

Access 2016数据库应用技术 案例教程

主 编 刘 垣 林敦欣

副主编 郭李华 徐沛然 王一蕾

清华大学出版社

北 京

内 容 简 介

在课程素养教育理念引领下,本书紧扣工程教育认证标准,致力于打造基于Access数据库的人工智能通识课程教材。本书共10章,第1章为数据库系统概述,介绍数据库系统的概念与结构、数据模型、关系数据库、国产数据库、数据库设计方法与步骤等数据库领域的基础核心内容。第2章为“人工智能+”数据技术,概述人工智能的基础理论、国家数据基础设施、我国人工智能产业和数据库产业,并推介一个AI+数据中心应用的优秀案例。第3章至第9章依次介绍Access集成环境中表、查询、窗体、报表、宏和模块等6种标准数据库对象的基本知识和基础应用。第10章为数据库应用系统开发,介绍软件工程基础理论和人工智能大语言模型,并利用国产DeepSeek大模型辅助数据库应用开发,还介绍上海盟威Access软件快速开发平台的应用。

通过学习本书,读者能够系统了解人工智能的发展历程与核心基础理论、我国人工智能产业的发展现状,以及国产大语言模型在数据库领域的应用实践;同时,还能掌握关系数据库系统的基本理论、Access 2016的操作技能与应用方法,从而具备运用“人工智能+”技术开发Access小型数据库应用系统的能力。

本书涵盖《全国计算机等级考试二级Access数据库程序设计考试大纲(2025年版)》的核心知识点,配套有《Access 2016数据库应用技术案例教程学习指导》(ISBN: 978-7-302-69316-1),可作为高等院校数据库课程的教材,也可作为全国计算机等级考试的备考用书、数据库应用技术的学习用书,以及人工智能爱好者的Access培训或参考用书。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。举报:010-62782989, beiqinquan@tup.tsinghua.edu.cn。

图书在版编目(CIP)数据

Access 2016数据库应用技术案例教程 / 刘垣, 林敦欣主编.

北京:清华大学出版社, 2025. 6. -- ISBN

978-7-302-69315-4

I. TP311.132.3

中国国家版本馆CIP数据核字第20250JP029号

责任编辑:王 定

封面设计:周晓亮

版式设计:思创景点

责任校对:成凤进

责任印制:杨 艳

出版发行:清华大学出版社

网 址: <https://www.tup.com.cn>, <https://www.wqxuetang.com>

地 址:北京清华大学学研大厦A座

邮 编:100084

社总机:010-83470000

邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:北京瑞禾彩色印刷有限公司

经 销:全国新华书店

开 本:185mm×260mm

印 张:21

字 数:550千字

版 次:2025年7月第1版

印 次:2025年7月第1次印刷

定 价:79.80元

产品编号:110645-01

前言

P R E F A C E

新故相推，日升不滞。人工智能(AI)浪潮以雷霆之势席卷全球，AI凭借其强大影响力，以前所未有的速度和广度，深刻改变着世界。2024年，国务院政府工作报告首次提出“人工智能+”行动，教育部启动了教育系统人工智能大模型应用示范行动。下一步，将打造人工智能通识课程体系，赋能理工农医文等各类人才培养。

我们在课程素养教育理念的引领下，深入研究了人工智能与数据库技术的融合发展路径，紧扣工程教育认证标准，重新编排了原版教材的内容架构，并在其中添加了更多的中国元素，力求使本书符合教育部倡导的人工智能通识课程体系要求，为研究和探索AI赋能的教育改革创新提供一些思路，帮助读者建立更加完善、更具前瞻性的数据库知识体系。

微软公司自1992年11月首次推出Access以来，已历经14个版本的变迁，拥有比较稳定的用户群体。本书不仅涵盖《全国计算机等级考试二级Access数据库程序设计考试大纲(2025年版)》的核心内容，还考虑了将来与国产数据库技术的衔接，以及人工智能相关理论与技术的融合应用。第1章为数据库系统概述，介绍数据库系统的概念与结构、数据模型、关系数据库、国产数据库、数据库设计方法与步骤等数据库领域的基础核心内容。初学数据库应用技术的读者可以先泛读本章，待学完后续各章后再精读，并完成第1章的实验案例。第2章为“人工智能+”数据技术，概述人工智能的基础理论、国家数据基础设施、我国人工智能产业和数据库产业，并推介一个AI+数据中心应用的优秀案例。第3章至第9章依次介绍Access集成环境中6种标准数据库对象的基本知识和基础应用。第10章为数据库应用系统开发，介绍软件工程基础理论和人工智能大语言模型，并利用国产DeepSeek大模型辅助数据库应用开发，还介绍上海盟威Access软件快速开发平台的应用。

