第 1 章

数据库概述

1.1 数据库概述

数据库（DataBase）是按照数据结构来组织、存储和管理数据的仓库，它产生于 20 世纪 60 年代。随着信息技术和市场的发展，特别是 20 世纪 90 年代以后，数据库不再仅限于存储和管理数据，而转变成开发用户所需要的各种数据管理的方式。数据库有很多种类型，从最简单的存储各种数据的表格到能够进行海量数据存储的大型数据库系统，在各方面得到了广泛的应用。

数据库技术从诞生以来，在半个世纪的时间里，形成了坚实的理论基础、成熟的商业产品和广泛的应用领域，并吸引了越来越多的研究者。数据库的诞生和发展给计算机信息管理带来了一场巨大的革命，它已成为企业乃至个人日常工作、生产和生活的重要工具。同时，随着应用的扩展与深入，数据库的数量和规模越来越大，数据库的研究领域也已经大大地拓广和深化了。数据库领域获得了三次计算机图灵奖（C.W. Bachman，E.F.Codd，J.Gray），更加充分地说明了数据库是一个充满活力和创新精神的领域。

传统上，为了确保企业持续扩大的 IT 系统稳定运行，一般用户信息中心往往要不断更新容量更大的IT 运维软硬件设备，极大地浪费了企业资源；更要长期维持一支由数据库维护、服务器维护、机房值班等各种人员组成的“运维大军”，维护成本也随之节节高升。为此，企业 IT 决策者开始思考：能不能像拧水龙头一样按需调节使用IT 运维服务？而不是不断增加已经价格不菲的运维成本。

1.2 数据库发展史

数据库的历史可以追溯到 50 多年前的 20 世纪 60 年代，那时的数据管理非常简单。通过大量分类、比较和表格绘制的机器运行数百万穿孔卡片来进行数据的处理，其运行结果将在纸上打印出来或者制成新的穿孔卡片。而数据管理就是对所有这些穿孔卡片进行物理存储和处理。1951 年雷明顿兰德公司（Remington Rand Inc.）的一种叫做 Univac I 的计算机推出了一种一秒可以输入数百条记录的磁带驱动器，从而引发了数据管理的革命。1956 年 IBM 生产出第一个磁盘驱动器 the Model 305 RAMAC。此驱动器有 50 个盘片，每个盘片直径为 2 英尺，可以储存 5MB 的数据。使用磁盘最大的好处是可以随机地存取数据，而穿孔卡片和磁带只能顺序存取数据。1951 年 Univac 系统使用磁带和穿孔卡片存储数据。

数据库发展阶段大致可分为如下几个阶段：人工管理阶段、文件系统阶段、数据库系统阶段、高级数据库阶段。

1.2.1 人工管理阶段

20 世纪 50 年代中期之前，计算机的软、硬件均不完善。硬件存储设备只有磁带、卡片和纸带，软件方面还没有操作系统。当时的计算机主要用于科学计算。这个阶段由于还没有软件系统对数据进行管理，程序员在程序中不仅要规定数据的逻辑结构，还要设计其物理结构，包括存储结构、存取方法、输入/输出方式等。当数据的物理组织或存储设备改变时，用户程序就必须重新编制。由于数据的组织面向应用，不同计算程序之间不能共享数据，使得不同的应用之间存在大量重复数据，很难维护应用程序之间数据的一致性，如图 1-1 所示。这一阶段的主要特征可归纳如下：

① 计算机中没有支持数据管理的软件；

② 数据组织面向应用，数据不能共享，数据重复；

③ 在程序中要规定数据的逻辑结构和物理结构，数据与程序不独立；

④ 数据处理方式为批处理。

例如，Acrobat 生成的数据与 Word 的数据无法共享，如图 1-2 所示。

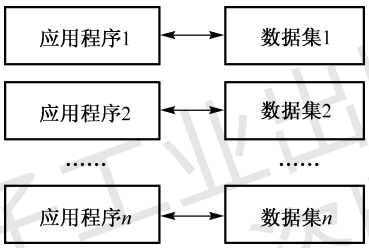

图 1-1 人工管理阶段

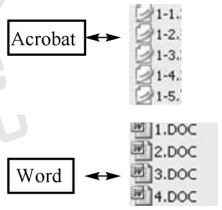

图 1-2 程序间不能共享数据

1.2.2 文件系统阶段

文件系统阶段的主要标志是计算机中有了专门管理数据库的软件——操作系统。20 世纪 50 年代中期到 60 年代中期，由于计算机大容量存储设备（如硬盘）的出现，推动了软件技术的发展，而操作系统的出现标志着数据管理步入一个新的阶段。在文件系统阶段，数据以文件为单位存储在外存，且由操作系统统一管理。操作系统为用户使用文件提供了友好的界面。文件的逻辑结构与物理结构脱钩，程序与数据分离，使数据与程序有了一定的独立性。用户的程序与数据可分别存放在外存储器上，各应用程序可以共享一组数据，实现了以文件为单位的数据共享，文件系统结构图如图 1-3 所示。

