数据库系统概论

吴保国

北京林业大学
林业资源学院
一九八九年二月

序言

数据库技术产生于60年代末70年代初，它的出现使得计算机应用渗透到工农业、商业、行政管理、科学研究、工程技术和国防等每一个部门。它是现代计算机系统提供的最重要的功能。事实上，其重要性已达到这样的程度，它已普遍成为购买计算机的主要出发点。它是软件方面的一个独立分支，目前它正处于发展中。

计算机应用中，数据处理占的比重最大。管理信息系统和决策支持系统中，数据库系统是数据处理的核心机构。它的功能往往决定整个计算机应用的经济效益。因此，我们除了撑握具体的数据库管理系统使用外，还须学习撑握数据库系统的原理、技术和数据库设计，用它解决实际工作中的问题，提高数据库应用水平。本教材就是为了这一需要而编写的。

教材的编写主要参考了萨师煊、王珊的《数据库系统概论》，冯玉才的《数据库系统基础》、钟恢扶、王仲东的《关系数据库的设计与应用》等书，教材还广泛参考了其它国内外出版的数据库方面的书籍、文献、资料和论文，吸取它们中大量的、好的思想内容。

教材尽量避免过深的理论，着重介绍数据库系统的基本概念，力求通俗易懂。选材上尽量反映这一领域的新动态和当前状况，使读者能全面地、综合地了解数据库及发展动态。

目前，关系数据库是世界上最受欢迎的数据库，尤其中、小型关系数据库。如 dBASE Ⅲ。因此教材在第五章，第六章重点地介绍了关系数据库和关系模式的规范理论，而对网状和层次数据库没作过多的介绍。

教材第一——四章介绍了数据库的基本概念和构成，第七章介绍了数据库设计的全过程，第八章介绍了数据库的保护。

由于编者水平有限，加之时间仓促，书中错误和不妥之处一定不少，恳切希望读者批评指正。

编者

1989. 2. 10

目　录

第一章　什么是数据库

§1.1　数据和数据处理

一、信息和数据

在计算机的用语中，经常出现信息和数据这两个词。因此，在介绍数据库之前，我们首先说明一下信息和数据这两个基本概念。

信息的最简单定义就是“通知”。说完整一些就是“关于生活主体同外部客体之间的有关情况的通知”。信息是伴随着生物的诞生而诞生的。生物为了维持本身的生存，不断对外部环境和事物进行观察和测量以获取与它范围有关的情况通知，并予以识别、评价，然后采取适应外部环境的行动。例如：人们对大气的观察，识别和评价而得到当地天气情况和信息，并根据这些信息采取相应的行动。而天气情况的信息又是通过一定的形式表示和传播的。这就是数据的概念。数据是从观察或测量中所收集到的事实。对于数据，往往有人把它误解成仅仅表示数值概念的数据。其实，两者是不能等同的，数值数据仅仅是数据的一个子集，因为信息不仅需要“数量”的表达，而且还要求有“陈述”的表达。而且这种表达是大量的。因此，广义而言，数据是一切文字、符号、声音、图像等及有意义的组合。所以，有人认为，“信息是向人们（或机器）提供关于现实世界中有关事物的知识；数据则是用以载荷信息的物理符号”；又有人认为，“数据是记录下来的而且可以鉴别的符号；信息则是数据加工的结果，是对数据的解释”。不论何种说法，两个名词都是不可分割的，但又有一定的概念区别。在图1.1中，过程（Ⅰ）是后一种解释，信息不随载荷它的物理设备形式改变，过程（Ⅱ）是前一种解释，两种说法都成立。数据却不然，由于载体的不同，数据的表现形式可以不同。

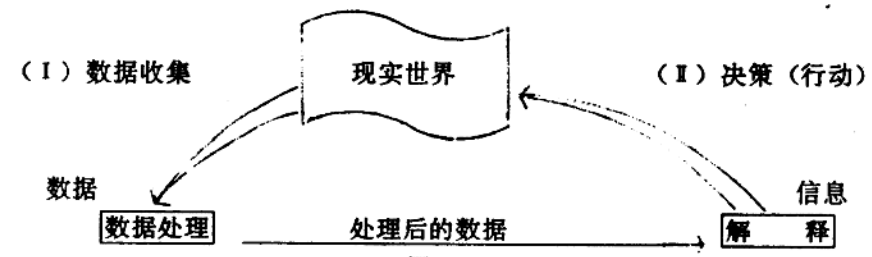

图 1.1

总之，信息和数据都是客观世界物理状态的反应，但数据则是反映“特定目的尚未作出评价的事实”而信息则是经过处理、加工后产生的有助于实现特定目的信息，即信息是消化了的数据。信息依赖于数据而存在，数据组合并且具体地表现信息。

但是，在一些不很严格的场合，人们把它们当作同义词。例如：数据处理和信息处理等等。

二、数据处理

从上面的讨论我们可以知道，信息用数据表示，对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：在森林资源管理中，我们通过外业调查获得了许多有关森林数据，对这些数据进行加工就得到了森林资源状况的信息，生产管理者就可以根据这些信息进行综合分析和评价，

