第 1 章 概 论

1.1 数据、信息与数据处理

人类的一切活动都离不开数据，离不开信息。随着科学技术的发展、生产技术的进步、商业和社会活动的复杂化，各行各业每时每刻都在产生大量的信息。

在计算机应用中，数据处理和以数据处理为基础的信息系统所占的比重最大，现代化水平越高，科学管理、自动化服务的需求就越大。

1.1.1 数据的概念

描述事物特性必须借助一定的符号，这些符号就是数据形式。例如，某人的出生日期是“二〇〇七年十月二十三日”，当然也可以将以上汉字形式改为用“10/23/2007”来表示。

所谓数据，通常指用符号记录下来的可加以鉴别的信息。数据的概念包括两个方面：其一，数据内容是事物特性的反映或描述；其二，数据是符号的集合。

“符号”不仅指数字、字母、文字和其他特殊字符，而且还包括图形、图像、声音等多种媒体数据；所谓“记录下来”也不仅是指印刷在纸上，还包括记录在存储介质中。

数据在空间上的传递称为通信，在时间上的传递称为存储。

1.1.2 信息的概念

信息是关于现实世界事物的存在方式或运动形态的综合反映，是人们进行各种活动所需要的知识。在不同的领域中，信息的含义有所不同。一般认为信息(information)是数据、消息中所包含的意义，是经过加工的数据。

数据与信息既有联系又有区别。数据是承载信息的物理符号或载体。数据能表示信息，但并非任何数据都能表示信息，正如人们常说的“如果计算机输入的是垃圾，输出的也会是垃圾”。同一数据也可能有不同的解释。信息是人们消化理解了的数据。信息是抽象的，不随数据设备所决定的数据形式而改变；而数据的表示方式却具有可选择性。数据和信息有时可以混用，例如，数据处理也称为信息处理；有时必须分清，例如，不能把信息系统称为数据系统。

信息是反映客观现实世界的知识,用不同的数据形式可以表示同样的信息。例如,同样一条新闻在报纸上以文字的形式刊登,在电台以声音的形式广播,在电视上以视频的形式放映,以及在计算机网络上以通信形式传播,其信息内容可以相同。

信息与数据的关系可以归纳为:

- 信息是有一定含义的数据。
- 信息是经过加工(处理)后的数据。
- 信息是对决策有价值的数据。

信息具有以下一些基本属性。

(1) 事实性。事实是信息的基本性质,也是信息的中心价值。因为不符合事实的信息不仅没有价值而且可能导致负价值,害人害己。因此,事实性是信息收集时最应注意的性质。

(2) 等级性。不同的使用目的要求不同等级的信息,例如有战略信息、策略信息、执行信息等。对于不同等级的信息,其保密程度、生命长短、使用频率、精度要求等都有不同。

(3) 可压缩性。可以对信息做浓缩处理,即进行集中、综合和概括而又不丢失信息的本义。例如,可以把大量实验数据总结成一个经验公式、剔除无用信息、减少冗余信息等。

(4) 可扩散性。信息可以通过各种渠道和手段向四面八方扩散,尤其是在计算机技术与通信技术飞速发展的今天,信息的可扩散性得到更加充分的体现。信息的可扩散性存在两面性,它有利于知识的传播,但又会造成信息的贬值以致产生无法弥补的利益损失。因此,人们采取了许多办法防止和制约信息的非法扩散,如制定有关法律、研究各种保密技术。

(5) 可传输性。信息可以通过多种形式迅速传输,如电话、计算机网络系统、书报、杂志、存储介质等。信息的可传输性优于物质和能源,它加快了资源传递,加速了社会的发展。

(6) 共享性。信息可以被多个用户共享而得到充分的利用。当然,共享信息时应该采取合法手段。

(7) 增值性与再生性。信息是有价值的,而且可以增值。信息的增值往往是信息从量变到质变的结果,是在积累的基础上可能产生的飞跃。信息再生还可能在“信息废品”中提炼有用的信息。