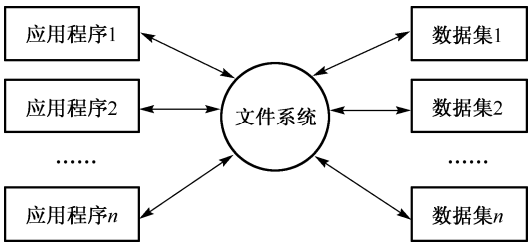

图 1-3 文件系统阶段

在文件系统中，一个 txt 文档既可以由 Word 软件打开，也可以由记事本软件打开，两种应用程序共享一个 txt 文档，如图 1-4 所示。由于文件系统中数据的组织仍然是面向程序的，存在大量数据冗余，数据的逻辑结构不能方便地修改和扩充，数据逻辑结构的每一点微小改变都会影响到应用程序。

图 1-4　文件系统中不同应用程序共享数据

另一方面，文件之间互相独立，不能反映现实世界中事物之间的联系，操作系统不维护文件之间的联系信息。如果文件之间有内容上的联系，也只能由应用程序处理。例如，在 Windows 操作系统中的一个 Word 文档，其中的文字数据不能够被另外一个 Word 文档直接使用，如图 1-5 所示。如果两个 Word 文档有相同的数据，则必须各自复制一份，这种重复就是冗余，不方便维护。当图 1-5 中任何一个学生更换电话号码时，必须同时修改两个文档。如果修改有遗漏，则号码不一致，其中必然有一个号码是错误的。

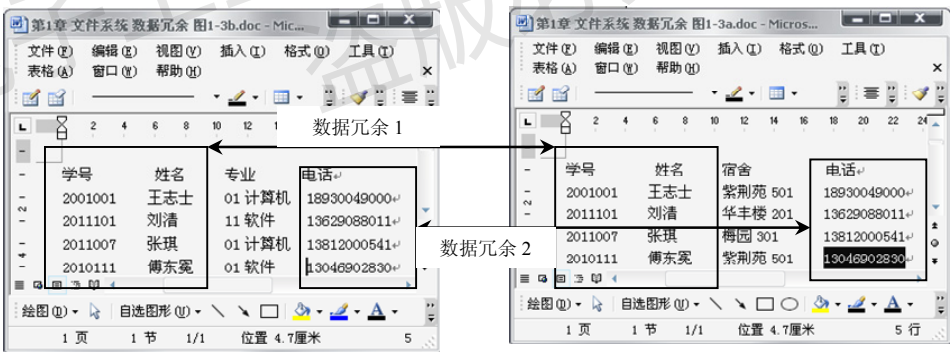

图 1-5　两个文档不能互相引用数据产生冗余

1.2.3　数据库系统阶段

20 世纪 60 年代后，随着计算机在数据管理领域的普遍应用，人们对数据管理技术提出了更高的要求：希望面向企业或部门，以数据为中心组织数据，减少数据的冗余，提供更高的数据共享能力，同时要求程序和数据具有较高的独立性，当数据的逻辑结构改变时，不涉及数据的物理结构，也不影响应用程序，以降低应用程序研制与维护的费用。数据库技术正是在这样一个应用需求的基础上发展起来的。