从而进行指导林业生产。当这种生产进行后，我们又会重新上述过程。

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或信息反馈。象此类围绕信息所作的一系列工作，称为信息处理。因而我们称该信息处理就是指信息的收集、整理、加工、存贮和传播等一系列活动的总和。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。所谓数据处理是各种类型的数据进行加工处理和综合分析，它包括对数据进行收集、输入、存贮、传输、分类、排序、计算，并以多种形式进行输出，也包括对已存贮的数据进行检索和更新。数据处理的基本目的，就是将大量数据进行加工整理变为对人们有价值的信息。

数据处理的历史可以追溯到远古时代，远古时代的结绳记录，累石记数便是数据处理的雏形。随着社会生产、文明和科学的发展，信息的概念就越来越复杂和深化，信息已经支配着人类的整个社会活动。比如国际贸易中，如我们获得信息是错误的，则会给我们造成重大经济损失。所以有人把现代社会称为信息社会。

三、计算机在数据管理中的作用及其发展

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。当计算机进入数据领域中后，原来的那套手工管理方式就不适应计算机的自动处理的需要了。为此，许多计算机专家，特别是软件工作者就数据管理技术进行了大量的研究工作，并且取得了重大进展，发展了许多卓有成效的数据管理技术，就发展情况看，大至经历了三个阶段：

(1)单项数据处理阶段(50—60 年代中)

早期的计算机，由于硬、软件的限制，用户上机处理数据，除编制自己所用的程序外，还需考虑数据的逻辑定义和组织及在记算机存贮设备内的物理存贮方式和地址。数据的引用是按物理地址进行处理，功能及差。这个阶段一般为批处理，主要代替部分手工数据处理，效率极低。

(2)文件管理方式(60—70 年代初)

这时计算机硬、软件有了很大发展，计算机配上了操作系统，并且有了文件管理系统。它将数据按一定的规则组织起来，成为一个有效的数据组合体，并赋予它一个名字，和一个为文件名或文件标识，供以后访问该文件时使用。

文件系统是应用程序和数据文件之间的接口，应用程序通过文件系统对文件中的数据或对文件中的某些数据进行修改，如图 1.2

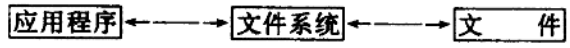

图 1.2

在文件管理方式中又可分为独立文件方式和共享文件方式二种。

独立文件方式中，各个用户建立各自的文件，如图 1.3 文件和文件应用程序之间有着密切的相应依赖关系。文件仅仅是数据的存贮，而数据的逻辑定义，物理存贮设备，组织方式和存取方法仍由程序选择、决定，并在程序中指出。数据与程序紧密相联，数据文件一旦离开了它所依赖的程序就会失去它存在的意义。一个应用程序所建立的数据文件，根本不能由另一个应用程序所共享。这就造成了大量数据重复，即冗余。它不能反应数据之间的内部联系，而这正是数据应具有的至关重要的性质。为了克服独立文件中的上述缺点，人们发展了共享文件方式。如图 1.4

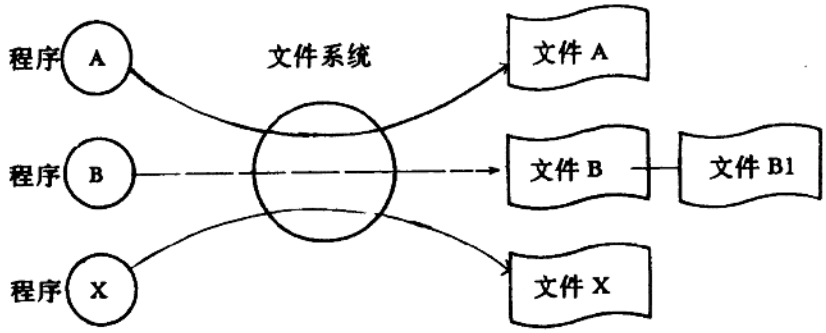

图 1.3

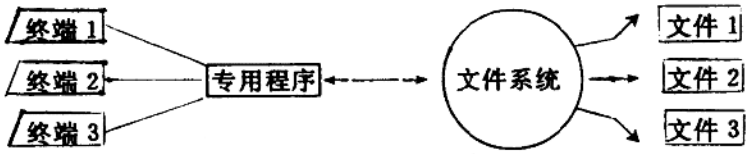

图 1.4

共享文件方式中，人们采用一个或几个综合性文件来存贮一组或一个部门内所有用户要用到的数据，而这种综合性文件的逻辑组织，物理特性由一个专门的公用用户程序来控制。任何一个有关应用处理都可以通过这个公用的用户程序去访问任何一个文件中的数据。它在一定程序上克服了独立文件方式中数据文件严重依赖于应用程序的缺点，在一定范围内实现了数据文件共享，这种共享文件方式已经具有了数据库的某些概念。