(8) 转换性。信息、物质和能源是人类的三项重要的宝贵资源,三位一体而又可以互相转换。现在很多企业利用信息技术大大节约了能源或获得合理的原材料,信息转换的目的是实现其价值。

1.1.3 数据处理

数据处理是指将数据转换成信息的过程。广义地讲,它包括对数据的收集、存储、加工、分类、检索、传播等一系列活动。狭义地讲,它是指对所输入的数据进行加工整理。其基本目的是从大量的、已知的数据出发,根据事物之间的固有联系和运动规律,通过分析归纳、演绎推导等手段,萃取出对人们有价值、有意义的信息,作为决策的依据。由此可见,信息是一种被加工成特定形式的数据,这种数据形式对于数据接收者来说是有意义的。对数据的加工可以相对比较简单,也可以相当复杂。简单加工包括组织、编码、分类、排序等;复杂加工

可以复杂到使用统计学方法、数学模型等对数据进行深层次的加工。

数据是原料，是输入；而信息是产出，是输出结果。当两个或两个以上数据处理过程前后相继时，前一过程称为预处理。预处理的输出作为二次数据，成为后面处理过程的输入，此时信息和数据的概念就产生了交叉，表现出相对性。如图 1-1 所示，人们有时说“信息处理”，其真正含义应该是为了产生信息而处理数据。

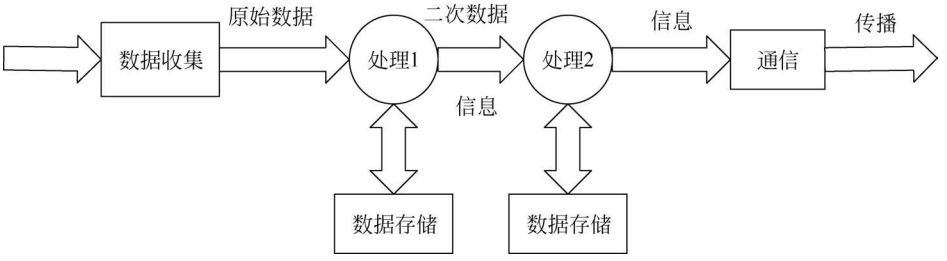

图 1-1 信息处理

例如，一个人的“出生日期”是有生以来不可改变的基本特征之一，属于原始数据，而“年龄”是当前年份与出生年份相减而得到的数字，具有相对性，可视为二次数据。同样道理，生产日期和购置日期是产品和设备的原始数据，失效日期和资产折旧是经过简单计算得出的结果。

又如，用手工或计算机填写的发货单，对于发货部门的工作人员来说即为照单发货的信息，但对于仓储部门的管理者来说，它只是核算、盘点库存量的原始数据。由于数据与信息之间存在着这种关系，因此这两个词有时被交替使用，其根本区别在于信息对当前或将来的行动或决策有价值。

1.2 计算机系统

计算机可以模拟人的大脑解决问题的思维过程和部分功能，因此它又被称为电脑。它的结构特点与人脑也有许多相似之处，应该具有接收（输入信息）、记忆（存储信息）、分析和处理（各种运算和判断）、按正确顺序逐步去做（控制）和得出结果（输出）这五部分功能以及实现这五部分功能的物质基础。

计算机系统是由人员（people）、数据（data）、设备（device）、程序（program）、规程（regulation）等几部分组成的有机整体，共同完成相关的数据采集、加工处理。其中，设备主要是指计算机及相关设备，中央处理器（central processing unit，CPU）、存储、输入设备和输出设备等硬件配置对运算速度和处理能力起到重要作用。程序是软件系统，应用管理技术、计算技术等对数据进行处理。