数据库技术有如下特点：

① 面向企业或部门，以数据为中心组织数据，形成综合性的数据库，为各应用共享。

② 采用一定的数据模型。数据模型不仅要描述数据本身的特点，而且要描述数据之间的联系。

③ 数据冗余小，易修改，易扩充。不同应用程序根据处理要求，从数据库中获取需要的数据，这样就减少了数据的重复存储，也便于增加新的数据结构，便于维护数据的一致性。

④ 程序和数据有较高的独立性。

⑤ 具有良好的用户接口，用户可方便地开发和使用数据库。

⑥ 对数据进行统一管理和控制，提供了数据的安全性、完整性及并发控制。

从文件系统发展到数据库系统，这在信息领域中具有里程碑的意义。在文件系统阶段，人们在信息处理中关注的中心问题是系统功能的设计，因此程序设计占主导地位；在数据库方式下，数据开始占据了中心位置，数据的结构设计成为信息系统首先关心的问题，而应用程序则以既定的结构为基础进行设计。本阶段文件系统结构图如 1-6 所示。例如，在某大学的网络服务软件中“教师管理系统”、“学生管理系统”、“教务管理系统”通过 SQL Server 管理系统共享教务数据库中的“学生表”、“课程表”、“教师表”、“成绩表”，如图 1-7 所示。

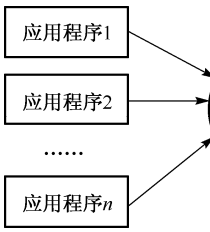

图 1-6 数据库系统阶段

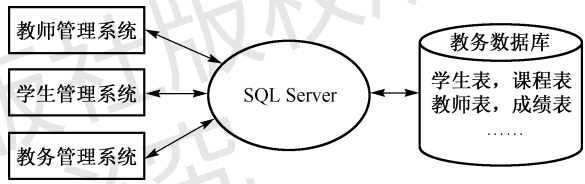

图 1-7 多个应用程序共享数据库

1.3 数据库系统的功能

数据库技术是计算机科学的重要分支。最初的数据管理采用的是人工管理方式，数据的存储结构、存取方法、输入/输出方式都要程序员亲自动手设计，数据管理的效率很低。随着大容量外存储器的出现，专门用于管理数据的软件“文件系统”应运而生，数据可以长期保存，程序员也不必过多地考虑物理细节，数据管理效率有所提高，但仍然不能共享数据，导致数据大量冗余。为了解决这个问题，20 世纪 60 年代中期出现了数据库技术，在数据库中可以实现应用程序间的数据共享，并最大限度地减少冗余，保证数据的正确性。由于数据库具有数据结构化好、冗余度小、数据独立性高、数据共享性高和易于扩充等优点，所以被广泛应用于数据处理中。

数据库是信息时代的产物，可实现大量信息的管理和处理。人们通过数据库可以方便地使用、查找所需要的信息。一个完整的数据库系统（DataBase System，DBS）由数据库（DateBase，DB）、数据库管理系统（DataBase Management System，DBMS）、数据库应用系统（DataBase Administrator System，DBAS）、数据库管理员（DataBase Administrator，DBA）及用户（User）组成。图 1-8 所示为数据库系统的组成，图 1-9 所示为数据库的角色访问层次。

