第 1 章 数据库系统概述

数据库技术是现代信息科学与技术的重要组成部分，是计算机数据处理与信息管理系统的核心。数据库技术研究和解决了计算机信息处理过程中大量数据有效地组织和存储的问题，在数据库系统中减少数据存储冗余、实现数据共享、保障数据安全以及高效地检索数据和处理数据。数据库技术的根本目标是要解决数据的共享问题。

数据库系统涉及许多基本概念，主要包括信息、数据、数据处理、数据库、数据库管理系统以及数据库系统等。

1.1 数据管理技术的发展

数据管理技术的发展是和计算机技术及其应用的发展联系在一起的，经历了由低级到高级的发展过程。

1.1.1 数据和数据管理

数据库系统的核心任务是数据管理。数据库技术是一门研究如何存储、使用和管理数据的技术，是计算机数据管理技术的最新发展阶段。数据库应用涉及数据（data）、信息（information）、数据处理（data processing）和数据管理（data management）等基本概念。

1．数据和信息

现代社会是信息的社会，信息以惊人的速度增长，因此，如何有效地组织和利用它们成为急需解决的问题。数据库系统的目的是高效地管理及共享大量的信息，而信息与数据是分不开的。

数据和信息是数据处理中的两个基本概念，有时可以混用，但有时必须分清。数据是信息的一种表现形式，数据通过能书写的信息编码表示信息。信息有多种表现形式，可通过手势、眼神、声音或图形等方式表达，但是数据是信息的最佳表现形式。由于数据能够书写，因而它能够被记录、存储和处理，从中挖掘出更深层的信息。因此，也可以说，数据是原材料，信息是产品，信息是数据的含义。例如数据 1、3、5、7、9、11、13、15，这是一组原始数据，对其进行分析便可以看出：这是一组首项是 1、公差为 2 的等差数列，我们可以比较容易地知道它的任意项的值和前 n 项的和，这便是一条信息。而从数据 1、3、–23、4、5、–1、41 中不能提炼出任何有用的内容，故它不是信息。

数据和信息可以混用，有时一些数据对某些人来说可能是信息，而对另外一些人而言则可能只是数据。例如，在运输管理中，运输单对司机来说是信息，这是因为司机可以从该运输单上知道什么时候要为什么客户运输什么物品。而对负责经营的管理者来说，运输

单只是数据，因为从单张运输单中无法知道本月经营情况，也不能掌握现有可用的司机、运输工具等情况。

2．数据处理与数据管理

数据处理是对数据的采集、存储、检索、加工、变换和传输，基本目的是从大量的、可能是杂乱无章的、难以理解的数据中抽取并推导出对于某些特定的人们来说是有价值、有意义的数据。例如，某省全体高考学生各门课程成绩的总分按从高到低的顺序进行排序、统计各个分数段的人数等，进而可以根据招生人数确定录取分数线。数据处理是系统工程和自动控制的基本环节。数据处理贯穿于社会生产和社会生活的各个领域。数据处理技术的发展及其应用的广度和深度，极大地影响着人类社会发展的进程。

数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、检查、维护并实现对数据的各种运算和操作。其目的在于充分有效地发挥数据的作用。实现数据有效管理的关键是数据组织。随着计算机技术的发展，数据管理经历了人工管理、文件管理、数据库管理 3 个阶段。

1.1.2 数据管理发展的 3 个阶段

计算机硬件、系统软件的发展和计算机应用范围不断扩大是促使数据管理技术发展的主要因素。随着数据管理技术的不断发展，标志着数据存储冗余不断减小、数据独立性不断增强以及数据操作更加方便和简单。

1．人工管理阶段

在这一阶段（20 世纪 50 年代中期之前），计算机主要用于科学计算，其他工作还没有展开。外部存储器只有磁带、卡片和纸带等，还没有磁盘等直接存取存储设备。软件也处于初级阶段，只有汇编语言，没有操作系统和数据管理方面的软件。数据处理方式基本是批处理。这个阶段有如下几个特点：

（1）数据不保存，数据也无须长期保存。

（2）计算机系统不提供对用户数据的管理功能。用户编制程序时，必须全面考虑好相关的数据，包括数据的定义、存储结构以及存取方法等。程序和数据是一个不可分割的整体。数据脱离了程序就无任何存在的价值，数据无独立性。