人在计算机系统中起着主导作用，系统发挥的作用在很大程度上取决于计算机使用人员素质的高低。

1.2.1 硬件系统

一般地，计算机硬件系统包括主机、外存储器、输入设备、输出设备、系统总线。

(1) 计算机的主机主要由 CPU 和内存储器(简称内存)两大部分组成。

- CPU 主要由控制器和运算器(以及一些寄存器)组成。其中,控制器是计算机的指挥与控制中心,主要作用是控制管理计算机系统。它按照程序指令的操作要求向计算机的各个部分发出控制信号,使整个计算机协调一致地工作。运算器是对数据进行加工处理的部件,负责完成各种算术运算、逻辑运算和比较等。CPU 的性能主要取决于它在每个时钟周期内处理数据的能力和时钟频率(主频)。
- 内存储器是 CPU 可以直接访问的存储器。

(2) 外存储器(简称外存),如磁盘、光盘等,一般用来存储需要长期保存的各种程序和数据。外存不能被 CPU 直接访问,其存储的信息必须先调入内存储器。

(3) 输入设备是向计算机中输入信息的设备,常用输入设备有键盘、鼠标、图形扫描仪、麦克风等。

(4) 输出设备负责把计算机处理数据完成的结果转换成用户需要的形式传送给人们,或传送给某种存储设备保存起来备用。常用输出设备有显示器、打印机、绘图仪等。

(5) 系统总线是计算机系统中 CPU、内存储器和外部设备之间传送信息的公用通道。包括:

- 数据总线(data bus):用于在 CPU、存储器和输入输出设备间传递数据。
- 地址总线(address bus):用于传送存储单元或输入输出接口地址信息。
- 控制总线(control bus):用于传送控制器的各种信号。

1.2.2 软件系统

计算机软件系统可以分为系统软件和应用软件。

(1) 系统软件是控制和协调计算机及其外部设备、支持应用软件的开发和运行的软件。一般包括操作系统、编译程序、诊断程序、系统服务程序、语言处理程序、数据库管理系统和网络通信管理程序等。

- 操作系统是一些程序的集合,它的功能是统一管理和分配计算机系统资源,提高计算机工作效率,同时方便用户使用计算机。它是用户与计算机之间的联系纽带,用户通过操作系统提供的各种命令使用计算机。
- 诊断程序是计算机管理人员用来检查和判断计算机系统故障,并确定发生故障的器件位置的专用程序。
- 语言处理程序是用于编写计算机程序的计算机语言,可分为机器语言、汇编语言和高级语言三大类。机器语言是用二进制代码(由 0 和 1 组成的计算机可以识别的代码)指令来表示各种操作的计算机语言;汇编语言是一种用符号表示指令的程序设计语言;高级语言是接近于人类自然语言和数学语言的程序设计语言,它是独立于具体的计算机而面向过程的计算机语言。用后两种语言编写的程序,必须通过相应的语言处理程序(编译系统),将它转换成机器语言才能执行。
- 数据库管理系统是一套软件,它是操纵和管理数据库的工具。
- 网络通信管理程序是用于计算机网络系统中的通信管理软件,其作用是控制信息的传送和接收。

(2) 应用软件是直接服务于用户的程序系统,一般分为两类:一类是为特定需要开发的实用程序,如订票系统、辅助教学软件等;另一类是为了方便用户使用而提供的软件工具,如图形处理软件、电子报表处理软件等。

1.2.3 计算机硬件与软件的关系

计算机硬件与软件的关系主要体现在以下三个方面。

(1) 相互依存。计算机硬件与软件的产生与发展本身就是相辅相成、相互促进的,二者密不可分。硬件是软件的基础和依托;软件是发挥硬件功能的关键,是计算机的灵魂。在实际应用中二者更是缺一不可,硬件与软件缺少哪一部分,计算机都无法使用。许多硬件所能达到的功能常常需要通过软件配合来实现,如中断保护,既要有硬件实现中断屏蔽保留现场,又要求有软件来完成中断的分析处理;又如操作系统诸多功能的实现,都需要硬件支持。

