第一章　数据库系统导论

1.1　引　　言

1.1.1　数据与数据处理

人们在日常生活中接触到大量数据，如银行帐目数据、学校的教学管理数据、企业的生产管理和产品销售数据等等。这些数据都是人们生活中所离不开的，它们可以被视为人类社会中一种极其重要的资源。

然而这些数据之所以有价值，就是因为这些数据是表现信息，是载荷信息的物理符号。简单地说，信息是向人们提供关于现实世界新的事实的知识，它反映了客观事物的物理状态。例如，某一国家的汽车产量是 30 万辆，这个 30 万辆是一个数据，而它又表示了这个国家生产汽车的能力这一信息。由此可见，信息是用数据来表示的，但二者是即不可分离，又有一定区别的概念。一方面并非任何数据都能表示信息，信息只是消化了的数据，信息是依赖于数据而存在的。另一方面，信息是更基本的、直接反映现实的概念，数据是信息的具体表现。

对于数据进行综合推导，得出新的数据，这些新的数据则表示了新的信息。例如，在商品生产中，生产管理者根据某种商品历年销售数量及最近的市场需求调查，获得了许多数据。再对这些数据进行加工，就得出了有关商品的市场预测信息。生产管理者就可以根据这些信息进行分析和评价，作出对该产品是增产、减产还是停产的决策，并指挥生产部门的生产活动。这种从收集数据，到加工成信息进行评价和决策，再指挥实践活动，从而产生新的数据的循环过程称为信息循环。

诸如此类围绕信息所做的系统化的操作，称为信息处理。一般地说，信息处理就是指信息的收集整理、加工、归并、分类、计算、汇编、编辑、存储和传播等一系列活动的总称。因为信息是用数据表示的，所以对信息的处理又具体地体现在对数据的处理上。通过对原始数据（即尚未作评价的各种信息）的处理，产生新的数据（新的信息），作为结果。这一处理，包括对数据的收集、记录、分类、排序、存储，计算、加工、传输、制表和递交等处理。经过处理的数据是精炼的数据，是能够反映事物或现象的本质和特征及其内在联系的数据。

数据处理的历史可以追溯到远古时代。原始人类的结绳记数，累石记数便是数据处理的雏形。随着社会生产和文明的日益发展，信息的概念就趋于越来越复杂和深化，信息已经支配着人类的整个社会活动，所以有人把现代社会称为信息社会。研究信息的形态、传输、处理和存储理论的信息科学也就应运而生。随着信息概念的深化和发展，数据处理也就得到了相应的发展，就整个数据处理的发展过程来看，大致可以分成三个阶段。

第一阶段是数据的手工处理阶段。这是自远古时代到 19 世纪末一直使用的数据处理方法。由于当时的社会生产力低下，科学技术不发达，因此，数据处理的方法也就保持着简单、低级的手工操作。人们只能借助于如算盘、手摇计算机、计算尺等低级的计算工具处理数据。

需要全本请访问 www.ertongbook.com

这种方法的效率低，能处理的数量少，而且可靠性差。随着生产力和科学技术的发展，依靠手工操作来处理数据的方法，已经不能适应社会的需要，因而人们着手进行计算工具的改革和发明创造。

第二阶段是数据的机械处理阶段。这一阶段的主要特征是使用了比第一阶段先进得多且比较有效的工具。例如，H·霍勒内斯在19世纪80年代发明了卡片制表机。这种机器能以自动方式进行卡片的穿孔、校验、分类、整理和制表等操作。这是跨入机械数据处理阶段的一个重要标志。机械设备的使用，使数据处理的能力有了很大的改进和提高。但是这种设备的性能和使用，仍然受着很大的限制。这一阶段仍然有着较多的手工操作。直到20世纪40年代中期，第一台电子计算机ENIAC问世以后，数据处理才进入了自动化的电子数据处理时代。

第三阶段是电子数据处理阶段。电子计算机（简称计算机）的出现，为数据处理展现了广阔的前景。它不仅处理速度快、存储容量大、输入输出灵活，而且把人的手工操作降低到最小程度。电子数据处理方法不仅适应了不断提高的社会生产力的需要，而且给社会生产力的发展以推动和促进。因此，计算机的应用是数据处理领域中的一场革命。

尽管这3种数据处理的方法一个比一个先进，但是目前世界上仍然存在着手工和机械处理方法，特别是手工处理方法还是屡见不鲜的。而机械方法相对来说目前很少使用了。这一方面是因为某些领域的某些部门还不适宜使用计算机，另一方面是因为世界上各个国家和地区的发展还不平衡。这种局面将随着超小型和微型计算机的发展而得到改善。

人类活动的整个历史，离不开对信息和数据的收集、保存和处理。像银行、统计、仓库等部门的工作人员，无时无刻不在和大量的数据打交道。然而，随着社会的不断进步，人类社会积累的信息，正以“几何级数”的速度增长。因此人们过去传统的、落后的数据处理方法，已经远远适应不了形势发展的需要。所以长期以来，如何妥善地保存和管理数据，一直是人们十分关注的课题。人们对数据处理现代化的要求日益迫切。

计算机的出现，为人类处理数据自动化提供了可靠的物质基础。初期，计算机只是单纯为科学计算服务，所以人们又把它称为数字计算机。随着计算机硬件和软件的发展，计算机硬件成本的不断降低和各种新型外部设备的问世，特别是大容量磁盘存储器的生产和应用，人们逐渐地认识到，除了科学计算之外，计算机还具有强大的处理数据的能力，因为计算机的外存储器是保存数据资料的理想场所。于是，计算机的应用走出了科学计算这一传统领域，在数据处理这个广阔的领域大显身手。 60年代后期，数据库技术出现了。数据库技术研究的对象，就是高效地、巧妙地进行数据处理，而又开销最小。

目前，在整个计算机应用领域中，数据处理还是最庞大、最活跃的应用领域。在发达国家中，数据处理约占80%～90%。计算机工业历史说明，计算机是在科学计算领域中诞生，却在数据处理领域中站稳了脚跟，找到了真正的立足点。而数据库系统是数据处理的核心机构，数据库管理系统的最新成果，反映了数据处理技术的最新水平。有人用这样一段话概括地说明了数据库系统的地位：“数据库管理是现代计算机系统提供的最重要的功能。事实上，其重要性已经达到了这样的程度，即它已普遍成为购买计算机的主要出发点”。

一个计算机化的电子数据处理系统（Electronic Date Processing System，简称EDP系统），包括了人和计算机系统。它执行输入、处理、存储、输出和控制功能，用于完成对数据的一系列操作过程。一个计算机化的电子数据处理系统，一般由三个基本系统组成，它们是电传