（3）只有程序的概念，没有文件的概念。数据的组织形式必须由程序员自行设计。

（4）数据不能共享。不同的程序均有各自的数据，这些数据对不同的程序通常是不相同的，也不可共享。即使不同的程序使用了相同的一组数据，这些数据也不能共享，程序中仍然需要各自加入这组数据，不能省略。基于这种数据的不可共享性，必然导致程序与程序之间存在大量的重复数据，浪费了存储空间。

（5）数据面向程序，不单独保存数据。基于数据与程序是一个整体，数据只为本程序所使用，数据只有与相应的程序一起保存才有价值，否则就毫无用处。所以，所有程序的数据均不单独保存。

例如，学校管理系统有人事管理、学生管理和课程管理三部分，在人事管理阶段应用程序与数据之间的依赖关系如图 1.1 所示。

图 1.1　应用程序和数据的依赖关系

2．文件管理阶段

在这一阶段（20 世纪 50 年代后期至 60 年代中期），计算机不仅用于科学计算，还用于信息管理。随着数据量的增加，数据的存储、检索和维护问题成为紧迫的需要，数据结构和数据管理技术迅速发展起来。此时，外部存储器已有磁盘、磁鼓等直接存取存储设备。软件领域出现了高级语言（如 Fortran、Algol60 等）和操作系统（如 GM-NAA I/O、IBSYS 等）。操作系统中的文件系统是专门管理外存的数据管理软件。数据处理的方式有批处理，也有联机实时处理。

这一阶段的数据管理有以下特点：

（1）数据以“文件”形式可长期保存在外部存储器的磁盘上。由于计算机的应用转向了信息管理，因此对文件要进行大量的查询、修改和插入等操作。

（2）数据的逻辑结构与物理结构有了区别，但比较简单。程序与数据之间具有“设备独立性”，即程序只需用文件名就可以进行数据操作，不必关心数据的物理位置。由操作系统的文件系统提供存取方法（读/写）。

（3）文件组织已多样化。有索引文件、链接文件和直接存取文件等。但文件之间相互独立、缺乏联系。数据之间的联系要通过程序去构造。

（4）数据不再属于某个特定的程序，可以重复使用，即数据面向应用。但是文件结构的设计仍然是基于特定的用途，程序基于特定的物理结构和存取方法。因此程序与数据结构之间的依赖关系并未根本改变。

在文件系统阶段，由于具有设备独立性，因此当改变存储设备时，不必改变应用程序。但这只是初级的数据管理，在数据的物理结构修改时，仍然需要修改用户的应用程序，即应用程序具有“程序-数据依赖”性。有关物理表示的知识和访问技术将直接体现在应用程序的代码中。