(2) 无严格界面。虽然计算机的硬件与软件各有分工,但是在很多情况下软硬件之间的界面是浮动的。计算机某些功能既可以由硬件实现,也可以由软件实现。随着计算机技术的发展,一些过去用软件实现的功能现在可以嵌入硬件系统来实现,而且速度和可靠性都大为提高。

(3) 相互促进。无论从实际应用还是从计算机技术的发展看,计算机的硬件与软件之间都是相互依赖、相互影响、相互促进的。硬件技术的发展会对软件提出新的要求,促进软件的发展;反之,软件发展又对硬件提出新的课题。

1.3 计算机数据管理技术发展过程

各类信息系统都需要大量的数据作为基础,数据处理的中心问题是数据管理。数据管理是指对数据的组织、分类、编码、存储、检索和维护。

与其他技术的发展一样,计算机数据管理也经历了由低级到高级的发展过程。计算机数据管理随着计算机硬件(主要是外存储器)、软件技术和计算机应用范围的发展而不断发展,多年来大致经历了如下四个阶段。

- 人工管理阶段。
- 文件系统阶段。
- 数据库系统阶段。
- 分布式数据库系统阶段。

1.3.1 人工管理阶段

计算机早期主要用于科学计算,当时在硬件方面,外存储器只有卡片、纸带、磁带,没有像磁盘这样可以随机访问、直接存取的外存储器。在软件方面,没有专门管理数据的软件,数据由计算或处理它的程序自行携带,数据处理方式基本是批处理。

这一时期数据管理的特点是：

(1) 数据与程序不具有独立性。

一组数据对应一组程序。这就使得程序依赖于数据，如果数据的类型、格式或者数据量、存取方法、输入输出方式等改变了，程序必须做相应的修改。

(2) 数据不共享。

由于数据是面向应用程序的，在一个程序中定义的数据，无法被其他程序利用，因此程序与程序之间存在大量的重复数据。

(3) 没有对数据进行管理的软件。

数据管理任务(包括存储结构、存取方法、输入输出方式等)完全由程序设计人员负责，这就给应用程序设计人员增加了很大的负担。

1.3.2 文件系统阶段

在这一阶段，程序与数据有了一定的独立性，程序和数据分开存储，有了程序文件和数据文件的区别。数据文件可以长期保存在外存储器上多次存取，如进行查询、修改、插入、删除等操作。数据的存取以记录为基本单位，并出现了多种文件组织形式，如顺序文件、索引文件、随机文件等。

文件系统阶段对数据的管理虽然有了进步，但一些根本性问题仍然没有彻底解决，主要表现在以下三个方面。

(1) 数据冗余大。

数据冗余是指不必要的重复存储，同一数据项重复出现在多个文件中。在文件系统下，数据文件基本上与各自的应用程序相对应，数据不能以记录和数据项为单位共享。即使有部分数据相同，只要逻辑结构不同，用户就必须各自建立自己的文件，这不仅浪费存储空间、增加更新开销，更严重的是，由于不能统一修改，容易造成数据的不一致性。

(2) 缺乏数据独立性。

文件系统中的数据文件是为了满足特定业务领域某部门的专门需要而设计的，服务于某一特定应用程序。数据和程序相互依赖，如果改变数据的逻辑结构或文件的组织方法，必须修改相应的应用程序。同样道理，如果修改应用程序，如改用另一种程序设计语言来编写程序，也将影响数据文件的结构。

(3) 数据无法集中管理。

除了对记录的存取由文件系统承担以外，文件没有统一的管理机制，其安全性与完整性无法保障。数据的维护任务仍然由应用程序来承担。

这些问题阻碍了数据处理技术的发展，不能满足日益增长的信息需求，这既是数据库技术产生的原动力，也是数据库系统产生的背景。应用需求和计算机技术的发展促使人们研究一种新的数据管理技术——数据库技术。

