第 一 篇

数据库原理与应用基础

第 1 章 数据库概论

本章要点：

- 数据库基本概念；
- 数据模型；
- 开发环境搭建过程；
- 数据库系统结构；
- 大数据、分布式数据库和并行数据库；
- 数据仓库、联机分析处理和数据挖掘。

数据库在金融、工业、农业、运输业、电信业、服务业、制造业、教育及政府部门等诸多行业的信息系统中有着广泛的应用，数据库建设的规模、信息量大小和使用的频度已成为衡量一个国家信息化程度的重要标志。数据库技术是信息系统的核心和基础，越来越多的应用领域采用数据库技术进行数据的存储和处理，成为计算机科学与技术中发展最快，应用最广的一项重要技术。本章从数据库基本概念出发，介绍数据模型、数据库系统结构、大数据、分布式数据库、并行数据库、数据仓库、联机分析处理和数据挖掘，它是学习以后各章的基础。

1.1 数据库基本概念

数据库是长期存放在计算机内的、有组织的、可共享的数据集合。数据库管理系统是一个系统软件，用于科学地组织和存储数据、高效地获取和维护数据。数据库系统是在计算机系统中引入数据库之后组成的系统，它是用来组织和存取大量数据的管理系统。

1.1.1 数据库

1. 数据

数据（Data）是事物的符号表示，其形式包括有数字、文字、图像、声音等，经数字化处理后以二进制形式存储于计算机。

日常生活中人们直接用自然语言描述事物，而在计算机系统中，则需要提取事物的特征并组织成记录来进行描述。例如，一个学生记录数据如下所示：

241001	孙俊松	男	2005-12-17	计算机	52

数据的含义称为信息，数据是信息的载体，信息是数据的内涵，是对数据的语义解释。

2. 数据库

数据库（Database，DB）是长期存放在计算机内的、有组织的、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，具有尽可能小的冗余度、较高的数据独立性和易扩张性等特点。

数据库具有以下特性：

- 共享性。数据库中的数据能被多个应用程序的用户所使用。
- 独立性。提高了数据和程序的独立性，有专门的语言支持。
- 完整性。指数据库中数据的正确性、一致性和有效性。
- 减少数据冗余。

数据库包含了以下含义：

- 建立数据库的目的是为应用服务。
- 数据存储在计算机的存储介质中。
- 数据结构比较复杂，有专门理论支持。

1.1.2 数据库管理系统

数据库管理系统（Data Base Management System，DBMS）是数据库系统的核心组成部分，它是在操作系统支持下的系统软件，是对数据进行管理的大型系统软件，用户在数据库系统中的一些操作都是由数据库管理系统来实现的。

- 数据定义功能。提供数据定义语言定义数据库和数据库对象。
- 数据操纵功能。提供数据操纵语言对数据库中的数据进行查询、插入、修改及删除等操作。
- 数据控制功能。提供数据控制语言进行数据控制，即提供数据的安全性、完整性及并发控制等功能。
- 数据库建立维护功能。包括数据库初始数据的装入、转储、恢复和系统性能监视、分析等功能。

1.1.3 数据库系统

数据库系统(Database System，DBS)是在计算机系统中引入数据库后的系统架构，主要由数据库、操作系统、数据库管理系统、应用程序、用户、数据库管理员(DataBase Administrator，DBA)组成，如图1.1所示。数据库系统在整个计算机系统中的地位如图1.2所示。

数据库应用系统分为客户/服务器架构和浏览器/服务器架构。

1. 客户/服务器（C/S）架构的应用系统

当应用程序需要处理数据库中的数据时，首先向数据库管理系统发送一个数据请求，数据库管理系统接收到这一请求后，对其进行分析，然后执行数据库操作，并把处理结果返回给应用程序。