（5）对数据的操作以记录为单位。这是由于文件中只存储数据，不存储文件记录的结构描述信息。文件的建立、存取、查询、插入、删除及修改等所有操作，都要用程序来实现。

文件系统阶段是数据管理技术发展中的一个重要阶段。在这一阶段，得到充分发展的数据结构和算法丰富了计算机科学，为数据管理技术的进一步发展打下了基础。

随着数据管理规模的扩大，数据量急剧增加，文件系统显露出了 3 个明显的缺陷：

（1）数据冗余（redundancy）。由于文件之间缺乏联系，造成每个应用程序都有对应的文件，有可能同样的数据在多个文件中重复存储。

（2）数据不一致（inconsistency）。这往往是由数据冗余造成的，在进行数据更新操作

时，稍不谨慎，就可能使同样的数据在不同的文件中不一样。

（3）数据联系弱（poor data relationship）。这是由于文件之间相互独立，相互之间又缺乏联系造成的。

例如，学校管理系统有人事管理、学生管理和课程管理三部分，在文件管理阶段应用程序与数据之间的依赖关系如图 1.2 所示。

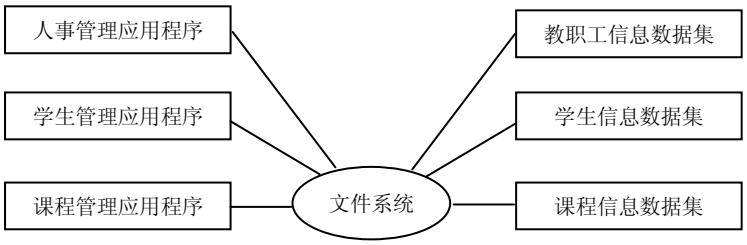

图 1.2　应用程序和数据的依赖关系

3．数据库管理阶段

在 20 世纪 60 年代末，磁盘技术取得了重要进展，具有数百兆容量和快速存取的磁盘陆续进入市场，成本也不高。同时，计算机在管理中应用规模更加庞大、数据量急剧增加，为数据库技术的产生提供了良好的物质条件。数据库系统克服了文件系统的缺陷，提供了对数据更高级、更有效的管理。概括起来，数据库阶段的管理方式具有以下特点：

（1）数据结构化。数据结构化是数据库系统与文件系统之间的根本区别。数据库中包含许多单独的数据文件，这些文件的数据具有特定数据结构（数据项不可再分、类型相同）。文件之间也存在相互的联系，在整体上也服从一定的组织形式，从而满足管理大量数据的需求。

（2）较高的数据共享性。数据共享是指数据不再面向某个应用而是面向整个系统，使得多个用户同时存取数据而互不影响。如同一企业中的不同部门，甚至不同企业、不同地区的用户都可以使用同一个数据库中的数据，减少了数据冗余。

数据共享性可以大大减少数据冗余，节约存储空间，还能够避免数据之间的不相容性与不一致性。

（3）统一管理和控制数据。利用专门的数据库管理系统实现对数据定义、操作、统一管理和控制。在应用程序和数据库之间保持高度的独立性，数据具有完整性、一致性和安全性，并具有充分的共享性，有效地减少了数据冗余。

例如，学校管理系统有人事管理、学生管理和课程管理三部分，在数据库管理阶段应用程序与数据之间的依赖关系如图 1.3 所示。

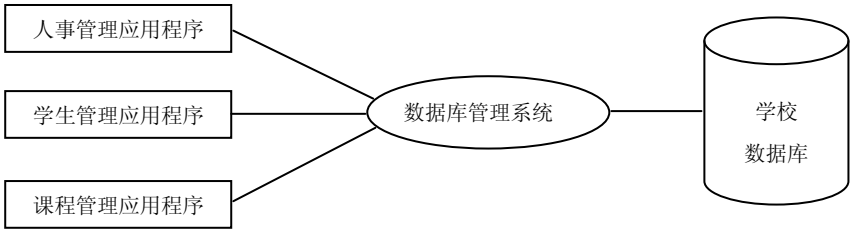

图 1.3　应用程序和数据的依赖关系

1.2 数据库系统

视频讲解

数据库系统（DataBase System，DBS）是为适应数据处理需要而发展起来的一种较为理想数据处理的核心机构。它是一个实际可运行的，用于存储、维护数据的软件系统并且可以向应用系统提供数据，是存储介质、处理对象和管理系统的集合体。

1.2.1 数据库系统的组成

数据库系统是指在计算机系统中引入数据库后的系统，一般由计算机硬件、数据库、数据库管理系统、数据库开发工具、数据库应用系统、数据库管理员和用户构成。数据库系统各元素的层次结构如图 1.4 所示。

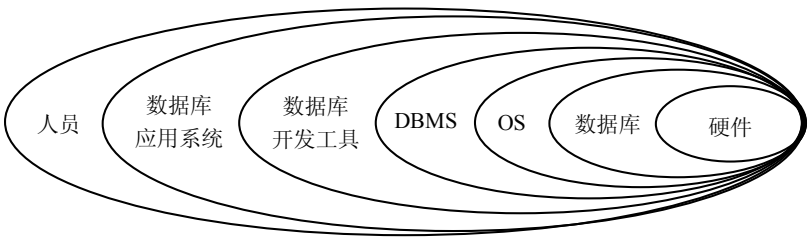

图 1.4　数据库系统层次图

1．计算机硬件

计算机硬件是数据库系统的物质基础，是存储数据库及运行数据库管理系统的硬件资源，主要包括主机、存储设备、输入输出设备以及计算机网络环境。

2．数据库

数据库（DataBase，DB）是指长期存储在计算机内、有组织的、统一管理的相关数据的集合。数据库能为各种用户共享，具有较小的数据冗余度，数据间联系紧密而又有较高的数据独立性等。