1.3.3 数据库系统阶段

从 20 世纪 60 年代后期开始，计算机应用于管理的规模更加庞大，需要计算机管理的数

据量急剧增长，并且对数据共享的需求日益增强。大容量磁盘系统的采用使计算机联机存取大量数据成为可能。同时，软件价格上升，硬件价格相对下降，使独立开发系统维护软件的成本增加，文件系统的数据管理方法已无法适应开发应用系统的需要。为解决数据的独立性问题，实现数据的统一管理，达到数据共享的目的，出现了数据库技术。

数据库(database，DB)是通用化的相关数据集合，它不仅包括数据本身，而且包括关于数据之间的联系。数据库中的数据不是只面向某一项特定应用，而是面向多种应用，可以被多个用户、多个应用程序共享。例如，某个企业、组织或行业所涉及的全部数据的汇集。其数据结构独立于使用数据的程序，对于数据的增加、删除、修改和检索由系统进行统一的控制，而且数据模型也有利于将来应用的扩展。

为了让多种应用程序并发地使用数据库中具有最小冗余的共享数据，必须使数据与程序具有较高的独立性。这就需要一个软件系统对数据实行专门管理，提供安全性和完整性等统一控制机制，方便用户以交互命令或程序方式对数据库进行操作。数据库系统是一个完整的解决方案，包括硬件、软件和人员等多个组成部分，旨在提供一个全面的数据存储、管理和应用环境。

为数据库的建立、使用和维护而配置的软件称为数据库管理系统(database management system，DBMS)，它是在操作系统支持下运行的。数据库已成为各类信息系统的核心基础，在数据库管理系统支持下数据与程序的关系如图 1-2 所示。

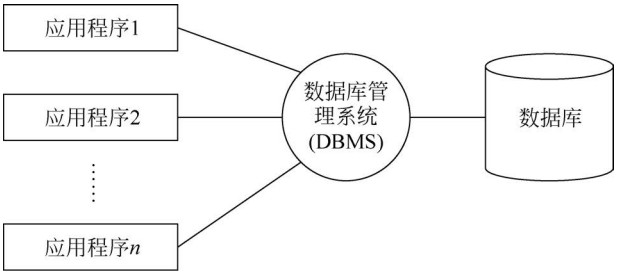

图 1-2 数据库系统中数据与程序的关系

数据库的主要特点是：

(1) 实现数据共享，减少数据冗余。

在数据库系统中，对数据的定义和描述已经从应用程序中分离开来，通过数据库管理系统来统一管理。数据的最小访问单位是数据项，既可以按数据项的名称存取库中某一个或某一组数据项，也可以存取一条记录或一组记录。

建立数据库时，应当以面向全局的观点组织库中的数据，而不能像文件系统那样仅仅考虑某一部门的局部应用。数据库中存放全部数据，某一类应用通常仅使用总体数据的子集，这样才能发挥数据共享的优势。

(2) 采用特定的数据模型。

数据库中的数据不是一盘散沙，必须表示出数据之间所存在有机的关联才能反映现实世界事物之间的联系。也就是说，数据库中的数据是有结构的，这种结构由数据模型表示出来。

文件系统只表示记录内部的联系，类似于属性之间的联系，而不涉及不同文件记录之间

的联系。要想在不同文件中查询相关的数据,必须编写一个程序。

例如,有三个文件:图书(图书 ID,分类号,书名,作者,出版单位,单价);读者(借书证号,姓名,性别,单位,职称,地址);借阅(借书证号,图书 ID,借阅日期,备注)。要想查找某人所借图书的书名、出版单位及借阅者的职称,则必须编写一段逻辑程序来实现。

数据库系统不仅表示属性之间的联系,而且表示实体之间的联系。只要定义好数据模型,上述询问可以非常容易地联机查到。关于数据模型将在数据库技术一章中详细介绍。

(3) 具有较高的数据独立性。

使用数据库系统后,应用程序对数据结构和存取方法有较高的独立性。数据的物理存储结构与用户看到的逻辑结构可以有很大差别。用户只以简单的逻辑结构来操作数据,无须考虑数据在存储器上的物理位置与结构。