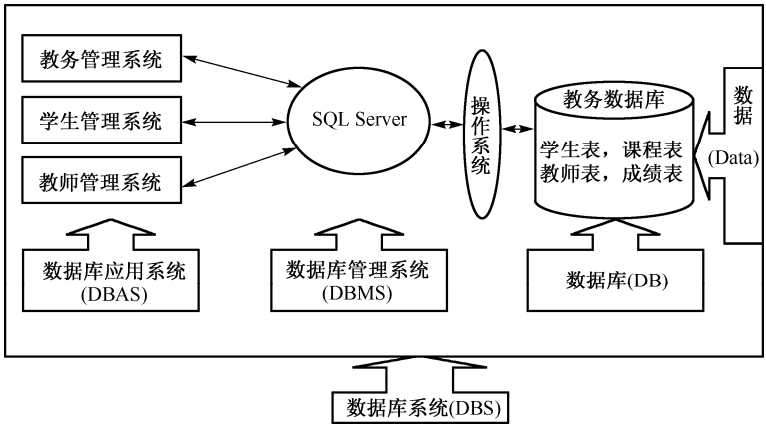

图 1-8　数据系统的组成

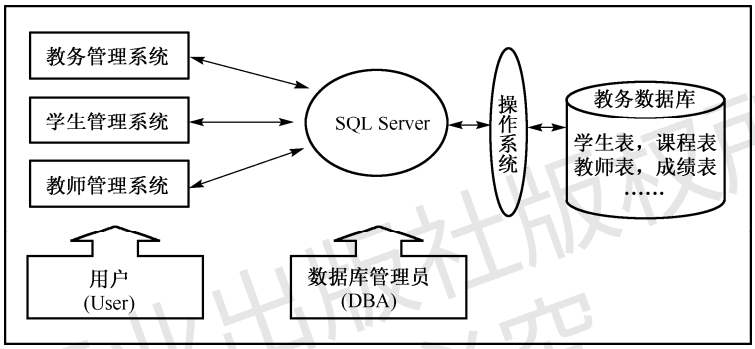

图 1-9　数据库的角色

在介绍数据库之前首先需要理解关于数据库的几个概念。

（1）数据

数据（Data）是信息的符号化表示，是记录事务的物理符号。数据的表示形式是多种多样的，可以是数值的、字符的、图形的、声音的等。为了了解世界、交流信息，人们需要描述这些事物。在日常生活中直接用自然语言（如汉语）描述。在计算机中，为了存储和处理这些事物，就要抽出这些事物的特征组成一个记录来描述。

例如，在学生档案中，如果人们最感兴趣的是学生的姓名、性别、年龄、出生年月、籍贯、所在系别、入学时间，那么可以这样描述（刘清，女，21，1990，福建，计算机系，2011），这里的学生记录就是数据。对于上面这条学生记录，了解其含义的人会得到如下信息：刘清是个大学生，1990 年出生，女，福建人，2011 年考入计算机系；而不了解其语义的人则无法理解其含义。可见，数据的形式还不能完全表达其内容，需要经过解释。所以数据和关于数据的解释是不可分的，数据的解释是指对数据含义的说明，数据的含义称为数据的语义，数据与其语义是不可分的。

（2）数据库

所谓数据库（DataBase，DB）就是长期存放在计算机内，以一定组织方式动态存储的、相互关联的、可共享的数据集合。数据库中的数据结构化好、冗余度小、独立性高、共享性高并易于扩充。数据库存储数据，是一个静态的存储结构。数据库中的数据是存放在外存储器中的永久性数据，使用时必须把它调入内存。

（3）数据库管理系统

数据库管理系统（DataBase Manage System，DBMS）是一个专门的管理软件，负责数据的检索、增加、删除与修改，维护数据的一致性与完整性，提供正确使用的各种机制。应用程序不能直接使用数据库中的数据，只能提出访问数据的请求，由 DBMS 完成对数据的操作。数据库管理系统是指建立在操作系统之上，支持数据库的建立、使用和维护的软件，如 Microsoft SQL Server 和 Oracle 等。它们建立在操作系统的基础上，对数据库进行统一管理和控制。利用数据库管理系统提供的一系列命令，用户可以建立各种数据库操作文件和辅助文件，定义数据及对数据进行增加、删除、更新、查找、输出等操作。用户对数据的操作要通过数据库管理系统实现。此外，数据库管理系统还承担着数据库维护的任务。

（4）数据库应用系统

数据库应用系统（DataBase Application System，DBAS）是指用 Visual Basic、FoxPro 等开发工具设计的、实现某种特定功能的应用程序，如学生成绩管理系统、工资管理系统、物资管理系统等。它利用数据库管理系统提供的各种手段访问一个或多个数据库，实现其特定的功能。

（5）数据库系统

数据库系统（DataBase System，DBS），是指由计算机硬件、操作系统、数据库管理系统，以及在其他对象支持下建立起来的数据库、数据库应用程序，用户和维护人员等组成的一个整体。

1.4 数据库系统的三级模式结构

从数据库管理系统角度看，数据库系统通常采用三级模式结构。从数据库最终用户角度看，数据库系统的结构分为集中式结构、分布式结构、客户/服务器结构和并行结构。

数据库系统的三级模式结构是指数据库系统是由外模式、模式和内模式三级构成的，如图 1-10 所示。用户级对应外模式，概念级对应模式（概念模式和逻辑模式），物理级对应内模式。在一个数据库系统中，只有唯一的数据库，因而作为定义、描述数据库存储结构的内模式和定义、描述数据库逻辑结构的模式，也是唯一的，但建立在数据库系统之上的应用则是非常广泛、多样的，所以对应的外模式不是唯一的，也不可能是唯一的。