特别需要指出的是，数据库中存储数据具有“集成”和“共享”的特点。

所谓“集成”，是指把某个特定应用环境中的与各种应用相关的数据及其数据之间的联系（联系也是一种数据）全部集中并按照一定的结构形式进行存储，或者说，把数据库看成若干个性质不同的数据文件的联合和统一的数据整体，并且在文件之间局部或全部消除了冗余，使数据库系统具有整体数据结构化和数据冗余小的特点。

所谓“共享”，是指数据库中的一块块数据可为多个不同的用户所共享，即多个不同的用户使用多种不同的语言，为了不同的应用目的同时存取数据库，甚至同时存取同一块数据。共享实际上是基于数据库具有“集成”特点这一事实的结果。

3．计算机软件

数据库系统中的软件包括操作系统（Operating System，OS）、数据库管理系统、数据库开发工具及数据库应用系统。操作系统给用户提供良好的应用接口，数据库系统必须在操作系统的支持下才能正常运转。

数据库管理系统（DataBase Management System，DBMS）为用户或应用程序提供访问数据库的方法，包括数据库的建立、查询、更新及各种数据控制。数据库系统各类用户对数据库的各种操作请求，都是由数据库管理系统来完成的，它是数据库系统的核心软件。

数据库管理系统有大小之分，常见的大中型数据库管理系统有 Oracle、DB2、SQL Server、Sybase 等，小型数据库管理系统有 Foxpro、Access、MySQL 等。

数据库管理系统的主要功能有以下五方面。

（1）数据库的定义功能：数据库管理系统提供数据定义语言（Data Definition Language，DDL）定义数据库的三级结构、两级映像，定义数据的完整性约束、保密限制约束等。

（2）数据库的操纵功能：数据库管理系统提供数据操纵语言（Data Manipulation Language，DML）实现对数据的操作。基本的数据操作有两类：检索（查询）和更新（包括插入、删除、更新）。

（3）数据库的保护功能：数据库中的数据是信息社会的战略资源，对数据的保护是至关重要的大事。数据库管理系统对数据库的保护通过如下四个方面实现。

- 数据库的恢复：在数据库被破坏或数据不正确时，系统有能力把数据库恢复到正确的状态。
- 数据库的并发控制：在多个用户同时对同一个数据进行操作时，系统应能加以控制，防止破坏数据库中的数据。
- 数据完整性控制：保证数据库中数据及语义的正确性和有效性，防止任何对数据造成错误的操作。
- 数据安全性控制：防止未经授权的用户存取数据库中的数据，以避免数据的泄露、更改或破坏。

（4）数据库的维护功能：这一部分包括数据库的数据载入、转换、转储，数据库的改组以及性能监控等功能。

（5）数据字典：数据库系统中存放三级结构定义的数据库称为数据字典（Data Dictionary，DD）。对数据库的操作都要通过访问数据字典才能实现。数据字典中还存放数据库运行时的统计信息，例如记录个数、访问次数等。管理数据字典的实用程序称为“数据字典系统”，访问数据字典中的数据是由数据字典系统实现的。在现有的大型系统中，把数据字典系统单独抽出来成为一个软件工具。

有两点需要说明：一是数据库管理系统功能强弱随系统而异，大系统功能较强、较全，小系统较弱、较少；二是应用程序并不属于数据库管理系统范围，应用程序是用主语言和 DML 编写的，程序中 DML 语句由数据库管理系统执行，而其余部分仍由主语言编译程序完成。

4．人员

开发、管理和使用数据库系统的人员主要是数据库管理员（DataBase Administrator，DBA）、系统分析员（System Analyst，SA）和数据库设计人员、应用程序员和用户。

1）数据库管理员

在数据库环境下，有两类共享资源：一类是数据库，另一类是数据库管理系统软件。因此需要有专门的管理机构来监督和管理数据库系统。DBA 则是这个机构的一个（组）人员，负责全面管理和控制数据库系统。具体职责包括：