由此，我门可知，后期的文件系统有了很多改进，但一些根本性的问题仍然没有解决。主要表现在以下三方面：

①冗余度大

文件系统下的用户各自建立自己的文件，数据不能共享，造成大量重复，不仅浪费空间，而且增加更新开支，更严重的是容易产生数据不一致性。

②缺乏数据独立性

数据和程序互相依赖，一旦数据结构改变，与这些数据有关的程序都必须重新编写和调试，无形中增加了开支。

③数据无集管理

由于各个文件没有统一的管理机构，其安全性、完整性等也就无法得到保障。

所有这些问题，文件系统本身已无法解决，因而造成效率低、成本高，还有很多潜在问题，严重的阻碍了数据处理技术的发展。这就是数据库技术产生的原动力，也就是数据库系统产生的背景。

(3)数据库方式(70 年代→现在)

数据库克服文件系统的弊病，解决冗余和数据独立性问题。试图提供一种更完美的高级的数据管理方式。它用一个软件系统对数据实行统一的、集中的、独立的管理，使数据的存贮独立使用数据的程序，从而实现了数据共享，对于完整性、安全性等问题也都得到了相应的解决。

关于什么是数据库，它的特点和文件系统的区别，我们将在下节中介绍。

§1.2 数据库是什么

一、数据库是什么

数据库系统是一种有组织地、动态地存储有密切联系的数据集合，并对其进行统一管理的记算机软件以及硬件资源所组织的系统。数据库系统将各有关部门中反映客观事物的大量信息，进行记录、分类、整理、定量化、规范化处理，并以记录为单位存储于数据库中，数据库实质上是一个记录保存系统。一个为多种应用提供共享数据的数据集合。许多传统的纸上文件可以很方便地保存在数据库里。在数据库系统的统一作用下，用户通过应用程序向数据库发出查询、检索等操作命令以得到各层次用户所需要的信息。例如，在我们的计算机上有一个（相当小的）数据库，它帮助我们记录学生的人事档案。它是用 DBASE Ⅲ 数据库建立的。我们利用 DBASE Ⅲ 数据库系统对这档案数据进行下面的四种操作，以帮助我们了解数据库。

①检索

如果我们想了解所有学生的情况，就输入下面的命令：

. USE STUDENT ↙

. DISPLAY ALL ↙

计算机会立即显示：

Record#	学号	姓名	性别	籍贯	出生日期	班名	入学平均分	助学金	学历
1	86201	张大山	男	山东	10/01/71	林业 89	80.00	25.00	memo
2	86202	刘萍	女	山东	08/09/70	林业 89	90.00	29.00	memo
3	86203	李红	女	四川	01/20/72	林业 89	80.00	31.00	memo
4	86204	刘勇	男	湖南	11/23/72	林业 89	80.20	32.00	memo
5	86205	周峰	男	四川	05/08/72	林业 89	81.10	25.00	memo
6	86207	赵明	男	吉林	07/20/71	林业 89	79.00	31.00	memo
7	86301	王晓	女	山西	06/21/73	信息 88	88.00	30.00	memo
8	86302	周新	男	河北	09/11/72	信息 88	80.00	28.00	memo
9	86303	许雷	男	四川	03/04/72	信息 88	75.50	30.00	memo

如果我们想知道籍贯是四川、男生、入学平均分成绩高于 80 分的同学情况，就输入下面的命令：

. DISPLAY FOR 籍贯="四川". AND. 性别="男". AND. 入学平均分>80 ↙

计算机显示：

Record#	学号	姓名	性别	籍贯	出生日期	班名	入学平均分	助学金	学历
5	86205	周峰	男	四川	05/08/72	林业 89	81.10	25.00	memo

②删除

如果某个学生被除名，那么在这个数据库中不必再保留该学生的情况，用下面的命令删除：

. DELETE FOR 姓名="刘勇" ↙

刘勇这个人的数据就被删掉了。

③更新：

如果对所有学生的助学金增加5元，则输入下面的命令：

. REPLACE ALL 助学金 WITH 助学金+5 ↙

④插入：

如果新增加一个学生，则输入下面命令：

. APPEND ↙

然后逐项输入该学生的数据。

现在，我们已看到对数据的四种基本操作，初步了解数据库。

由于数据库发展很快，已成为大家所熟悉的术语，但由于发展时间还处于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发展和变化。另一方面，由于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的概括其全部特征。因此，对于什么是数据库目前还没有一个统一的公认的定义。

目前较流行的定义有三个：

(1)DBTG(Data Base Task Group)的定义：

数据库是由一个特定模式(schema)控制的所有记录、系(set)和域组成的。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彼此无关。

(2)C. J. Data 的定义：

存贮在磁介质或其它存贮介质上的数据集合一指数据库本身，存在以这种数据为背景运行的若干应用程序，对其进行检索、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；数据库是集成的，包含许多用户的数据，每个用户只享用其中的一小部分，且不同的用户使用的部份以各种方式重叠，也就是单独的数据片能够被许多不同的用户所共享。