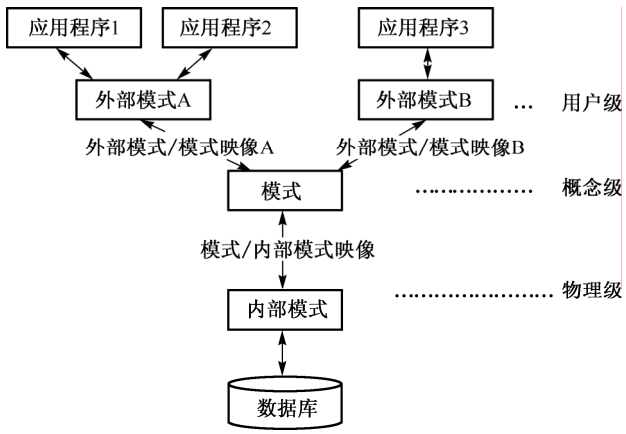

图 1-10　数据库系统的三级模式结构

（1）模式

模式又称概念模式或逻辑模式，对应于概念级。它是由数据库设计者综合所有用户的数据，按照统一的观点构造的全局逻辑结构，是对数据库中全部数据的逻辑结构和特征的总体描述，是所有用户的公共数据视图（全局视图）。它是由数据库管理系统提供的数据模式描述语言（Data Description Language，DDL）来描述、定义的，反映了数据库系统的整体观。

（2）外模式

外模式又称子模式，对应于用户级。它是某个或某几个用户所看到的数据库的数据视图，是与某一应用有关的数据的逻辑表示。外模式是从模式导出的一个子集，包含模式中允许特定用户使用的那部分数据。用户可以通过外模式描述语言来描述、定义对应于用户的数据记录（外模式），也可以利用数据操纵语言（Data Manipulation Language，DML）对这些数据记录进行操纵。外模式反映了数据库的用户观。

（3）内模式

内模式又称存储模式，对应于物理级，是数据库中全体数据的内部表示或底层描述，是数据库最低一级的逻辑描述，它描述了数据在存储介质上的存储方式和物理结构，对应着实际存储在外存储介质上的数据库。内模式由内模式描述语言来描述、定义，反映了数据库的存储观。

（4）三级模式间的映射

数据库的三级模式是数据库在 3 个级别（层次）上的抽象，使用户能够逻辑地、抽象地处理数据而不必关心数据在计算机中的物理表示和存储。实际上，对于一个数据库系统而言，已有的物理级数据库是客观存在的，它是进行数据库操作的基础（内模式），概念级数据库中不过是物理数据库的一种逻辑的、抽象的描述（即模式），用户级数据库则是用户与数据库的接口，它是概念级数据库的一个子集（外模式）。

不同级别的用户对数据库形成不同的视图。所谓视图，就是指观察、认识和理解数据的范围、角度和方法，是数据库在用户“眼中”的反映。很显然，不同层次（级别）的用户“看到”的数据库是不同的，图 1-11 所示为一个三级模式映射的例子。用户应用程序“教务管理系统”根据外模式“学生视图”和“教师视图”进行数据操作，通过“外模式/模式映射”定义和建立某个外模式与模式间的对应关系，将外模式“学生视图”“教师视图”与模式“学生表”“教师表”联系起来，当模式发生改变时，只要改变其映射，就可以使外模式保持不变，对应的应用程序也可以保持不变。另一方面，可通过“模式/内模式映射”定义建立数据的逻辑结构（模式）“学生表”“成绩表”与存储结构（内模式）“学生表”“成绩表”间的对应关系。当数据的存储结构发生变化时，只需改变“模式/内模式映射”，就能保持模式不变，因此应用程序也可以保持不变。