（1）决定数据库中的信息内容和结构。数据库中要存放哪些信息，DBA 要参与决策。因此 DBA 必须参加数据库设计的全过程，并与用户、应用程序员、系统分析员密切合作，搞好数据库设计。

（2）决定数据库的存储结构和存取策略。DBA 要综合各用户的应用需求，与数据库设计人员共同决定数据的存储结构和存取策略，以求获得较高的存取效率和存储空间的利用率。

（3）定义数据的安全性要求和完整性约束条件。DBA 的重要职责之一是保证数据库的安全性和完整性。因此 DBA 负责确定各个用户对数据库的存取权限、数据的保密级别和完整性约束条件。

（4）监控数据库的使用和运行。DBA 还有一个重要职责就是监视数据库系统的运行情况，及时处理运行过程中出现的问题。例如，当系统发生各种故障时，数据库会因此遭到不同程度的破坏，DBA 必须在最短时间内将数据库恢复到正确状态，并尽可能地不影响或少影响计算机系统其他部分的正常运行。为此，DBA 要定义和实施适当的备份和还原策略。如周期性地转储数据、维护日志文件等。

（5）数据库的改进、重组或重构。DBA 还负责在系统运行期间监视系统的空间利用率、处理效率等性能指标，对运行情况进行记录、统计分析，依靠工作实践并根据实际应用环境，不断改进数据库设计。不少数据库产品都提供了对数据库运行状况进行监视和分析的工具，DBA 可以使用这些软件完成这项工作。

另外，在数据运行过程中，大量数据不断插入、删除、修改，时间一长，会影响系统的性能。因此，DBA 要定期对数据库进行重组织，以提高系统的性能。

当用户的需求增加和改变时，DBA 还要对数据库进行较大的改造，包括修改部分设计，即数据库的重构造。

2）系统分析员和数据库设计人员

系统分析员负责应用系统的需求分析和规范说明，要和用户及 DBA 相结合，确定系统的硬件和软件配置，并参与数据库系统的概要设计。

数据库设计人员负责数据库中数据的确定、数据库各级模式的设计。数据库设计人员必须参与用户需求调查和系统分析，然后进行数据库设计。在很多情况下，数据库设计人员就由数据库管理员担任。

3）应用程序员

应用程序员负责设计和编写应用系统的程序模块，并进行调试和安装。

4）用户

这里用户是指最终用户。最终用户通过应用系统的用户接口使用数据库。常用的接口方式有浏览器、菜单驱动、表格操作、图形显示及报表打印等。

综上所述，数据库中包含的数据是存储在存储介质上的数据文件的集合；每个用户均可使用其中的数据，同一组数据可以为多个用户共享；数据库管理系统为用户提供对数据的存储组织、操作管理功能；用户通过数据库管理系统和应用程序实现对数据库系统的操作与应用。

1.2.2 数据库系统结构

为了有效地组织、管理数据，提高数据库的逻辑独立性和物理独立性，美国国家标准协会（American National Standards Institute，ANSI）数据库管理系统研究小组于 1978 年提出了标准化的建议，为数据库设计了一个严密的三级模式体系结构，它包括外模式（external schema）、概念模式（conceptual schema）和内模式（internal schema）。三级结构之间差别往往很大，为了实现这 3 个抽象级别的联系和转换，数据库管理系统提供了外模式与概念模式和概念模式与内模式的两级映像。数据库系统体系结构如图 1.5 所示。