处理系统、数据管理系统及科学计算系统。如图 1.1 所示。

电传处理系统的基本功能是，在适当的地点和时间提供必要的数据，即由它完成数据提供、通信（传输）和传递。

数据管理系统完成对数据的组织、编目、定位、存储、检索和维护等功能。

科学计算系统作为数据处理系统的一部分，会大大地加强系统的处理能力。它强调使用推理过程产生信息，好象纸和笔一样，成为人们在分析问题时的有力工具。但是有些复杂的推理问题，单靠纸和笔是不能解决的，而要靠科学计算系统来完成，因为它有助于提高人们的逻辑推理和通信的功能。例如在一个企业中，它能够用来解决某些管理决策、计划编制和资源最优分配等问题。

由于数据处理系统的核心是数据，因此数据管理系统是计算机数据处理系统中最重要的组成部分，而数据库管理系统则是一种新的数据管理软件，它提供了一种新的数据管理方法。

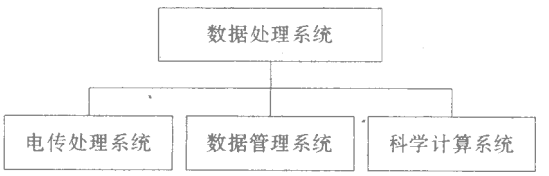


图1.1 数据处理系统

1.1.2 数据管理技术的发展过程

数据以及数据之间的关系可以从逻辑的物理两方面进行描述和组织。数据的逻辑描述是指按用户要求的数据间逻辑关系来组织和表达数据的，而物理描述则将涉及数据在计算机内的存储方式，是以存储数据的时空概念来描述数据间的关系的，前者可称为数据的逻辑结构，后者则称为数据的物理结构。

如何用计算机对数据进行组织、编目、定位、存储、检索和维护是数据管理的任务。数据管理是数据处理的中心问题。数据管理技术与数据处理方式有着密切的关系，并且直接影响着数据处理的效率。在数据处理的手工处理阶段和机械处理阶段，数据管理技术也是手工的。但是，在计算机进入数据处理领域之后，原来的那套手工管理方式就不能适应计算机自动处理数据的需要了。而且，在跨入信息社会的今天，需要管理的数据越来越多，如果仍然沿用原来那套方式，不仅需要耗费大量的人力，而且很难使这些数据发挥应有的作用。为此，许多计算机专家，特别是软件工作者，对数据管理技术做了大量的研究工作，使数据管理技术经历了一个由简单到复杂的不断完善的发展过程。

1.1.2.1 人工管理阶段（50 年代中期以前）

在这一阶段中，计算机除了硬设备之外没有任何软件可供使用，放在用户面前的只有硬件的裸机。用户在这种裸机上进行数据处理，除了编制自己的课题程序外，还必须考虑数据的逻辑定义和组织，以及数据在计算机存储设备内的物理存储方式和地址。数据的引用是按物理地址进行的。这时，表示处理流程的程序和作为处理对象的数据，相互结合成一个整体。数据的管理基本上是手工的、分散的，计算机无法有在数据管理中发挥应有的作用。因此，严重地影响了计算机的使用效率。这一时期数据和程序之间的关系可以由图 1.2 表示。

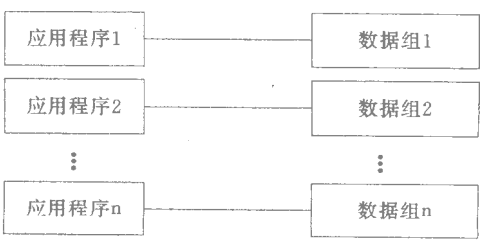


图1.2 人工管理阶段数据和程序之间的关系

1.1.2.2 文件系统阶段（50年代后期到60年代中期）

把数据组织成文件的形式，是计算机数据管理的重大进展，也是计算机在数据管理中直接发挥作用的开始。众所周知，无组织的一大堆数据是毫无意义的，也是没有实用价值的。该软件将数据按照一定的规则组织起来，成为一个有效的数据组合体，赋予它一个名字，称为文件名或文件标识，用户通过文件名来访问该文件。这时用户就不必过多地考虑物理细节了，而是由文件系统充当应用程序和数据文件之间的接口。这一时期应用程序和数据的关系如图1.3所示

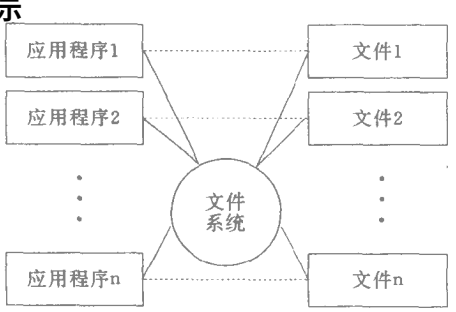


图1.3 文件系统阶段数据和程序之间的关系

虽然这一阶段较人工管理阶段有了很大的改进，但仍存在很大的弱点，即文件本身还是基本上对应于一个或几个应用程序，或者说数据还是面向应用的。从逻辑结构上讲，还与前一阶段没有什么变化，尽管程序不必要直接与文件打交道而有软件做为接口，但它仍然是一个不具有弹性的无结构的信息集合，存在着冗余度大、空间浪费及文件不易扩充等缺点，因此反映不了现实世界事物之间的广泛的内在联系。

1.1.2.3 数据库系统阶段（60年代后期开始）

这一时期，磁盘技术已经成熟，并作为主要外存而广泛使用。计算机硬件的价格大幅度下降，可靠性增强，为数据管理技术的发展奠定了物质基础。另外，计算机用于管理，其规模更加庞大，从而使数据量急剧增加。对数据进行集中控制，充分提供数据共享的要求日益迫切。

在这样的背景下产生了一种新的数据管理技术即数据库技术。数据库技术克服了以前所有管理方式的缺点，试图提供一种完善的、更高级的数据管理方式。它的基本思想是解决多用户数据共享的问题，实现对数据的集中统一管理，具有较高的数据独立性，并为数据提供各种保护措施。