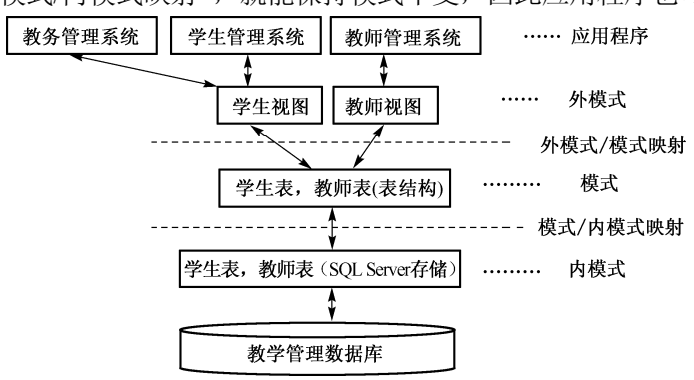

图 1-11　教务系统的三级模式结构

例如，图 1-12 所示的教务系统的三级模式实例是图 1-11 的一个例子。其中，图 1-12(c)“学生文件”、“课程文件”、“选课文件”是显示内模式表示数据的底层结构，即物理结构；图 1-12(b)中显示的模式“学生关系”、“选课关系”、“课程关系”只与逻辑结构有关，与物理结构无关。通过模式与内模式的映射，可以保证逻辑结构与底层物理存储具有相对独立性，也就是说，当物理存储改变时（例如存储位置发生改变），不需要修改逻辑结构，只要修改映射，就可以保证数据库能够运行。图 1-12(a)外模式“成绩单”显示与用户交互的部分，也就是说，可以通过外模式隐藏不希望用户看到的信息，同时代码通过外模式访问数据，可以使得代码具有可移植性。也就是说，代码可以脱离数据库的逻辑结构，当数据库逻辑结构发生改变时，只要修改模式与内模式的映射，不需要修改代码即可运行。

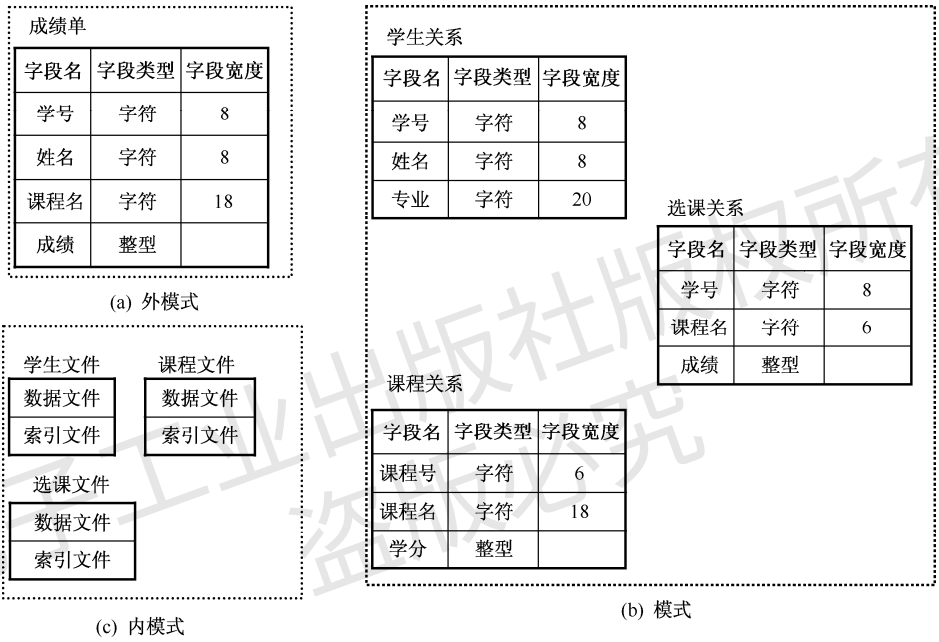

图 1-12 教务系统的三级模式实例

扩展阅读

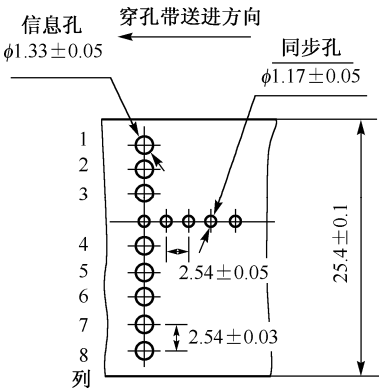

图 1-13 穿孔卡

穿孔卡是早期计算机的信息输入设备，通常可以存储 80 列数据。它是一种很薄的纸片， 面积为 190×84 mm^2，见图 1-13。首次使用穿孔卡技术的数据处理机器，是美国统计专家赫曼·霍列瑞斯（H.Hollerith）博士的伟大发明。

公元 1880 年，美利坚合众国举行了一次全国性人口普查，为当时 5000 余万的美国人口登记造册。当时美国经济正处于迅速发展的阶段，人口流动十分频繁，再加上普查的项目繁多，统计手段落后，从当年元月开始的这次普查，花了 7 年半的时间才把数据处理完毕。也就是说，直到快进行第二次人口普查时，美国政府才能得知第一次人口普查期间全国人口的状况。