本书采用“目标导向+思政引领”的双重教学设计：每章开篇设置“知识目标”明确学习要点，“素质目标”聚焦能力培养，“学习指南”提供方法论指导，并辅以可视化“思维导图”，通过国产人工智能大模型辅助学习和知识图谱构建，帮助读者系统掌握知识体系。每章小结特别设置“拓展阅读”板块，有机融入坚持和发展中国特色社会主义的理论精髓、实现中华民族伟大复兴中国梦的实践路径、新时代科学家精神与工匠精神的内涵诠释，同时结合《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规，培养读者的科技伦理意识与法治素养，实现专业知识学习与社会主义核心价值观培育的深度融合。

2024年12月30日，国家数据局发布了第一批数据领域常用名词解释，这一重要文件为数据相关领域的标准化与规范化发展提供了权威指导。我们据此对教材中涉及的相关名词、概念与表述进行了全面梳理与修订，以确保教材内容与国家数据领域的标准规范保持一致，为读者提供一个准确、权威的知识体系。

通过学习本书，读者可以了解人工智能的发展历程、核心基础理论和大语言模型在数据库中的应用，掌握关系数据库系统的基本理论、Access 2016的基本操作和应用技能，并具备运用人工智能+技术开发Access小型数据库应用系统的能力。本书案例均在64位Win10/Win11+Access 2016环境下调试通过。

本书为福建省“十四五”普通高等教育本科规划教材建设项目的成果。其编写人员均为高校计算机类教学一线教师，对Access数据库教学进行过较深入的探索。书中案例大多是编者多年教学和实践的总结。本书由刘垣、林敦欣任主编，郭李华、徐沛然、王一蕾任副主编，参与编写工作的还有苏备迎、温馨、韩宜航、郑兆铨和陈治杰等人，本书的完成也离不开原版教材作者连贻捷、刘琰、林铭德和张波尔等老师的贡献。本书的编写得到了福建理工大学、福州大学、湖北工程学院、福建农林大学等多所院校的大力支持，在此一并表示感谢！

在本书的编写过程中，我们始终秉持严谨的态度，对内容进行了多番梳理与雕琢，但由于水平有限，书中难免有疏漏之处，在此，恳请广大读者批评指正。

本书提供教学课件、教学大纲、电子教案、案例源文件以及思考与练习参考答案，读者可扫下列二维码获取。



教学课件



教学大纲



电子教案



案例源文件



思考与练习参考答案

编者

2025年3月于榕城旗山



C O N T E N T S

第1章 数据库系统概述 1	
1.1 数据库技术..... 3	
1.1.1 数据管理技术的产生与发展..... 3	
1.1.2 基本术语..... 9	
1.1.3 数据库系统的结构..... 13	
1.1.4 国产数据库..... 15	
1.2 数据模型..... 17	
1.2.1 数据模型的分层..... 17	
1.2.2 概念模型的表示方法..... 22	
1.2.3 关系模型的基本术语及性质..... 26	
1.2.4 关系运算..... 29	
1.2.5 关系的完整性..... 30	
1.3 数据库设计..... 32	
1.3.1 数据库系统的需求分析..... 32	
1.3.2 概念结构设计..... 33	
1.3.3 逻辑结构设计..... 34	
1.3.4 物理结构设计..... 37	
1.3.5 数据库的实施..... 38	
1.3.6 数据库的运行和维护..... 38	
1.4 本章小结..... 39	
1.5 思考与练习..... 40	
1.5.1 选择题..... 40	
1.5.2 填空题..... 41	
1.5.3 简答题..... 41	
第2章 “人工智能+”数据技术 43	
2.1 人工智能概述..... 45	
2.1.1 人工智能的定义..... 45	
2.1.2 人工智能的起源与发展..... 46	
2.1.3 人工智能的主要学派..... 49	
2.1.4 机器学习和深度学习..... 50	
2.1.5 机器学习的类型..... 51	
2.1.6 AI的技术发展方向..... 53	
2.2 我国人工智能产业..... 54	
2.2.1 AI产业发展现状..... 54	
2.2.2 AI标准化体系建设总体要求..... 54	
2.2.3 人工智能标准体系结构..... 55	
2.2.4 人工智能治理..... 56	
2.2.5 AI+数据中心“绿·智·弹性” 解决方案..... 57	
2.3 国家数据基础设施..... 59	
2.3.1 数据基础设施内涵..... 59	
2.3.2 发展愿景目标..... 59	
2.3.3 总体技术架构..... 59	
2.4 我国数据库产业..... 60	
2.4.1 数据库产业发展现状..... 60	
2.4.2 数据库支撑体系..... 61	
2.4.3 数据库关键技术发展趋势..... 61	
2.5 本章小结..... 65	
2.6 思考与练习..... 66	
2.6.1 选择题..... 66	
2.6.2 填空题..... 66	
2.6.3 简答题..... 66	
第3章 数据库和表 67	
3.1 Access数据库..... 69	
3.1.1 Access的发展历程..... 69	
3.1.2 建立新数据库..... 70	
3.1.3 数据库的基本操作..... 72	
3.2 创建表..... 75	