这一时期，数据库管理软件作为用户与数据的接口、程序和数据的关系如图1.4所示。

1.1.3 什么是数据库

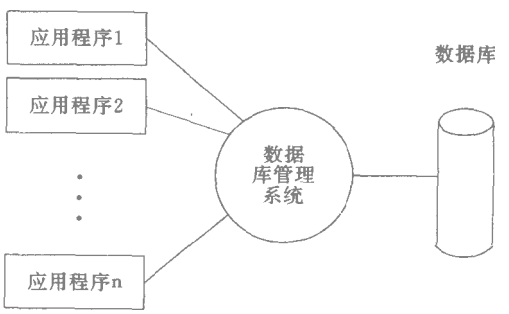


图1.4 数据库阶段数据和程序之间的关系

数据库系统是计算机软件的一个重要分支，是近20年来发展起来的一门新兴学科。它和计算机网络、人工智能被称为当今计算机技术界的三大热门技术。目前，虽有人对数据库的设计原则和方法进行总结和探讨，使之通用化、标准化和理论化。但总的说来，它的概念、原理和方法还处于从工程实践向理论过渡的阶段，它的概念、原理和方法还在继续变化和发展。另外，数据库是一个很复杂的系统，涉及面很广，难以用简练的语言准确地概括其全部特征。因此，给它下一个确切的定义是困难的，所以我们先从简单分析入手，逐步认识什么是数据库。

库，这个名词在日常生活中经常接触到。我们知道各种各样的库。例如：仓库、书库、金

库、血库等。稍微留心一下，我们就会发现，这些库都具有这样几个特点：

1. 可以存放现实世界中一定数量的东西。

2. 这些东西是有条理、有组织地存放在库中的，并由专人来进行管理。

例如，书库中的书就是有次序地摆在书架上的，并由图书管理员对它们进行管理。显然，库中的东西如果不是按照一定的次序存放，并由专人管理，那么就不能有效地利用库中的空间，也不可能方便地找出所需要的东西。

顾名思义，数据库就是存储数据的“仓库”。但它和我们前面讲的库是有所不同的。首先，数据不是存放在容器或空间中，而是存放在计算机的外存储器上（如磁盘），并且是有组织地存放的。数据的管理和利用是通过计算机的数据管理软件——数据库管理系统来完成的。因此，我们讲的数据库，不单是指存有数据的计算机外存，而是指存放在外存上的数据集合，以及管理它们的计算机软件的总和。这通常称为数据库管理系统。

目前，对于什么是数据库还没有一个统一的、公认的定义。比较流行的有关数据库的文献中常见的数据库定义有如下 3 种：

1. DBTG 的定义。数据库是由一个特定的模式控制的所有记录、系和城组成的。如果有多个数据库，则每一个数据库必须有自己的模式。并假定不同数据库的内容是彼此无关的。

2. C. J. Date 的定义。它从三个方面描述数据库：①存储在磁鼓、磁盘或其它外存介质上的数据集合——这是指数据库自身；②存在以这种数据为背景而运行的若干个应用程序，可进行检索、修改、插入和删除等操作，另外可能有一些联机用户，利用远程终端与数据库相互作用；③数据库是集成的，即包括许多用户的数据，每个用户只享用其中的一小部分，且不同用户使用的部分以多种方式重叠，也就是单独的数据片能够被许多不同用户所共享。

3. J. Martin 的定义。数据库是存储在一起的相关数据的集合，这些数据去掉了有害的或不必要的冗余，为多种应用服务，数据的存储独立于使用它的程序；对数据库插入新数据，修改和检索原有数据均能按一种公用的和可控制方式进行；数据被结构化，为今后的应用研究提供基础。当某个系统中存在结构上完全分开的若干数据库时，则说该系统包含一个数据库集合。

综合上述 3 种定义，通常一个数据库系统包含下列内容：

1. 有一个结构化的相关数据的集合。在这个数据集合中没有有害的或不必要的冗余，能够为多种应用服务，它独立于应用程序而存在。这种结构化的数据集合就是数据库本身，是数据库系统的核心和管理对象。

2. 有一个负责数据库管理和维护的软件系统，称为数据库管理系统（DataBase Management System—DBMS）。它对数据库中数据的各种操作，提供一种公用的方法。它接受并完成用户程序或终端命令提出的访问数据库的各种请求，负责数据库的维护，保护数据库中的数据不受破坏。

3. 有一个（或一组）负责整个数据系统的建立、维护和协调工作的专门人员，这就是数据库管理员（Data Base Administrator—DBA）。他们负责对数据库系统的全面管理与监督，并对运行状况进行统计分析，不断改进数据库设计。

图 1.5 表示了数据库系统的用户、数据库管理系统、数据库管理员和数据库之间的相互关系。

另外，还应该指出，数据库保存的是属于企业和事业部门，团体和个人的有关数据。例

需要全本请访问 www.ertongbook.com

如，银行帐目数据、学校的教学管理数据、政府部门的管理、统计和计划数据等。特别应该强调的是，设计数据库保存这些数据的目的，不仅仅是为了扩展人们的记忆，而主要是帮助人们去控制与之相关的事物。因此，数据库往往不是孤立存在的，通常是一个更大的信息系统的一部分，两者的关系如图 1.6 所示。该图表示，人们从观测客观事物中得到大量信息，对这些信息进行记录、整理和归类(总称规范),然后将规范信息数据化，送入数据库中保存起来，其中一部分信息可能直接送入控制决策机构。另一方面，控制决策机构（它既可由一些人组成，又可是一个自动控制系统）向数据库发出询问，并利用数据库响应后提供的信息（以及其它有关信息）作出决策，再行控制客观事物。例如，就一个企业的产品销售信息系统而言，客观事物可以是产品的销售情况和用户对该产品的反映情况等，控制决策机构指企业领导机关，策略是指提高产品质量扩大产品销售的措施。当数据库的管理人员把观测客观事物得到的信息规范化、数据化送入数据库后，企业领导机关可以通过询问数据库得到产品的销售情况和用户对产品的反映。根据这些情况，企业领导就可以提出改进产品销售,提高产品质量的措施,并在今后的生产经营中付诸实行(控制客观事物)。

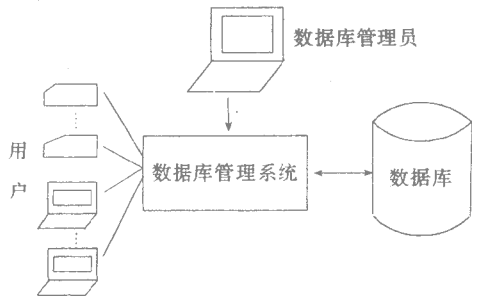


图 1.5 数据库系统

数据库系统作为软件的一个分支，与其它基础软件和系统软件有密切的关系。它几乎涉及到软件的所有知识，是许多重要软件技术的综合应用。例如，数据结构、操作系统、编译技术、程序设计等知识在数据库都会用到。所以数据库系统是一门综合性的软件技术，是一门很有意义很有趣味的学科。要更好的研究掌握它，必须了解软件的各个方面，以便更加理解和认识到这些知识的内在联系，并在一种观念上将它们统一起来。