人口普查需要处理大量数据，如用调查表采集的项目年龄、性别等，并且还要统计出每个社区有多少儿童和老人，有多少男性公民和女性公民等。这些数据是否也可以由机器自动进行统计？采矿工程师霍列瑞斯想到了纺织工程师杰卡德 80 年前发明的穿孔纸带。杰卡德提花机用穿孔纸带上的小孔控制提花操作的步骤，即编写程序。霍列瑞斯则进一步设想要用它来存储和统计数据，于是他想发明一种自动制表的机器。两年后，霍列瑞斯博士离开了人口局，到专利事务所工作了一段时间，也曾任教于麻省理工学院，他一边工作，一边致力于自动制表机的研制。

霍列瑞斯首先把穿孔纸带改造成穿孔卡片，以适应人口数据采集的需要。由于每个人的调查数据有若干不同的项目，如性别、籍贯、年龄等。霍列瑞斯把每个人所有的调查项目依次排列于一张卡片上，然后根据调查结果在相应项目的位置上打孔。例如，穿孔卡片“性别”栏目下有“男”和“女”两个选项，“年龄”栏目下有从“0 岁”到“70 岁以上”等系列选项，等等。统计员可以根据每个调查对象的具体情况，分别在穿孔卡片各栏目的相应位置打出小孔。每张卡片都代表着一位公民的个人档案。

霍列瑞斯博士巧妙的设计在于自动统计。他在机器上安装了一组盛满水银的小杯，穿好孔的卡片就放置在这些水银杯上。卡片上方有几排精心调好的探针，探针连接在电路的一端，水银杯则连接于电路的另一端。与杰卡德提花机穿孔纸带的原理类似：只要某根探针撞到卡片上有孔的位置，便会自动跌落下去，与水银接触从而接通电流，启动计数装置前进一个刻度。由此可见，霍列瑞斯穿孔卡表示的也是二进制信息：有孔处能接通电路计数，代表该调查项目为“有”（“1”），无孔处不能接通电路计数，表示该调查项目为“无”（“0”）。

直到 1888 年，霍列瑞斯博士才完成了自动制表机的设计并申报了专利。他发明的这种机电式计数装置，比传统纯机械装置更加灵敏，因而被 1890 年后的历次美国人口普查选用，获得了巨大的成功。例如，1900 年进行的人口普查全部采用霍列瑞斯制表机，平均每台机器可代替 500 人工作，全国的数据统计仅用了 1 年多时间。虽然霍列瑞斯发明的并不是通用计算机，除了能统计数据表格外，它几乎没有别的什么用途，然而制表机穿孔卡第一次把数据转变成二进制信息。在以后的计算机系统里，利用穿孔卡片输入数据的方法一直沿用到 20 世纪 70 年代，数据处理也发展成为计算机的主要功能之一。

依托自己发明的制表机，霍列瑞斯博士创办了一家专业制表机公司，但不久就因资金周转不灵陷入困境，被另一家 CTR 公司兼并。1924 年，CTR 公司更名为“国际商业机器公司”，英文缩写为 IBM，专门生产打孔机、制表机一类的产品。到了 1950 年，IBM 的卡片已被业界与政府机构广泛使用，为了让卡片可作为证明文件重复使用，卡片上都印有“请勿折叠、卷曲或毁损”的警示词，这行警示词后来还成为后二次大战时期的流行标语。

FORTRAN 程序穿孔卡的使用直到 20 世纪 70 年代为止，不少计算机设备仍以卡片作为处理媒介，世界各地都有科学系或工程系的大学生拿着大叠卡片到当地的计算机中心交作业程序，一张卡片代表一行程序，然后耐心排队等着自己的程序被计算机中心的大型计算机处理、编译并执行。一旦执行完毕，就会打印出附有身份识别的报表，放在计算机中心外的文件盘里。如果最后打印出一大串程序语法错误的信息，学生就得修改后重新再一次执行程序。穿孔卡直到今日仍未绝迹，其特殊的尺寸（80 行的长度）在世界各地仍使用在各式表格、记录和程序中。

杰卡德和霍列瑞斯分别开创了程序设计和数据处理之先河。以历史的目光审视他们的发明，正是这种程序设计和数据处理，构成了计算机软件的雏形。

习题

1．数据库系统的发展过程分成哪几个阶段？

2．数据管理技术发展过程中，文件系统与数据库系统的重要区别是什么？

3．数据库系统的功能有那些？

4．数据库系统的模式结构是怎样的？有什么样的用途？

5．通常所说的数据库系统（DBS）、数据库管理系统（DBMS）和数据库（DB）三者之间的关系是什么？