(3)J. Martin 的定义：

数据库是存贮在一起的相关的数据集合，这些数据去掉了有害的或不必要的冗余，为多种应用服务，数据的存贮独立使用它的程序；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的方法进行；数据被结构化，为今后的应用研究提供基础。

综合上述三种定义，我们可知道一个数据库系统包含五部分内容：

〈1〉有一个结构化的相关数据集合。在这个数据集合中没有有害的或不必要的冗余，能为多种应用服务，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。

〈2〉有一组用户，有使用数据库中数据的请求，对数据库进行检索、插入、删除和修改等操作。

〈3〉数据库管理员，是负责整个系统管理的建立、维护、协调工作的专门人员。

〈4〉负责数据库管理和维护软件系统，称为数据库管理系统(DatabaseManagement System 简称 DBMS)，它对数据库中数据的各种操作提供了一种公用的方法。它接受并完成用户程序或终端命令提出的建立数据库和访问数据库的各种请求，维护、保护数据库中数据不受破坏。

〈5〉存储数据库和运行数据库管理系统的硬件资源。

为了更方便了解数据库，我们不妨把它与图书馆作一比较，见表 1.1

数据库	图书馆
数据	图书
外存	书库
用户	读者
用户标识	借书证
数据模型	书卡格式
数据管理系统	图书馆管理员
数据的物理组织方式	图书馆物理存放办法
用户对数据库的操作	读者对图书馆的访问
(使用计算机语言检索、插入、删除、修改)	(用普通语言借书、还书)

表 1.1

也就是说,数据库要完成类似图书馆的工作,正象图书馆是存储和借阅图书的部门,而数据库是存储数据并负责用户访问数据库的机构,我们要把数据库理解成一个系统。

二、数据库的主要特征

数据库技术之所以在短短十几年内如此快速的发展,受到计算机科学界的重视,成为引人注目的一门新兴学科,是因为它有其独特的特征。主要特征如下:

1. 实现数据共享。这是促成发展数据库技术的一个重要原因,也是技术先进的一个重要体现。数据共享体现在:当前的所有用户可以同时存取数据库中的数据,未来的多个新用户也可以同当前用户同时存取数据库中的数据。数据库留有同其它高级语言的接口,也就是可以用多种语言使用数据库。

2. 减少了数据存储的冗余度。冗余,浪费了存储的空间。为消除冗余,需浪费大量的机时,并可能给出不一致的信息。数据库是从整体观念来组织数据。数据不面向个别应用,而为共享,从而避免了不必要的冗余。从理论上讲,数据库没有冗余的,但实际上,许多数据库为了改善访问时间或为了较简单的寻址,还需存在着某种程度的冗余性。但它是去掉了有害的或不必要的冗余。

3. 维护了数据的一致性。数据的一致性是指数据的不同相容性或矛盾性。例如,对于同一个人的工资如果出现在财务科的工资单上的数字与其人事部门的档案中的不同时,就出现了不一致性。数据的不一致性主要是由于数据库冗余所引起的。数据库即使存在某些冗余,但由于数据库系统提出了对数据的各种控制和检查,以保证在更新数据时,同时更新所有付本,从而维护了数据的一致性。

4. 加强了对数据的保护,数据的安全可靠是一个数据库能否实用的关键问题,这也是用户非常关心的问题。数据保护有如下四方面的内容:

(1)安全性控制。主要是数据的保密性控制。采用方法有:将需要保密的部分与其它共用部分数据隔离开来,建立一些规则。如身份号、权限,或将数据以密码的形式存放于数据库内。

(2)完整性控制。完整性包括正确性、有效性和相容性。例如,一个数值型数据包含了诸如字母、特殊符号字符,则是错误的,是失去完成性的例子。

(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:某用户修改一个数据正准备写入库中,而另一用户却将该未修改的数据读出来使用,就会引起错误。并发控制功能将排除和避免这种错误的发生。

(4)故障的发现和恢复。数据库使用中，不可能保证其不受破坏，全部性和局部性的破坏随时可能发生。以及软件和硬件的故障。数据库采用下述措施来保障数据库的正确性：

①定期备份。

②对使用数据的过程进行登录。

③修改前的备份。

5. 维护数据的独立性。数据库的一个目的就是要使数据与使用它们的各个应用程序相互独立，数据与应用程序之间不存在依赖关系。独立性分二级：

(1)物理独立性。即数据的物理结构变化，如设备的更换，物理位置的更变，存取方法等等，不影响数据库的逻辑结构，从而不影响应用程序的修改。

(2)逻辑独立性。数据库逻辑结构的改变。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，不致影响到用户的原有应用程序的修改。遗憾的是逻辑独立性到目前为止还没能完全彻底地得到实现。

6. 对数据实行集中控制。由于文件管理方式，使数据库的管理处于一种分散的状态中。各个用户或同一用户的各个不同的处理文件之间通常是毫无关系的。而数据库对数据进行集中的控制和管理，对数据进行结构化。