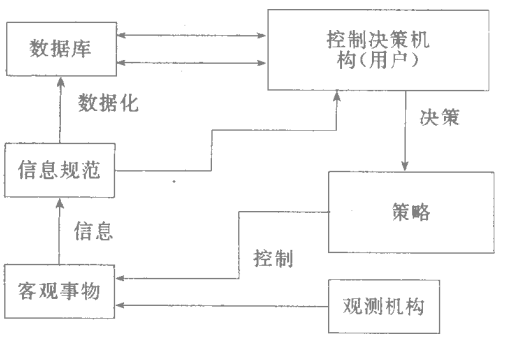


图 1.6 信息控制系统

1.1.4 数据库系统的主要特征

数据库技术之所以能够在近 20 年内如此快速的发展，受到计算机科学界普遍的重视，成为引人注目的一门新兴学科，是因为它具有独特的特征。

1.1.4.1 实现数据共享

实现数据共享是促成数据库技术发展的一个重要原因，也是数据库技术先进性的一个重要体现。数据的共享性包括 3 个方面：

1. 当前的所有用户可以同时存取数据库的数据，这是十分清楚的。因为数据库的研制正是以当前用户为现实基础的。

2. 未来用户与当前用户可以同时存取数据库中的数据。也就是说，数据库不仅可以为当前已知的多个用户服务，也可为将要使用该数据库的任何新用户服务，而且可以同时为多个新用户服务。

3. 用户可以通过多种程序设计语言或命令语言与数据库接口。如 COBOL、PL/1、FORTRAN，以及终端命令和数据库管理系统提供的数据操纵语言等等，从而可以用多种语言使用数据库。

1.1.4.2 减少数据的冗余度

在数据库技术产生以前的文件系统中，用户根据各自的应用建立文件，即使不同的应用程序需要使用的数据中也有许多是相同的，也必须建立各自的文件，其原因就在于文件系统中数据不能充分共享（文件系统中只提供文件一级的共享，不提供记录级的共享）。因此可能使存储的数据存在着大量的重复，这种现象称为数据的冗余。

下面我们通过一个例子来说明这个问题。

假设现在要管理的对象是一个学校，它由人事子系统、业务子系统、总务子系统组成，各子系统根据自己的特点及需要选择相应的数据格式，存放自己感兴趣的数据。它由这样 3 种记录组成，如图 1.7 所示。



图 1.7 学校的 3 个记录

在文件系统阶段，由这 3 个记录的集合组成 3 个文件：人事文件、业务文件、总务文件。这 3 个文件对应于 3 个不同的应用系统，它们之间是互不联系的。在这 3 个文件中，有许多项目如部门、职工号、姓名、性别、年龄、职务、工资是相同的。但是在文件系统阶段中，数据与应用系统是一一对应的，所以必须分别存储，这样势必造成很多的数据冗余，浪费存储空间。而这种冗余还会带来其它问题，如在更新操作时，必须要进行多次更新操作，以更新某些冗余的副本，增加不必要的机器时间。另外，由于多个副本的存在，系统容易出现不一致的信息。而在数据库中，是从整体观点来组织数据的，其数据不是面向个别的应用，而是为多种应用所共享的，从而避免了不必要的冗余。例如上面的例子在数据库中 3 个文件就不再是互不相干的了，而是将这 3 个文件视为统一的整体。也就是说，把这 3 个文件进行分解，重新组合，去掉重复部分，构成一个整体。图 1.8 给出经过重新组合形成的数据结构。

从图中可以看出，我们设计了各部门都要用的职工记录，并带上了人事记录、业务记录和总务记录等，这样就把整个组织的数据结构化了。这就要求在描述数据时，不仅要描述数据本身，还要描述数据之间的联系，这种联系往往是通过存取路径来表现的。在这种情况下就大大地减少了数据的冗余度，节省空间，提高存取速度，避免不相容。此外，由于从整体

的观点看数据，所以数据不是面向应用的，而是面向系统的。这样它的弹性大，可以扩充，应用数据时可以有灵活的方式，可以取整体模型的各种合理的子集，用于不同的应用系统。

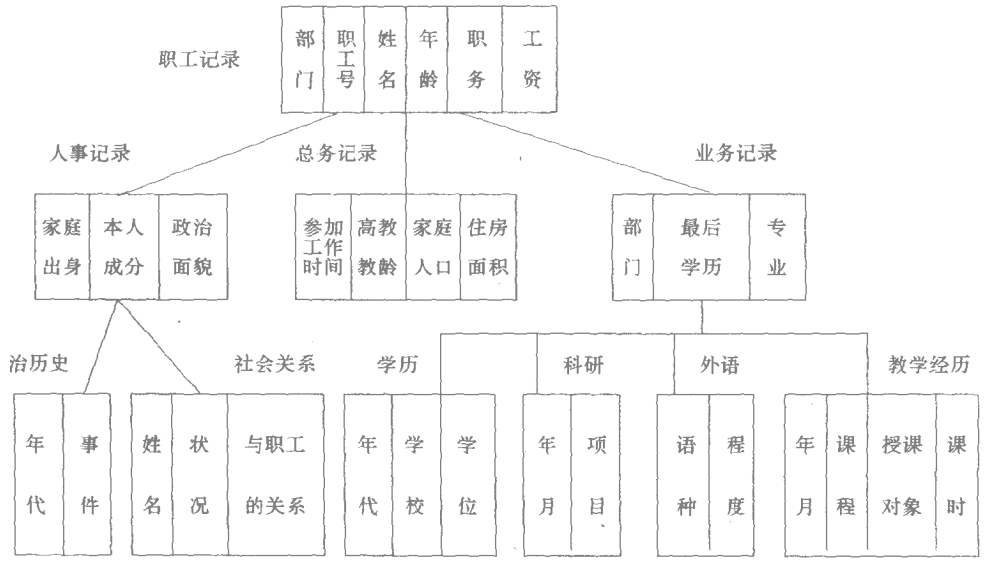


图 1.8 组合后的数据结构

从理论上说数据库中是没有冗余的，但实际上，许多数据库为了改善访问时间，或者由于采用比较简单的存取方法，数据库中还存在着某种程度的冗余性。但是它去掉了有害的或不必要的冗余。