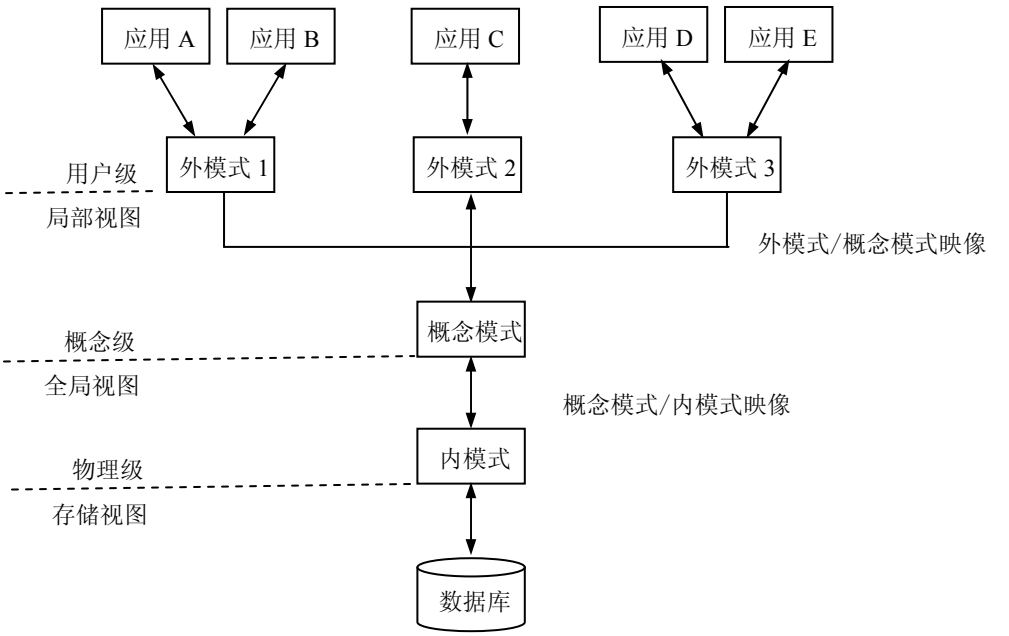

图 1.5　数据库系统体系结构

1．外模式

外模式又称为子模式或用户模式，是用户与数据库系统的接口，是用户能够看见和使用的局部数据逻辑结构和特征的描述。外模式是从概念模式中导出与某一应用有关的数据子集的逻辑表示。一个数据库可以有多个外模式。如果不同用户在应用需求、看待数据的方式以及对数据保密的要求等方面存在差异，则其外模式描述就是不同的。即使对概念模式中的同一数据，其在外模式中的结构、类型、长度、保密级别等都可以是不同的。

外模式是保证数据库安全性的一个有力措施，每个用户只能看见和访问所对应的外模式中的数据，数据库中的其余数据是不可见的。

2．概念模式

概念模式又称为模式或逻辑模式，是数据库中全部数据逻辑结构和特征的描述，是所有用户的公共数据视图。一个数据库只有一个概念模式，通常以某种数据模型为基础，综合地考虑所有用户的需求，并将这些需求有机地结合成一个逻辑整体。定义概念模式一方面要定义数据的逻辑结构，例如数据记录由哪些数据项构成，数据项的名称、类型、取值

范围等；另一方面还要定义数据项之间的联系，定义数据记录之间的联系以及定义数据的完整性、安全性等要求。

3．内模式

内模式也称为存储模式或物理模式，是对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。内模式是数据库最低一级的逻辑描述，它定义所有内部数据类型、索引和文件的组织方式，以及数据控制等方面的细节。例如，记录的存储方式是顺序存储，还是按照 B 树存储或按 Hash（哈希）方法存储；又如数据是否压缩存储、是否加密等。

4．两级映像

为了能够实现数据库的 3 个抽象层次的联系和转换，数据库管理系统在这三级模式之间提供了两级映像。

外模式/概念模式映像：对于每一个外模式，数据库系统都有一个外模式/概念模式映像，它定义了该外模式与概念模式之间的对应关系。当概念模式改变时，如增加新的关系、新的属性或改变属性的数据类型等，只需要数据库管理员对各个外模式/概念模式映像做相应的改变，就可以使外模式保持不变。应用程序是依据数据的外模式编写的，从而应用程序可以不必修改，保证了数据与程序的逻辑独立性。

概念模式/内模式映像：该映像是唯一的，它定义了数据库的全局逻辑结构与存储结构之间的对应关系。当数据库的存储结构发生改变的时候，如数据库选用了另一种存储结构，此时只需数据库管理员对概念模式/内模式映射做相应的改变，就可以使概念模式保持不变，从而应用程序也不必修改，保证了数据与程序的物理独立性。