§1.3 数据库系统在管理信息系统中的作用

数据库是管理信息系统（Management Information System，简称 MIS）的一部分，MIS 是进行信息的收集、转换、加工，利用信息进行预测和控制的系统，它是辅助企业或组织的领导进行决策的系统。

MIS 是以计算机为工具，数据库系统为基础的管理信息系统。如图 1. 5，图 1. 6。

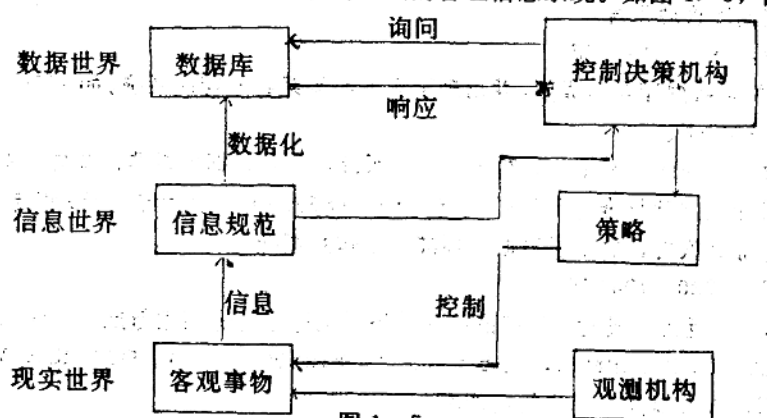

图 1. 5

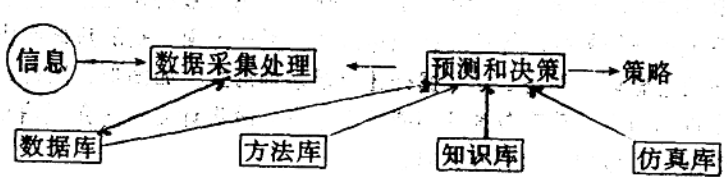

图 1. 6

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）。然后将规范后的信息数据化，并输入数据库中保存起来，其中一部分信息直接送入控制决策机构。控制决策结构向数据库提取数据作出决策，控制客观事物。对于一个 MIS 系统进行预测，决策除依靠数据库、还可依靠方法库、知识库、仿真库等。信息处理是其中重要的基础，而数据库系统是数据处理的核心。

§1.4 数据库技术的发展及前景

数据库自从 60 年代后期出现以来，已成为计算机科学技术中发展最快的重要部分之一，这十几年的发展，大至可分为以下几个阶段：

第一阶段：七十年代之前

这一阶段出现的是非关系型的数据库管理系统。如 IMS 公司研制成的世界上第一个数据库管理系统 IMS（层次）和美国 CODASYL（Confenceon Data System Languages）组织的 DBTG（Data Base Task Group）小组的 DBTG 报告。

第二阶段：1970—1974 年

这一阶段，美国 IBM 公司的研究员 E. F. Codd 提出了数据的关系模型，开创了数据库关系方式和数据规范化理论的研究，将数据方法引入到数据库研究中，使数据库技术有了理论基础。非关系型的数据库管理系统方面，IBM 公司对层次系统 IMS 作了许多改进，使之成为数据库层次方法的典型代表。对于网状系统，CODASYL 组织的 DBTG 小组又陆地发表了改进本，功能越来越强。由于 CODASYL 所作的工作，澄清了许多概念，建立了若干权威性的观点，极大地推动了数据库的发展，为数据库走向成熟奠定了基础。

此后，美国许多大学开始了数据库技术的研究。美国的很多公司也先后研制了各种各样大大小小的数据库管理系统。

第三阶段：1975—1979 年

此阶段，关系数据库走向成熟，并出现了完备的关系数据管理系统。如 1976 年宣布的 INGRES 和 SYSTEM—R。

这一阶段出现的新技术是分布数据库管理系统，以及数据库机器。此外，关系数据库理论的逐渐形成和数据库标准化问题的提出，ANSI/X3/SPARC 先后于 1975 年与 1977 年发表的中间报告与最终报告，奠定了数据库标准化工作的基础。

第四阶段：1980—1984 年

这一阶段中，巨大的成就是许多商品化关系数据库系统的问世与推广。如 SOL/DS，DB2，INGRES，ORACLE，INFORMIX，R：base400，R：base500，UNIFY，DB3，KNOWLEDGEMAN 等，广泛使用，大大提高了用户的生产率。

此阶段出现了两个最值得注意的动向，其一是人工智能数理逻辑与数据相结合，涌现出专家数据库系统，知识库系统和演绎数据库等这类智能化系统。另一方面，是随数据库应用于新的领域，出现了多介质信息管理，即格式化数据外还包括多种形式。如图形、图像、正文和声音等非格式化数据表现的信息。现有数据库不能满足需求，故设计支持多介质信息管理的数据库管理系统成为数据库领域中新的研究方向。