由于应用程序直接与用户交互并向数据库管理系统提出服务请求，因此被称为“前台”“客户端”“客户程序（Client）”；而数据库管理系统不直接与用户打交道，且专为

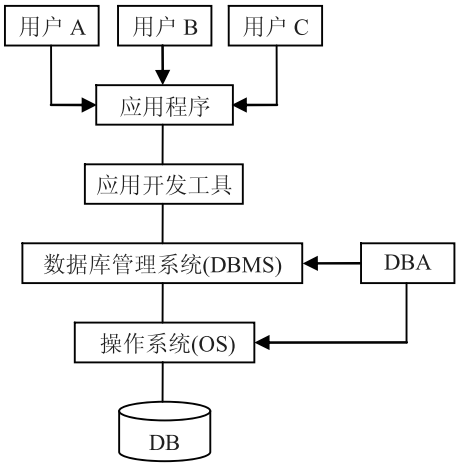

图 1.1　数据库系统

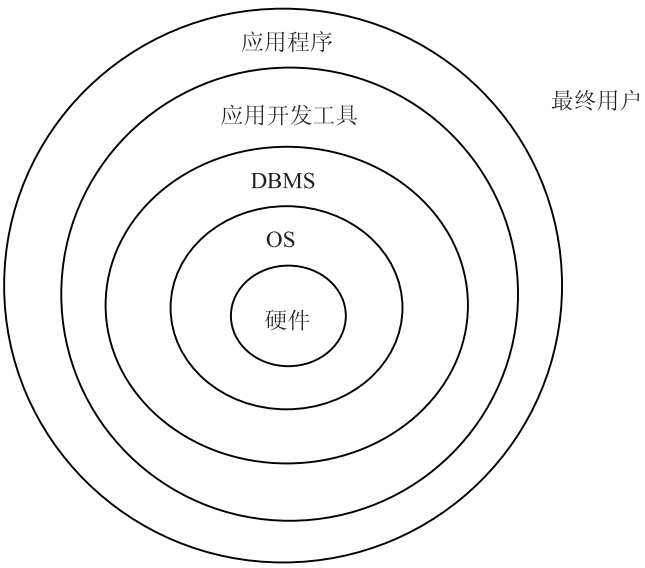

图 1.2　数据库在计算机系统中的地位

应用程序提供服务，故被称为“后台”“服务器”“服务器程序（Server）”。这一操作数据库的模式称为客户/服务器（Client/Server，C/S）架构，如图 1.3 所示。

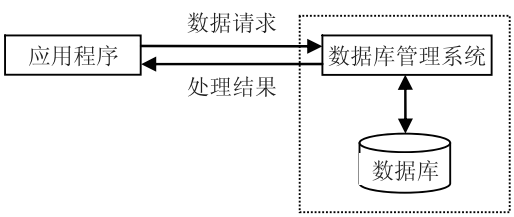

图 1.3　客户/服务器（C/S）架构

客户程序的开发，目前流行的工具主要有 Visual C++、.NET 框架、Visual Basic 等。

2. 浏览器/服务器（B/S）架构的应用系统

浏览器/服务器（Browser/Server，B/S）架构是一种基于 Web 应用的客户/服务器架构，又称为三层客户-服务器架构（浏览器/Web 服务器/数据库服务器），如图 1.4 所示。

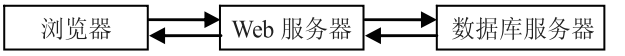

图 1.4　浏览器/服务器（B/S）架构

在图 1.4 中，浏览器（Browser）是用户输入数据和显示结果的交互界面，用户在浏览器表单中输入数据，然后将表单中的数据提交并发送到 Web 服务器。Web 服务器接收并处理用户的数据，通过数据库服务器，从数据库中查询需要的数据（或把数据录入数据库）回送 Web 服务器。Web 服务器把返回的结果插入 HTML 页面，传送给客户端，在浏览器中显示出来。

目前开发数据库 Web 界面的工具主要有 PHP、Java EE、ASP.NET(C#)等。

1.1.4　数据管理技术的发展

数据管理是指对数据进行分类、组织、编码、存储、检索和维护等项工作，数据管理技术的发展经历了人工管理阶段、文件系统阶段、数据库系统阶段，现在正在向更高一级的数据库系统发展。

1. 人工管理阶段

20 世纪 50 年代中期以前，人工管理阶段的数据是面向应用程序的，一个数据集只能对应一个程序，应用程序与数据之间的关系如图 1.5 所示。

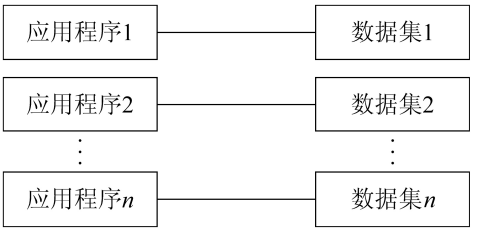

图 1.5　人工管理阶段应用程序与数据之间的关系

人工管理阶段的特点如下。

（1）数据不保存。

只是在计算某一课题时将数据输入，用完即撤走。

（2）数据不共享。

数据面向应用程序，一个数据集只能对应一个程序，即使多个不同程序用到相同数据，也得各自定义。

（3）数据和程序不具有独立性。

数据的逻辑结构和物理结构发生改变，必须修改相应的应用程序，即要修改数据必须修改程序。

（4）没有软件系统对数据进行统一管理。

2. 文件系统阶段

20 世纪 50 年代后期到 60 年代中期，计算机不仅用于科学计算，也开始用于数据管理。数据处理的方式不仅有批处理，还有联机实时处理。应用程序和数据之间的关系如图 1.6 所示。