(4) 有统一的数据控制功能。

数据库作为多个用户和应用程序的共享资源,对数据的存取往往是并发的,即多个用户同时使用同一个数据库。数据库系统必须提供并发控制功能、数据的安全性控制功能和数据的完整性控制功能。

1.3.4 分布式数据库系统阶段

分布式数据库系统是数据库技术和计算机网络技术相结合的产物。分布式数据库系统是一个逻辑上统一、地域上分布的数据集合,是计算机网络环境中各个节点局部数据库的逻辑集合,同时受分布式数据库管理系统的控制和管理,如图 1-3 所示。

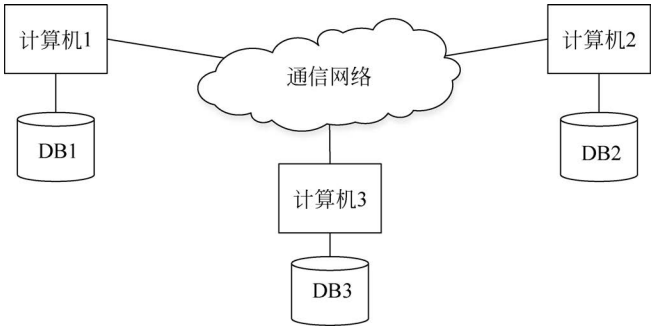

图 1-3 分布式数据库系统

分布式数据库系统在逻辑上像一个集中式数据库系统,实际上数据存储在处于不同地点的计算机网络的各个节点上。每个节点都有自己的局部数据库管理系统,它有很高的独立性。用户可以由分布式数据库管理系统(网络数据库管理系统)通过网络通信相互传输数据。分布式数据库系统有高透明性,每台计算机上的用户并不需要了解他所访问的数据究竟在什么地方,就像在使用集中式数据库一样。其主要优点有:

(1) 局部自主。

网络上每个节点的数据库系统都具有独立处理本地事务的能力(大量的),而且各局部节点之间也能够互相访问、有效地配合处理更复杂的事务。因此,分布式数据库系统特别适合各个部门的地理位置分散的组织机构。例如,银行业务、飞机订票、企业管理等。

(2) 可靠性和可用性。

分布式系统比集中式系统有更高的可靠性,在个别节点或个别通信链路发生故障的情况下可以继续工作。一个局部系统发生故障不至于导致整个系统停顿或破坏,只要有一个节点上的数据备份可用,则数据是可用的。可见,支持一定程度的数据冗余是充分发挥分布式数据库系统优势的先决条件之一。

(3) 效率和灵活性。

分布式系统分散了工作负荷,缓解了单机容量的压力。数据可以存储在邻近的常用节点上,如果本节点的数据子集包含了要查询的全部内容,显然比集中式数据库在全集上查找节省时间。

系统易于实现扩展。例如,一个单位要增加新的机构,分布式数据库系统能够在对现有系统影响较小的情况下实现扩充。由此,扩大系统规模比集中式系统更加方便、经济、灵活。

1.3.5 信息系统发展历程

信息系统是指为了某些明确的目的而建立的由人员、设备、程序和数据集合构成的统一整体。信息系统的主要功能是提供信息,以支持一个组织机构的运行、管理和决策。更确切地说,信息系统将不适用的数据形式加工成可利用的形式。

一个信息系统的质量取决于它是否能及时地为用户提供所需要的信息。在一个组织机构中,不同阶层的管理人员因其管理的目标不同,所需要的信息也不相同。信息系统针对各个层次的需求,通过计算机实现信息支持,达到辅助管理的目的。

信息系统可分为如下三类。

- 电子数据处理系统。
- 管理信息系统。
- 决策支持系统。

(1) 电子数据处理系统(electronic data processing system,EDPS)。

电子数据处理系统是用计算机代替繁杂的手工事务处理工作,其目的是提高数据处理的准确性、及时性,节约人力,提高工作效率。例如,计算机运行会计核算软件,对会计的"簿记"事务进行常规处理,提供数据查询、会计报表等功能,使会计部门的日常工作自动化。