第五阶段：1985 年的以后

分布数据库普遍推广使用，新颖技术如知识库系统、多介质数据库等，已被广泛接受，成为当前热门的前沿课题。在各方面都取得研究成果，其发展大有方兴未艾之势。

我国数据库技术起步较晚，1975年后才开始接触。1977年11月，在国家计委支持下，由中国科技大学主办在黄山召开了第一次全国数据库学术会议，这次会议对在我国宣传和推广数据库技术起到了开创作用。通过十多年的艰苦努力，培养了一批研究人员，及时地引进新的技术，紧跟国际前沿。并开始研制分布数据库系统，考虑智能化的数据库系统，并实现了一些多介质数据库系统。使数据库在国民经济重大领域和国家经济信息系统方面已获得普遍应用，以DBASE Ⅲ，DBASE Ⅱ为代表的微机数据库应用系统有如雨后春笋。自行研究了一些微机数据库管理实验系统。数据库设计工具和环境的开发，数据库技术和人工智能技术的结合，大型知识库系统的研究已经起步，但是另一方面，我们基础比较薄弱，还没有自己的商品化的数据管理系统，对于新颖领域虽能及时进行探讨，但不能深入。故要进一步努力，竭尽全力缩短我国的数据库技术在国际上的差距。

目前数据库系统的研究主要集中在以下几个方面：

（1）探讨数据库设计的方法论。目前数据库设计仍然停留在经验与尝试阶段。工程规范程度不高，缺乏理论指导，设计的好坏在很大程度上取决于主要设计者的个人知识和实践经验，数据库设计方法论包括：

①数据模型。包括用户模型与概念模型，设计方法，解决从现实世界到数据库的逻辑描述问题。

②数据存取方法的设计。根据模型设计数据的物理存储结构，确定用户存取数据的方式。

③数据库的管理与保护。

（2）数据库规范理论的研究。研究数据的语言问题（即数据元素之间的关系），构造规范的数据模型，使存储数据能正确反映现实世界的联系，防止导出与客观实际矛盾的结果。

（3）实现数据库标准化，使系统更加统一和通用，使用户接口尽量简单。使同一系统既能与关系模型的用户接口，又能与网络模型的用户接口。另一方面，对现行数据库系统进行简化和统一，建立通用的，标准化的数据库，类似程序语言的标准化。为各种数据库提供统一的基础，建立共用数据库结构。

（4）数据库机器系统的研制，DBM（Data Base Machine）

由于现行数据库系统中，DBMS软件的运行需占用CPU的处理50%以上时间，使系统效率低，为此，人们研究用硬件去完成一部分数据库软件的工作，目前仍处于实验阶段。

（5）分布式数据库系统DDBS（Distributed Data Base System）它是数据库技术与计算机网络相结合的产物。DDBS实际上是一个逻辑的数据库，而它的物理数据库是分布在计算机网络的多个结点上的物理数据段所组成，当前有两种研究方向：①对计算机网络系统各个结点上已有的数据库，建立一个负责分布和协调的管理系统，使现在的集中式数据库系统向分布式数据库过渡。②在计算机网络上建立全新的，统一的分布式数据库管理系统。

（6）研究多介质数据库系统，知识库系统，将数据库技术与人工智能技术相结合。这是80年代开始的一股最引人瞩目的新潮流。许多新的探讨表明，人工智能研究正强烈地影响着新一代数据管理系统研制和开发。

（7）数据库设计工具和环境开发。

第二章　实体 信息 数据

数据库是一个统一的集中的数据管理机构，因此它必须能反映现实世界中各种复杂的数据关系。现实世界中的复杂关系是怎样转换成数据呢？要实现这种转换，我们需要做什么呢？

§2.1　三个世界理论

将我们现实世界中各种复杂关系输入到数据库中，要分三个阶段，称之为三个世界，即现实世界(Read world)、信息世界(Information world)和计算机世界(Computer world)。

现实世界 —加工→ 信息世界 —加工→ 计算机世界

三个世界中所用的术语和概念是不同的，以下分别给予介绍：

一、现实世界

现实世界中存在着各种各样千差万别的事物。所谓事物，就是能够相互区别开来的东西，如每棵树都是一个事物，因为它们都能彼此相区别的。当然，某些事物表面上看起来似乎无法加以区别，但是它们的区别是客观存在的，只要我们仔细观察研究，还是能发现区别的。

每个事物都有一些特征，我们正是利用这些特征将它们区分开来。比如，树的特征是：树种、树高、胸径、年龄等等。人的特征是：姓名、性别、年龄、籍贯、学历等等。对于一个事物，存在的特征很多。但我们一般选取那些我们感兴趣的特征，如工资管理中，往往只选取工资级别、工资金额等，对籍贯，政治面貌不感兴趣，而人事管理中，则对姓名、年龄、籍贯、政治面貌、社会关系等感兴趣。