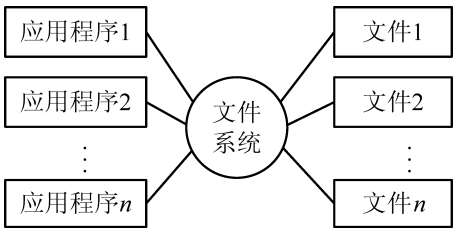

图 1.6　文件系统阶段应用程序与数据之间的关系

文件系统阶段数据管理的特点如下。

（1）数据可长期保存。

数据以文件的形式长期保存。

（2）数据共享性差，冗余度大。

在文件系统中，一个文件基本对应一个应用程序，当不同应用程序具有相同数据时，也必须各自建立文件，而不能共享数据相同数据，数据冗余度大。

（3）数据独立性差。

当数据的逻辑结构改变时，必须修改相应的应用程序，数据依赖于应用程序，独立性差。

（4）由文件系统对数据进行管理。

由专门的软件（文件系统）进行数据管理，文件系统把数据组织成相互独立的数据文件，可按文件名访问，按记录存取，程序与数据之间有一定的独立性。

3. 数据库系统阶段

20 世纪 60 年代后期开始，数据管理对象的规模越来越大，应用越来越广泛，数据量快速增加。为了实现数据的统一管理，解决多用户、多应用共享数据的需求，数据库技术应运而生，出现了统一管理数据的专门软件——数据库管理系统。

数据库系统阶段，应用程序和数据之间的关系如图 1.7 所示。

数据库系统与文件系统相比较，具有以下的主要特点：

（1）数据结构化。

（2）数据的共享度高，冗余度小。

（3）有较高的数据独立性。

（4）由数据库管理系统对数据进行管理。

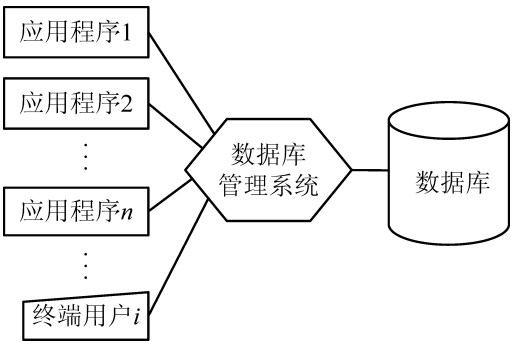

图 1.7　数据库系统阶段应用程序与数据之间的关系

在数据库系统中，数据库管理系统作为用户与数据库的接口，提供了数据库定义、数据库运行、数据库维护和数据安全性、完整性等控制功能。

1.2　数 据 模 型

模型是对现实世界中某个对象特征的模拟和抽象，数据模型（Data Model）也是一种模型，它是对现实世界数据特征的抽象，用来描述数据、组织数据和对数据进行操作的。数据模型是数据库系统的核心和基础，数据库管理系统的实现都是建立在某种数据模型基础上的。

1.2.1　数据建模


视频讲解

将现实世界中的具体事物抽象、组织为某一数据库管理系统支持的数据模型，这个过程称为数据建模（Data Modeling）。

数据建模有以下两个步骤。

1. 建立概念模型

将现实世界抽象为信息世界，即将现实世界中的客观事物抽象为某种信息架构，该信息结构不依赖于具体的计算机系统，不是某一数据库管理系统支持的数据模型，而是概念级的模型，称为概念模型（Conceptual Model）。