1.1.4.3 维护数据的独立性

所谓数据的独立性，就是应用程序不必因为数据的存储结构的变化而修改，即应用程序与数据的结构之间不存在依赖关系，这是数据库系统所努力追求实现的一个目标。数据库系统结构之所以复杂，这是一个重要原因。数据库系统的数据独立性分为两级：

1 物理独立性。数据库的物理结构的变化（如物理设备的更换、物理位置的变化、存取方法的改变等等），不影响数据库的逻辑结构，从而也就影响不到应用程序，不会导致应用程序的修改。

2 逻辑独立性。数据库逻辑结构的变化（如数据定义的修改，新数据类型的增加，数据间联系的变更等等），不致影响到用户原有应用程序的修改。

这两种独立性统称为数据独立性。数据独立性的目的，就是应用程序尽可能不受数据的影响。这两种数据独立性是靠数据库管理系统来实现的，从而大大减轻了程序员的负担。

1.1.4.4 加强对数据的保护

数据的用户非常关心数据的安全性和可靠性。数据的安全可靠是一个数据库能否实用的关键问题。一般较完整的数据库系统对数据的保护有如下 4 个方面的内容：

1 安全保密控制。它主要是数据的保密性控制，即保护库中的数据以防止不合法的使用。一般可以采取如下 3 种方法，来保证数据的安全性：①将数据库中需要保密的部分与其它公用部分的数据隔离开来；②建立一些访问规则；③将数据以密码的形式存放于数据库内。

2 完整性控制。数据的完整性包括数据的正确性、有效性和一致性。一个数值型数据包

含了诸如字母、特殊符号、字符，显然是不正确。此外，形式正确的数据也并非一定是有效的，例如月份必须是 1～12 的正整数，否则是无效的。表示同一事实的两个数据项之间的不一致性也是缺乏完整性的例子。数据库系统对数据的使用是集中控制的，因而有利于控制数据的完整性。

3. 并发控制。数据的不相容性主要是由于数据的共享而引起的。不同用户同时使用数据库，可能引起对数据的干扰。例如，当某个用户需要修改某一数据时，为此他必须先读出这个数据，修改后再写回去，如果在这个用户把被修改的数据写回数据库之前，别的用户再去读出这个正在被修改的数据就不合理了，就会引起错误。数据库管理系统的并发控制功能，将排除和避免这种错误的发生，保证数据的正确性。

4. 数据库恢复。数据库在运行过程中很难保证不受破坏，全局性的和局部性的破坏随时都可能发生。因为硬件和软件故障以及用户操作上的错误都是不可避免的。为了防止由于这些故障而破坏数据的正确性，数据库系统通过下述 4 项措施来保证数据的正确性：①定期将数据库中的内容拷贝到磁带上加以保存；②对每次应用数据库的过程进行记录，以便出错时检查错误的来源；③每当必须对数据库的内容进行修改时，将修改前的内容及修改后的内容记到磁带上保存起来，以备检查或重作时使用；④当发现对数据库的某个处理有错误时，消除有关处理的影响，以恢复正确的数据库。

1.1.4.5 对数据实行集中控制

由前述可知，文件管理方法使数据的管理处于一种分散的状态中。对数据无法按照一种统一的方法来控制、维护和管理。在数据库中克服了文件系统的缺点，对数据进行集中控制和管理。一个组织或一个企业，可以把它们所有要涉及到的数据都组织到一个数据库中。也就是说，数据库是“集成”的。但这里所说的“集成”并不是把数据简单地集中起来，而是对数据进行结构化，使其能适应所有用户的要求。

1.2 现实世界的数据描述

对用户来说，数据库是模拟现实世界中某些企业活动的信息集合。数据库中所存储的数据，来源于现实世界的信息流，是用来描述现实世界中一些事物的某些方面的特征及其相互联系的。例如，在一个工厂的生产活动中，除了购买原材料与设备、组织人力、加工生产、产品出厂、资金周转的活动外，还有生产计划、购销合同、生产进度表、产品清单、各种发票、财务帐目等信息流。我们在处理这些信息之前，必须先分析它，选择一种方法描述这些处理对象，并将这种描述转换成计算机所能接受的数据形式。

从图 1.6 中可以看出，在信息系统中，信息从客观事物出发，流经数据库，通过控制决策机构，最后又回到客观事物。信息的这一循环经历了 3 个领域：现实世界、信息世界、计算机世界。

1.2.1 现实世界

现实世界是指存在于人脑之外的客观世界，泛指客观存在的事物及其相互间的联系。一个实际存在并且可以识别的事物称为个体。个体可以是一个具体的事物，如一个学生、一台计算机，一辆汽车等等，也可以是一个抽象的概念，如年龄、性格、爱好等等。

每个个体都有自己的特征，用以区别其它个体，如学生有姓名、性别、年龄、身高、体

重等许多特征来标识自己，但是在我们研究个体时，往往只选择其中对之有意义的特征。例如，对于人事管理，选择的特征可以是姓名、性别、年龄、工资、职务等特征，而在描述一个人健康情况时，可以选用身高、体重、血压等特征。

我们把具有相同特征要求的个体称为同类个体，所有同类个体的集合称为总体。例如所有的“学生”、所有的“课程”、所有的“汽车”都是一个总体。

所有这些客观事物是信息的源泉，是设计数据库的出发点。

1.2.2 信息世界

现实世界中的事物反映到人们的头脑里，经过认识、选择、命名、分类等综合分析而形成了印象和概念，产生认识，这就是信息，即进入信息世界。在信息世界中，对每一个被认识了的个体称为实体，这是具体事物（个体）在人们头脑中产生的概念，是信息世界的基本单位。另外，个体的特征在头脑形成的知识称为属性。所以属性是事物某一方面的特征，即属性是反映实体的某一特征的。换句话说，一个实体是由它所有的属性表示的。例如，一本书是一个实体，可以由书号、书名、作者、出版社、单价 5 个属性来表示。

在信息世界里，主要研究的不是个别的实体，而是它们的共性。我们把具有相同属性的实体称为同类实体，同类实体的集合为实体集。

1.2.3 计算机世界

信息世界中的有些信息，可以直接用数字表示，如学生成绩、年龄、书号等；有些是由符号、文字或其它形式来表示的。在计算机中，所有信息只能用二进制数表示，一切信息进入计算机时，必须是数据化的。可以说，数据是信息的具体表现形式。在计算机世界中用到如下术语：