世界上事物千千万万，关系千差万别，但它们之间都有联系。例如人、树、房三者间差别明显。但有联系。人植树、树造房、房子为人所用。

二、信息世界

为了将存在现实世界中客观现象最终在计算机数据库中反映出来，必须要对这些现象进行认真地分析，去粗取精，去伪存真，最后得到一些基本概念和基本关系。这些基本概念与基本关系术语如下。

1. 实体(Entity)

现实世界中的事物可以抽象成实体。实体是客观存在的并且能相互区别的。比如一个人、一匹马、一次演出都可以称之为实体。

2. 实体集(Entity set) 性质相同的同类实体的集合叫实体集。比如所有的男同学，所有的树，所有的书。

3. 属性(Attribute)

现实世界中的事物都有一些特性，这些特性我们用属性来表示。因此，属性刻化了实体的特性。一个实体往往可以用几个属性刻化。也就是一个实体可以用一个属性集表示。如人事档案管理中每个人可以用一组属性：人员代码、姓名、性别、年龄、籍贯等表示。每个属性都有值，如年龄可以是20岁，30岁，50岁等。并且每个属性的取值都有一定变化范围，称为属性域。

4. 实体标识符

能将一个实体与其它实体区别开来的属性叫实体标识符。

5. 联系

实体集间的联系。

三、计算机世界

现实世界中的客观现象最后在计算机中的表现是以数据形式来表现，因而计算机世界又可以叫数据世界。

计算机中的数据是按文件形式组织的。而文件是记录(Rcord)所组成。记录又可分为若干个字段(数据项)(Field)。

(1)字段(Field)

字段也称数据项，是数据结构中最小的单位。相应于信息世界中的属性。如学号、姓名、年龄、班级等均为学生的字段(数据项)。

(2)记录(Record)

记录由若干个字段组成，记录内的各字段间是有逻辑联系的。

(3)文件(File)

文件是记录的集合，一般讲，一个文件包括的记录都是同型的，每个文件有一个文件名。

(4)关键字(Key)

能唯一标识记录的一个或多个数据项的记录值称为文件的关键字。

上述各术语的对应关系如图2.1所示：

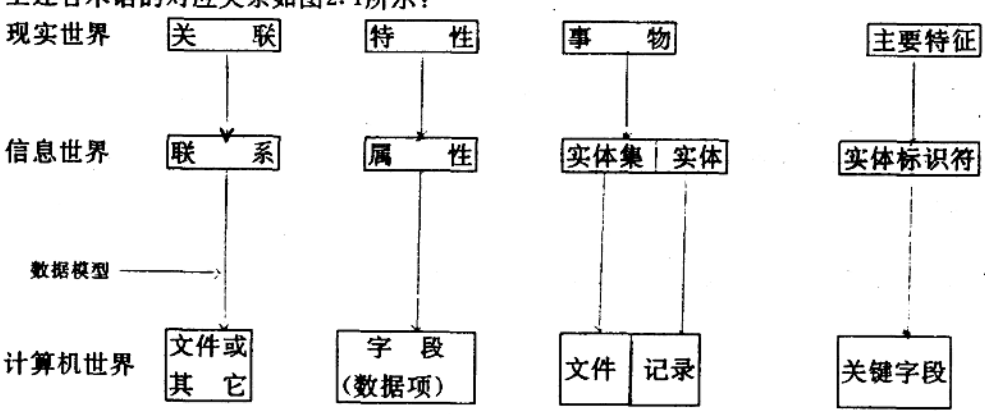

图2.1

数据项、记录、文件是数据世界中的基本数据存取结构。但是，如何将信息世界中的属性实体，联系转化成基本的数据存取结构呢？这里就需要建立数据模型。以数据模型为中间媒介建之起信息世界中的属性，实体联系到数据世界中的基本数据存取结构间的转换。目前一般的数据模型有三种，它们是网络模型，层次模型以及关系模型。这三种模型下章介绍。数据库管理系统实际上就是关系数据模型的描述。以数据操纵为其主要任务的软件，而其中数据操纵是以数据模型为转移的。因此，对数据模型的研究(及其操作)将成为数据管理系统的任务。

§2.2 实体间的联系

现实世界中的事物是彼此关联的，任何一个实体都不能孤立存在，描述实体的数据也是

互相联系的。联系有两种，一是实体内部的联系，反映在数据上是记录内部的字段间的联系，另一种是实体与实体间的联系，反映在数据上就是记录之间的联系。

文件系统只考虑记录内部的联系，而不考虑记录与记录这间的联系和文件与文件间的联系，因而从整体上看数据是无结构的。这就是文件系统存在各种弊病的根由。

数据库系统中除考虑记录内部的联系外，还必须考虑记录间的联系，文件与文件的联系。这种联系比较复杂，这就是数据库系统复杂的原因。这些都是实体间关系的复杂性引起的。

实体间的关系虽然复杂，但可以把它们归结为三类：

(1) 1—1 (one—to—one) 关系

定义：如果两个实体集 E1、E2中每一个实体至多和一个实体集的一个实体有联系，则 E1、E2 叫做“一对一”关系。例如，一个公司只有一个经理，同时一个经理只能在一个公司任职。