(2) 管理信息系统(management information system,MIS)。

管理信息系统是由若干子系统构成的一个集成的人机系统,从组织的全局出发,实现数据共享,提供分析、计划、预测、控制等方面的综合信息。其主要目的是发挥系统的综合效益,提高管理水平。

例如,某企业管理信息系统由技术管理子系统、人事管理子系统、财务管理子系统、物资管理子系统、生产管理子系统、设备管理子系统、销售管理子系统组成。实现计算机管理能够迅速、准确地提供有关信息,不仅有力地支持各个职能部门的组织管理,并且通过信息共享加强了各子系统之间的协同,使整个系统有机地联系起来,同时为企业领导制订计划、确定经营目标、指挥生产提供信息支持,从而大大提高企业的综合效益,增强市场竞争能力。

(3) 决策支持系统(decision support system,DSS)。

决策支持系统是为决策过程提供有效的信息和辅助决策手段的人机系统。其主要目的

是帮助决策者提高决策的科学性及有效性。

计算机辅助决策必须积累大量的数据、案例、方法、模型,更进一步地,还可以利用知识库系统、专家系统。决策支持系统的服务对象是面向某种决策问题的管理人员,它协助决策者在求解问题的过程中方便地检索出相关数据,对多种可选方案进行比较测试,然后做出决定。

这里需要强调指出,决策支持系统只能对决策提供支持,并不能由计算机代替人,自动化地做出决定,人是决策行动的主体。例如,不同的管理人员运行同一套决策支持系统软件时,可能做出不同的决策结果。

1.4 计算机软件开发技术发展过程

在计算机出现的初期,人们主要着力于计算机硬件的研制,仅用机器指令来编制可运行的程序,程序只是作为硬件的附属品存在。随着硬件的发展以及使用范围的扩大,为使系统正常工作且能充分发挥硬件的效率和潜力,必须配备完善的软件系统,软件技术作为一个独立的分支得到迅速发展。从狭义上理解,软件即是程序设计;从广义上讲,软件应包括程序、相应的数据(数据库)和文档三个方面。因此,软件技术是随着硬件的发展而发展的,而软件的发展与完善又促进硬件技术的新发展,硬件和软件组成一个相互依存、相互促进的有机整体。

1.4.1 高级语言阶段

20 世纪 50 年代末,John Backus 首先完成 FORTRAN 的编译系统,此后十年中,针对不同的应用领域出现了 ALGOL 60、COBOL、LISP 等高级语言。直到 20 世纪 60 年代末出现的 PL/1 和 ALGOL 68 对这一时期的语言特征做了一次总结。这一时期,编译技术代表了整个软件技术,软件工作者追求的主要目标是设计和实现在控制结构和数据结构方面表现能力强的高级语言。如为了避免语句的二义性,提出语义形式化要求,1959 年 Backus 提出一种描述高级语言语法和语义的方法(BNF),1960 年 K. Samelsen 与 F. L. Bauer 提出用先进后出的栈的技术实现表达式翻译。1963 年 R. W. Floyd 提出优先算子法,引入优先顺序概念,它与栈的技术结合起来可以实现高级语言的语法分解。但在这一时期内,编译系统主要是靠手工编制,自动化程度很低。

1.4.2 结构化程序设计阶段

20 世纪 70 年代是计算机技术蓬勃发展的时代。由于磁盘的问世,操作系统迅速发展;商业数据处理等非数值计算应用的发展,使数据库成为独立发展的领域;通信设备的完善,又促成计算机网络的发展;同时,由于大规模集成电路的飞速发展,硬件造价的下降,计算机应用范围的扩大,使软件的规模增大,软件的复杂性增加,由此产生了软件可靠性差的问题,许多耗资巨大的软件项目由于软件的错误导致巨大的经济损失,从而出现了所谓的“软件