概念模型是按用户观点对数据建模，用于数据库设计。从现实世界到概念模型的建模任务由数据库设计人员完成，可以通过数据库设计工具辅助设计人员完成。

2. 将概念模型转换为数据模型

将信息世界转换为机器世界，就是将概念模型转换为计算机上某一数据库管理系统支持的数据模型。

数据模型是按计算机的观点对数据建模，用于数据库管理系统的实现。从概念模型到数据模型的转换由数据库设计人员完成，可以通过数据库设计工具辅助设计人员完成。

数据建模的步骤如图 1.8 所示。

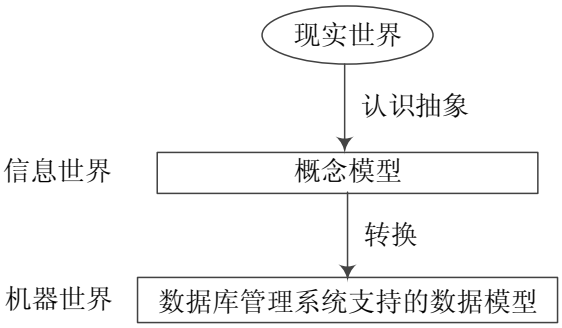

图 1.8　数据建模的步骤

1.2.2　概念模型

1. 概念模型的基本概念

概念模型是对现实世界的第一层抽象，又称信息模型，是数据库设计人员和用户之间交流的工具，仅需考虑领域实体属性和联系，要求有较强的语义表达能力，且简单清晰、易于理解，其基本概念如下：

（1）实体。客观存在并可相互区别的事物称为实体。实体可以是具体的人、事、物或抽象的概念，例如，在教学管理系统中，“学生”就是一个实体。

（2）属性。实体所具有的某一特性称为属性。属性采用椭圆框表示，框内为属性名，并用无向边与其相应实体连接。例如，在教学管理系统中，学生的特性有学号、姓名、性别、出生日期、籍贯、总学分、专业代码，它们就是学生实体的 7 个属性。

（3）实体型。用实体名及其属性名集合来抽象和刻画同类实体，称为实体型。例如，学生（学号, 姓名, 性别, 出生日期, 籍贯, 总学分, 专业代码）就是一个实体型。

（4）实体集。同型实体的集合称为实体集，例如全体学生记录就是一个实体集。

（5）联系。在现实世界中，事物内部和事物之间的联系，在概念模型中反映为实体（型）内部的联系和实体（型）之间的联系。

2. 实体之间的联系

实体之间的联系，可分为一对一的联系、一对多的联系和多对多的联系。

（1）一对一的联系（1∶1）。

例如，一个班只有一个正班长，而一个正班长只属于一个班，班级与正班长两个实体间具有一对一的联系。

（2）一对多的联系（1∶n）。

例如，一个班可有若干学生，一个学生只能属于一个班，班级与学生两个实体间具有一对多的联系。

（3）多对多的联系（m∶n）。

例如，一个学生可选多门课程，一门课程可被多个学生选修，学生与课程两个实体间具有多对多的联系。

1.2.3 数据模型的三要素

数据模型是现实世界数据特征的抽象，一般由数据结构、数据操作和数据完整性约束三部分组成。

1. 数据结构

数据结构用于描述系统的静态特性，是所研究对象类型的集合，数据模型按其数据结构分为层次模型、网状模型和关系模型等。数据结构所研究的对象是数据库的组成部分，包括两类：一类是与数据类型、内容、性质有关的对象，例如关系模型中的域、属性等；另一类是与数据之间联系有关的对象，例如关系模型中反映联系的关系等。

2. 数据操作

数据操作用于描述系统的动态特性，是指对数据库中各种对象及对象的实例允许执行的操作的集合，包括对象的创建、修改和删除，对对象实例的检索、插入、删除、修改及其他有关操作等。

3. 数据完整性约束

数据完整性约束是一组完整性约束规则的集合，完整性约束规则是给定数据模型中数据及其联系所具有的制约和依存的规则。

数据模型三要素在数据库中都是严格定义的一组概念的集合。在关系数据库中，数据结构是表结构定义及其他数据库对象定义的命令集；数据操作是数据库管理系统提供的数据操作（操作命令、语法规定、参数说明等）命令集；数据完整性约束是各关系表约束的定义及操作约束规则等的集合。

注意：

（1）这里讲的数据模型都是逻辑上的，即逻辑模型（Logical Data Model），也称为逻辑数据模型。

（2）这些数据模型将以一定的组织方式存储在数据库管理系统中，它是数据模型在数据库管理系统内部的物理存储结构，称为物理模型（Physical Data Model）。

1.2.4 层次模型、网状模型和关系模型

数据模型按照数据库系统的发展进程可分为层次模型、网状模型、关系模型、面向对象数据模型、对象关系数据模型和半结构化 XML 数据模型等。下面介绍其中的三种模型：层次模型、网状模型和关系模型，关系模型是应用最广泛、最重要的一种数据模型。

1. 层次模型

用树状层次结构组织数据，树状结构每个节点表示一个记录类型，记录类型之间的联系是一对多的联系。层次模型有且仅有一个根节点，位于树状结构顶部，其他节点有且仅有一个父节点。某大学按层次模型组织数据的示例如图 1.9 所示。

层次模型简单易用，但现实世界很多联系是非层次性的，如多对多联系等，表达起来比较笨拙且不直观。