/7/3	1996/7/7

（B）

图 1—6　选择操作结果

2. 投影

投影（projection）操作是从关系 R 中选择出若干属性列组成新的关系，记作