1—1

(2) 1—m (one—to—many) 关系

定义：有两个实体集 E1和 E2，如果 E2中每个实体与 E1中任意个实体（包括零个）有关，而 E1中每个实体至多和 E2中一个实体有关，则称该关系为“从 E2到 E1的1对多关系”。如母子关系、公司与职员关系。

1—m

(3) m—m (many—to—many) 关系

定义：如果两个实体集 E1、E2中的每一个实体都和另一个实体集中任意个实体（包括零个实体）有关，则称这两个实体集是“多对多关系”。例如师生关系，学生与选课关系。

注意：1—1关系是1—m 关系的特例，而1—m 关系又是 m—m 关系的特例。它们之间的关系是包含关系。如图 2. 2

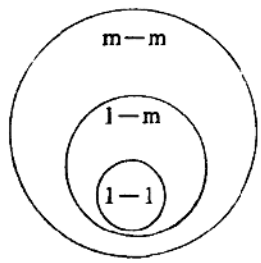

图2. 2

§2.3 实体—联系方法(ER 方法)

三个世界及转换过程中起关键作用的是信息世界。现实世界中是很难抽象成计算机世界的，必须通过中间阶段→信息。对信息世界需进行模式设计。目前较为流行的方法叫实体——联系方法 (Entity—relationship Approach)。

E—R 方法是由 P. P. S. chen1976年提出来的，用一种叫 E—R 图 (Entity—Relationship Diagram) 去描述信息模型。

这种方法直接列出所有的实体，实体属性以及实体间的联系，这种联系用一些抽象的命名表示如下：

(1) 实体集，用长方形表示，长方形内写上实体名。

(2) 属性，用椭圆形表示，属性名写在椭圆形内。

(3) 实体间联系，用菱形表示，在菱形框内写上联系名。

而在实体集之间的连线采用的是无向联线或有向联线.其中：

(1) 1—1关系用双向箭头分别指向有关实体。

(2) 1—m 关系用单向箭头指向 1 的实体。

(3) m—m 关系不用箭头。

示例：用 E—R 方法表示图书信息管理的数据模型.见图2. 3

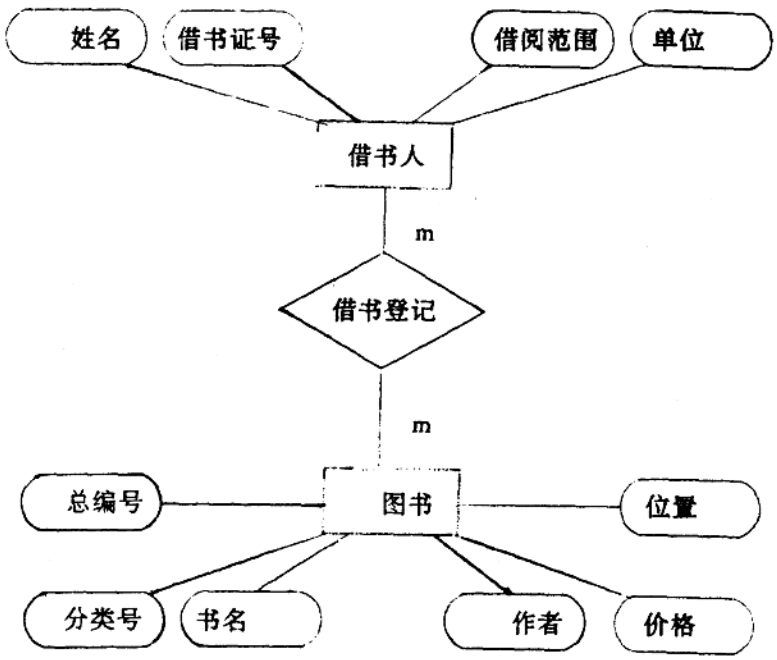

图2. 3

这种方法的一个关键是如何划分实体属性和实体间的关系。因为有些信息可用属性表示。有些信息可用实体关系表示。例如：部门和雇员这两个实体集是1—m 关系，这里有三种划分方法

(1) 雇员作为实体，部门作为它的属性。

(2) 雇员和部门都作为实体，再建两个实体间的联系。

(3) 部门作为实体，雇员作为它的属性。

采用哪一种方法，要具体情况具体分析。比如：对于工资管理，每个人只需一个部门名或部门号，则方法 (1) 较好；如果部门还有其它信息（所在地点，电话号码等等），则方法 (2) 较好。但无论何种情况，方法 (3) 总是不好的。(1) 和 (2) 总是可以的。结论：在 1—1 关系和 1—m 关系中表示 1 的实体如果只有一个信息需要表示，那么把它作为另一个实体的属性比较合适，而所有其它情况下都应作为两个实体来考虑，并用两实体间的关系表示所需信息。