1. 数据项。数据项是实体属性的数据表示。
2. 记录。记录是实体的数据表示。
3. 文件。文件是同类记录的集合，如所有职工的登记表组成一个文件。
4. 数据模型。现实世界中，个体与个体之间总是存在着某些联系。反映到信息世界中，是实体与实体之间的联系，由此构成实体模型。反映到计算机世界中，是记录与记录之间的联系。实体模型在该领域的数据化表示，就是数据模型。

由此可见，现实世界、信息世界、计算机世界，这 3 个领域是由客观到认识、由认识到使用管理的 3 个不同层次，而且后一领域是前一领域的抽象描述。

关于 3 个领域之间的术语对应关系，可由图 1.9 表示。

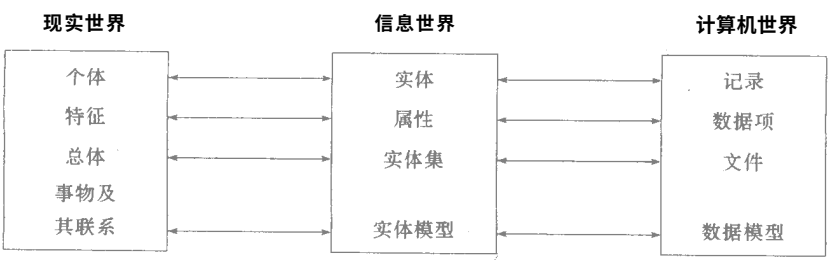


图1.9 3个世界的术语对应说明

1.2.4 常用术语的说明

1. 属性。属性为事物的某一方面特征的抽象描述。如学生这个事物，可以通过学生的姓名、学号、性别、年龄和政治面貌等特征来描述，称姓名、学号、性别、年龄和政治面貌为属性。

2. 属性值。属性值为属性的具体取值。如学生丁力，其姓名为丁力，学号为 79201，性别为男，年龄为 21，政治面貌为党员，这些具体值称为属性值。

3. 域。域为属性的取值范围，称之为属性的域。如学生的年龄为 16～30 范围内的正整数，其域为（16，30）。

4. 实体。若干个属性的属性值组成的集合，用来表征一个实体。如丁力、 79021 表示学生丁力。同类实体的集合则组成了实体集，例如描述全部学生的实体就构成了学生实体集，反映了一个学校全部学生的情况。

5. 关键字。关键字为唯一标识实体的属性或属性组合。例如在学生实体集中，学号可以唯一地标识每个学生实体，所以学号为关键字。在有些实体集中，可以有多个关键字，例如学生实体集，假设学生姓名没有重名，那么属性“姓名”也可以作为关键字。我们通常选定其中一个，被选的那一个关键字称为主关键字，其它的关键字为候选关键字。

6. 次关键字。次关键字，在实体集中经常选择一些不能唯一标识实体的属性来标识实体，这样的属性叫次关键字。例如“年龄”，“政治面貌”这些属性都可以充当次关键字。

1.3 数据模型

现实世界中的事物是彼此关联的，任何一个实体都不是孤立存在的，因此描述实体的数据也是相互联系的。通常，这种联系表现在两个方面：一是反映实体内部的联系，反映在数据上便是记录内部各数据项之间的联系，关于这个问题已经在《数据结构》一书中做了详尽的讨论，这里不再赘述；一是实体之间的联系，反映在数据上则是记录之间的联系，这种联系在数据模型中给予描述。

数据系统的一个核心问题，就是研究如何表示和处理实体之间的联系。而我们把表示实体及其实体之间联系的数据库的数据结构称为数据模型。换句话说，我们把数据库系统中所包含的所有记录类型，按照它们之间的联系组合在一起，构成一个整体。这个整体的结构就称之为数据库的数据模型，它是数据库中数据全局逻辑结构的描述。

数据模型是数据库中一个关键的概念，它是实体间联系的一个轮廓视图。数据模型的所有术语都用在模型一级上，整个模型就像一个框架，给它填上具体的数据值就是数据模型的一个实例。

通常，数据模型可以表示成由这样两个集合组成：

数据模型 DM＝（R，L）

其中 R 表示记录型的集合；L 表示记录型之间联系的集合。

如果按照记录间联系的表示方式，对数据模型加以分类的话，那么数据模型一般可以分成三种：层次模型、网状模型及关系模型。前两种又称为格式化数据模型。

数据模型的好坏，直接影响数据库的性能。数据模型的选择，是设计数据库的一项首要任务。

需要全本请访问 www.ertongbook.com

1.3.1 层次模型

在数据库中，把可以用图（这里用有向图）来表示的数据模型，称为格式化数据模型。而格式化数据模型又按照图的特点分为层次模型和网状模型两种。这两种模型都是以记录型作为图中的结点，而图中的边则用来表示记录型之间的联系。

层次模型是以记录型为结点的有向树。在树中，把无双亲的记录称为根记录，其它记录称为从属记录。除根记录外，任何记录只有一个父记录。一个父记录可以有多个子记录。从根记录开始，一直到最下一层的记录为止，所具有的层次称为该数据模型的层次，或称层次模型（见图 1.10）。

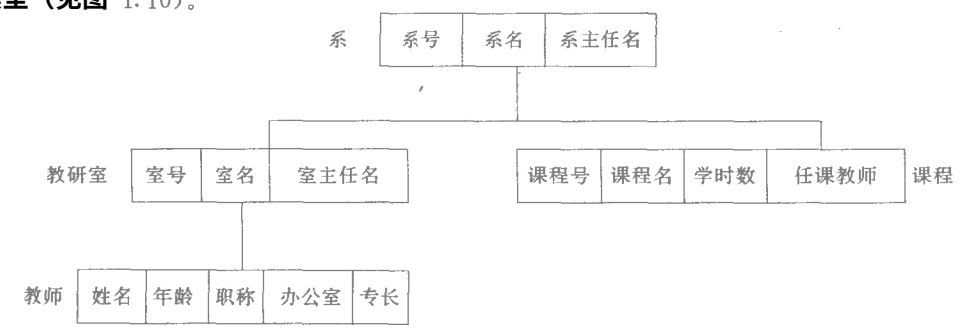


图 1.10 教育层次模型

在层次模型中，每个记录（除根记录外）只有一个双亲，记录之间的联系可以唯一地用双亲表示。在层次模型中，总是从双亲记录指向子记录。所以说，记录之间的联系可以不用命名，只要指出其双亲，就可以找到其子。在层次模型中，从根记录开始，按照双亲-子联系，依次连接的记录序列称为层次路径。在层次模型中，数据是按层次路径存取的。