外模式/概念模式映像一般在外模式中描述，概念模式/内模式映像一般在内模式中描述。

对于数据独立性有如下两点说明：一方面，由数据与程序之间的独立性，使得数据的定义和描述可以从应用程序中分离出去，由于数据的存取由数据库管理系统管理，用户不必考虑存取路径等细节，从而简化了应用程序的编制，大大减少了应用程序的维护和修改工作；另一方面，由于数据库三级模式结构使数据库系统达到了高度的数据独立性，但同时也给系统增加了额外的开销，也就是说，需要在系统中保存和管理三级结构、两级映像的内容，并且数据传输要在三级结构中来回转换，从而增加了存储空间和运行时间的开销。

1.3 数 据 模 型

视频讲解

计算机不能直接处理现实世界中的客观事物，所以人们必须事先将客观事物进行抽象，将之组织成为计算机最终能处理的某一数据库管理系统支持的数据模型。数据模型（data model）包括数据库数据的结构部分、操作部分和约束条件，它是研究数据库技术的核心和基础。

1.3.1 数据处理的 3 个阶段

人们把客观存在的事物以数据的形式存储到计算机中，经历了对现实生活中事物特性的认识、概念化到计算机数据库中的具体表示的逐级抽象过程，这就需要进行两级抽象，

首先把现实世界转换为概念世界，然后将概念世界转换为某一个数据库管理系统所支持的数据模型，即现实世界—概念世界—数据世界 3 个阶段。有时也将概念世界称为信息世界，将数据世界称为机器世界。其抽象过程如图 1.6 所示。

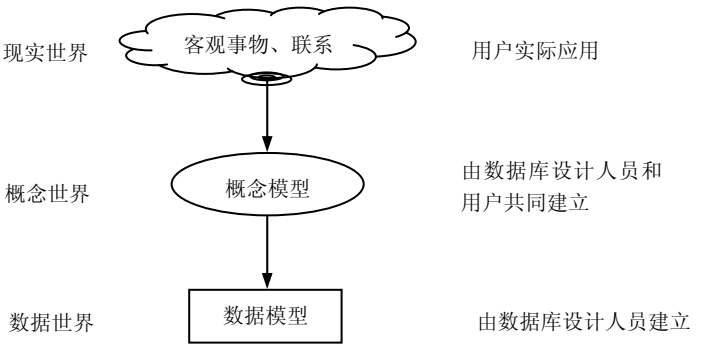

图 1.6　现实世界到数据世界的抽象过程

数据模型是现实世界中数据特征的抽象，它表现为一些相关数据组织的集合。在实施数据处理的不同阶段，需要使用不同的数据抽象，包括概念模型（conceptual model）、逻辑模型（logic model）和物理模型（physical model）。

1．概念模型

概念模型也称为信息模型，是对现实世界的认识和抽象描述，按用户的观点对数据和信息进行建模，不考虑在计算机和数据库管理系统上的具体实现，所以被称为概念模型。概念模型是对客观事物及其联系的一种抽象描述，它的表示方法很多，目前较常用的是美籍华人陈平山（Peter Chen）于 1976 年提出的实体-联系模型（Entity-Relationship Model，E-R 模型）。E-R 模型是现实世界到数据世界的一个中间层，它表示实体及实体间的联系。涉及的基本术语有：

（1）实体（entity）。客观存在、可以相互区别的事物称为实体。实体可以是具体的对象，例如一名男学生、一辆汽车等；也可以是抽象的对象，例如一次借书、一场足球比赛等。

（2）实体集（entity set）。性质相同的同类实体的集合，称为实体集。例如所有的男学生，全国足球锦标赛的所有比赛等。有时，在不引起混淆的情况下也称实体集为实体。

（3）属性（attribute）。实体有很多特性，每一个特性称为一个属性。每一个属性有一个值域，其类型可以是整数型、实数型、字符串型等。例如，实体学生有属性学号、姓名、年龄、性别等。

（4）实体标识符（entity identifier）。能唯一标识实体的属性或属性集，称为实体标识符。有时也称为关键码，或简称为键（key）。例如，学生的学号可以作为学生实体的标识符。

例如，学生实体由学号、姓名、性别、年龄、学院等属性组成，具体学生“吕占英”的信息（'S15'，'吕占英'，'女'，21，'CS'）是一个实体，其中 S15 表示学生的学号，CS 表示“计算机学院”。2017 级计算机学院的全体同学的数据集为一个实体集。其中学号、姓名、性别、年龄、学院等是实体的属性，S15 是实体标识符。