目前流行的大型数据库系统中，有许多采用的是层次模型。其中最著名的是 IBM 公司研制的 IMS 系统。该系统之所以应用较广，除了它问世较早外，还由于 IBM 公司具有强大的竞争能力。除了 IMS 系统之外，还有许多采用层次模型的系统，如系统 SYSTEM2000 就是其中的一个。

1.3.2 网状模型

数据库的网状模型，是以记录类型为结点的网状结构。这种结构必须满足如下条件：

1. 可以有一个以上的结点无双亲。
2. 至少有一个结点有多于一个的双亲。

图 1.11 所示的是学生的网状模型。

网状模型和层次模型的差别是：

1. 一个子结点可以有两个或多个父结点。
2. 在两个结点之间可以有两种或多种联系。

采用网状模型的数据库系统，大多依据 DBTG 的文本。DBTG 文本是在 1968 年 1 月推出的，目前已用在许多机器上。几乎除了 IBM 公司以外，其它大厂商生产的中型以上的计算机系统都配有网状模型的数据库系统。例如，Cullinane 公司的 IDMS 系统，就是依据 DBTG 文本实现的。

教师情况表

职工号	姓　名	年　龄	系　号	职　称	工　资
⋮	⋮	⋮	⋮	⋮	⋮
0613	张三	45	3	讲师	27
⋮	⋮	⋮	⋮	⋮	⋮
0895	李四	50	4	副教授	131
⋮	⋮	⋮	⋮	⋮	⋮
1234	王林	25	6	助教	70

学生情况表

学　号	姓名	年龄	系号	班　　级
80101	赵文	20	3	机械 86
80100	张生	19	6	计算机 87
80001	刘芳	21	1	电机 85-2
⋮	⋮	⋮	⋮	⋮

学生学习情况表

学号	课程号	分数
80101	J101	90
80100	J212	85
80001	J101	90
80101	J211	90
⋮	⋮	⋮

课程情况表

课程号	课程名	课时
J101	高等数学	144
J201	专业英语	54
J212	程序设计	72
⋮	⋮	⋮

教师任课情况表

职工号	课程号	班　级
⋮	⋮	⋮
0613	J101	机械 85
⋮	⋮	⋮
1234	J212	计算机 86

图 1.13　学院教务数据库

关系是元组的集合。如果表格有 n 列，则称该关系为 n 元关系。关系模型中的关系应有如下特点：

1. 关系中的每一列属性，都是不能再分的基本字段。
2. 各列被指定一个相异的名字。
3. 各行相异，不允许重复。
4. 行、列次序无关紧要。

关系模型与层次、网状模型相比，具有许多优点。C·J·DATE 说过这样一段话：关系方法可以看作是数据库理论的论点，它为理解和比较其它两种方法提供了极好的理论基础，也为评价任何现有系统提供了方便的标准或尺度。即使关系方法没有别的任何优点，仅仅理论上的正确性和永久性，也使之成为数学上的理想工具。

1.3.4　各模型的比较

在上面我们讨论了 3 种数据模型，即层次模型、网状模型和关系模型。这 3 种模型由于结构上的区别，它们都有各自的特点。为了评估这 3 种模型，必须先给出评价它们的标准，由此才能看出一个模型的好坏。通常我们主要关心这样两个方面：

1. 方便程度。由于数据库系统的用户是各式各样的，那么，提供给用户一个良好接口的数据库系统，将是十分受欢迎并具有生命力的。所以在选择模型时，则应选择一个用户使用方便的（即能方便程序设计和表达询问的）模型。

2. 实现效率。这方面要考虑系统的实现难易和系统的效率如何。通常我们需要这样一个模型：它允许 DBMS 方便地把概念模式和概念到物理的映象，转换成一种既能节省空间又能快速响应询问的实现。

就使用方便来说，毫无疑问，关系模型是最佳的。它对用户的要求很低。后面我们会看

到，它有功能丰富的、容许表达对关系数据库进行各种询问的高级语言。这些语言对于那些不太熟悉程序设计的人来说，是十分合适的。

相比较而言，格式化数据模型要求用户既要了解记录类型存取路径，又要了解它们之间的相互关系，这就势必增加了用户的负担。

从实现效率考虑，层次和网状模型要优于关系模型。格式化数据模型的存取路径，事先都是规定好了的，链技术就可大显身手。这样存取的效率就比较高，而且实现起来，相对地说也容易些。

从存储空间上说，格式化模型较关系模型更能合理利用空间。例如，在关系模型中就是靠冗余来实现连接的。

过去，商品化的系统几乎全都是基于层次和网状模型的。因为这样的系统着眼于大型数据库的维护，而这些模型又很容易支持它们对高效率实现的要求。然而，现在关系模型已经被人们重视了。这是因为：第一，用于设计大型数据库的概念，也适用于小型、中型数据库，但小型数据库比大型的要多得多。随着小型数据库的应用，关系模型固有的容易使用的特点将越来越重要；第二，关系模型的那些表面上的低效性，许多是能够消除的。例如，效率低的问题，可以通过优化得以提高等。另外，用层次和网状方法设计的数据库系统，是通过指针链查找数据的，而用关系方法设计的数据库系统，则是通过查表来查找数据的。由于改进指针链的查找收效不大，而加快查表速度还大有潜力可挖，所以促使许多人采用关系模型。

可以说，随着关系模型有关优化问题的进一步解决，关系模型将会更加吸引人。

1.4 数据库系统结构

1.4.1 数据库系统结构的基本概念

经过上面学习之后，这里给出一个数据库系统的整体结构，以便加深理解和掌握前面所学的各种概念。现在人们所习惯的并容易理解的数据库系统结构，是由 ANSI/X3/SPARC 的数据库管理系统研究组 1975 年提出的。目前世界上大多数数据库系统尽管种类不同，但它们的结构基本上是与下面介绍的 3 级结构相一致的。所以可以认为图 1.14 是数据库系统的一般结构。

数据库系统结构的 3 个抽象描述级，定义了数据库的 3 个层次，反映了看待数据库的 3 种不同角度。

下面我们对系统结构中用到的一些概念加以解释。

在文件系统中，用户对于他所用文件的物理组织、存储细节等要有相当的了解，这给用户带来了很大不便。数据库系统的一个目标，就是解决这个问题。它把一切琐碎事务都交给 DBMS 来处理。DBMS 的主要功能就是，允许用户逻辑地、抽象地处理数据，而不必涉及这些数据在计算机中是怎样存放的。

在数据库系统中，用户看到的数据和计算机存放的数据是两回事，当然这中间是有联系的。实际上它们之间已经有过了两次变换。第一次是系统为了减少冗余，实现数据共享，把所有用户的数据进行综合，抽象成一个统一的数据视图。第二次是为了提高存取效率，改善系统性能，把全局视图的数据按照物理组织的最优形式来存放。当计算机向用户提供数据时，则做相反的变换。

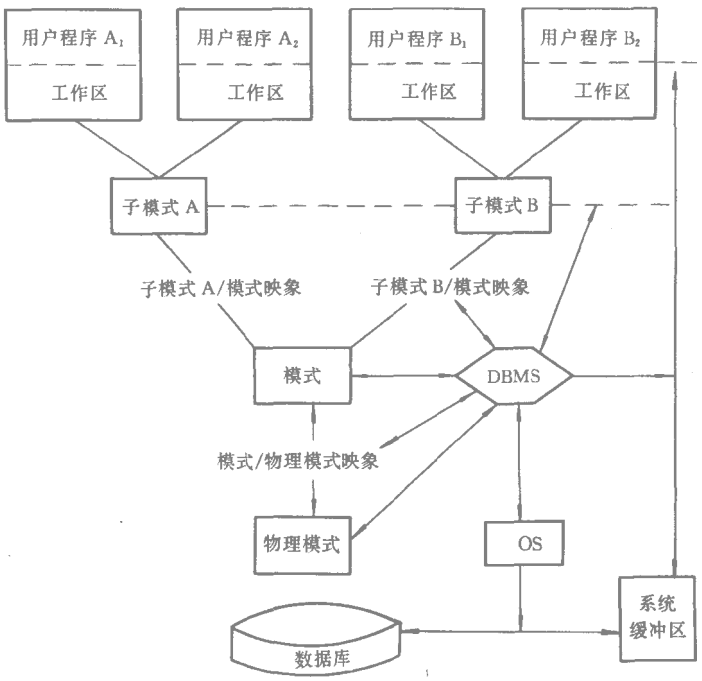


图 1.11 数据库系统结构

用户使用的数据视图叫子模式。子模式是一种局部的逻辑数据视图，它表示了用户所理解的实体、实体的属性和实体间的关系。

描述物理数据存储的模式叫内模式（物理模式）。

整个系统分为 3 层：外层、概念层和内层。用户只看到外层，其它 2 层他们是看不到的。子模式有多个，而概念模式和内模式只有一个。内模式是整个数据库的最低层。

用数据描述语言，精确地定义数据模型的程序叫模式。

1.4.1.1 子模式

定义用户视图的模式叫子模式，又叫外模式，用子模式数据描述语言来定义。具有相同数据视图的用户，共享用一个子模式，一个子模式可以为多个用户所使用。从逻辑关系上看，子模式是模式的一个逻辑子集，从一个模式可以推导出许多不同的子模式。

设立子模式的好处是：

1. 方便了用户的使用，简化了用户的接口。用户只要依照子模式，编写应用程序或在终端敲入命令，无须了解数据的存储结构。

2. 保证数据的独立性。由于在 3 级模式之间存在 2 级映象，使得物理模式和概念模式的变化，都反映不到子模式一层，从而不用修改应用程序，提高了数据的独立性。

3. 有利于数据共享。从同一模式产生不同的子模式，减少了数据的冗余度，有利于为多种应用服务。

4. 有利于数据的安全和保密。用户程序只能操作其子模式范围内的数据，从而把其与数据库中的其余数据隔离开来，缩小了程序错误传播的范围，保证了其它数据的安全。

1.4.1.2 模式

定义概念模型的模式叫概念模式，简称为模式，用模式数据描述语言来定义。它是数据库的整个逻辑描述，并说明了一个数据库所采用的数据模型。同时它还给出了实体和属性的名字，并说明了它们之间的关系，是一个可以放进数据项值的框架。

目前，模式中通常还包括寻址方式、存取控制、保密定义、安全性和完整性等方面的内容。

1.4.1.3 物理模式

内模式也叫物理模式，是用设备介质描述语言来定义的。它规定数据项、记录、数据集、索引和存取路径在内的一切物理组织方式，以及优化性能、响应时间和存储空间需求。它还规定记录的位置、块的大小与溢出区等。

无论哪一级模式都只能是处理数据的一个框架，而按这些框架填入的数据才是数据库的内容。但要注意的是框架和数据是两回事，它们放在不同的地方。所以模型、模式、具体值是三个不同的概念。此外还应知道：

1. 以子模式为框架的数据库是用户数据库；
2. 以模式为框架的数据库是概念数据库；
3. 以物理模块为框架的数据库为物理数据库。

这里必须强调指出，对一个数据库系统来说，只有物理数据库才真正存在，它是存放在外存上的实际数据，而其它两个数据库在外存上是不存在的。概念数据库只不过是物理数据库的一种抽象（逻辑）的描述，用户数据库则是用户需要使用的数据库子集的逻辑描述，是用户与数据库的接口。用户根据子模式进行操作，通过子模式／模式的映象与概念数据库联系起来，又通过模式／物理模式映象与物理数据库联系起来。所谓映象是一种对应规则，指出映象双方如何进行转换。数据库的 3 级结构就是靠映象来联结的。数据库管理系统的一项重要工作，就是完成 3级数据库之间的映象，把用户对数据库的操作，转化成对物理数据库的操作。

用户数据库是概念数据库的部分抽取；概念数据库是物理数据库的抽象表示；物理数据库是概念数据库的具体实现。

设计数据库时，设计者们主要考虑的是整个数据库的轮廓框架。使用数据库时，用户关心的是数据库的内容。数据库的框架是相对稳定的，而数据库的内容是经常变化的。

1.4.2 数据库系统的组成

在计算机系统中引入数据库技术之后，该系统就是以数据处理作为主要的应用特征。这里讲的数据库系统不是指数据库本身，也不仅指数据库管理系统，而是指计算机系统中引入数据库后的系统构成。在这种情况下，构成该系统通常有这样几部分：存储大量数据（包括系统数据和用户数据）的计算机系统；运行在该计算机系统上的各种数据管理软件和各种应用软件以及同该系统打交道的各种身份的人（这里统称为用户）。用计算机术语来说，该系统通常由 4 部分组成：硬件、软件、数据和用户。

有了数据库之后，整个计算机系统的软、硬件层次如图 1.15所示。

由此可以看出，在硬件外面包上操作系统是计算机系统的一般配置，而数据库管理系统是包在操作系统外面的一层软件，它是在操作系统的支持下工作的。

需要全本请访问 www.ertongbook.com