3.2.1	使用内置模板创建表	76	4.2.2	使用“查询设计视图”创建 查询	124
3.2.2	使用表设计创建表	77	4.3	选择查询	125
3.2.3	通过输入数据创建表	79	4.3.1	创建不带条件的查询	125
3.2.4	使用SharePoint列表创建表	80	4.3.2	查询条件	126
3.2.5	通过获取外部数据创建表	81	4.3.3	创建带条件的选择查询	127
3.2.6	主键的设置	84	4.3.4	在查询中使用计算	128
3.3	数据类型与字段属性	85	4.4	参数查询	130
3.3.1	数据类型	85	4.5	交叉表查询	131
3.3.2	字段属性	86	4.6	操作查询	132
3.4	建立表之间的关系	95	4.6.1	追加查询	132
3.4.1	创建表关系	96	4.6.2	更新查询	133
3.4.2	查看与编辑表关系	97	4.6.3	删除查询	134
3.4.3	实施参照完整性	98	4.6.4	生成表查询	134
3.4.4	设置级联选项	98	4.7	SQL查询	135
3.5	编辑数据表	99	4.7.1	SQL概述	135
3.5.1	向表中添加与修改记录	99	4.7.2	SQL数据查询语句	136
3.5.2	选定与删除记录	100	4.7.3	SQL数据操纵语句	138
3.5.3	数据的查找与替换	101	4.7.4	使用SQL语句创建查询	139
3.5.4	数据的排序与筛选	102	4.8	本章小结	140
3.5.5	行汇总统计	106	4.9	思考与练习	140
3.5.6	表的复制、删除与重命名	107	4.9.1	选择题	140
3.6	设置数据表格式	108	4.9.2	填空题	141
3.6.1	设置表的行高和列宽	108	4.9.3	简答题	142
3.6.2	设置字体格式	109	第5章	窗体	143
3.6.3	隐藏和显示字段	109	5.1	窗体概述	145
3.6.4	冻结和取消冻结	110	5.1.1	窗体的功能	145
3.7	本章小结	111	5.1.2	窗体的类型	146
3.8	思考与练习	112	5.1.3	窗体的视图	147
3.8.1	选择题	112	5.1.4	窗体的构成	149
3.8.2	填空题	113	5.2	创建窗体	150
3.8.3	简答题	114	5.2.1	自动创建窗体	151
第4章	查询	115	5.2.2	使用向导创建窗体	153
4.1	查询概述	117	5.2.3	使用“空白窗体”按钮创建 窗体	156
4.1.1	查询的功能	117	5.2.4	使用设计视图创建窗体	157
4.1.2	查询的类型	117	5.3	设计窗体	157
4.1.3	查询视图	118	5.3.1	窗体设计视图的组成与主要	
4.2	创建查询的方式	119			
4.2.1	使用向导创建查询	119			

功能	157	6.5 本章小结	203
5.3.2 为窗体设置数据源	158	6.6 思考与练习	204
5.3.3 窗体的常用属性与事件	159	6.6.1 选择题	204
5.3.4 在窗体中添加控件的方法	161	6.6.2 填空题	205
5.3.5 常用控件及其功能	161	6.6.3 简答题	205
5.3.6 常用控件的使用	164	第7章 宏	206
5.4 修饰窗体	172	7.1 宏的概述	208
5.4.1 主题的应用	172	7.1.1 什么是宏	208
5.4.2 条件格式的使用	173	7.1.2 宏的类型	208
5.4.3 窗体的布局及格式调整	174	7.1.3 宏的设计视图	209
5.5 定制用户入口界面	175	7.1.4 常用的宏操作	210
5.5.1 创建导航窗体	175	7.1.5 宏的结构	211
5.5.2 设置启动窗体	176	7.2 独立宏的创建与运行	212
5.6 本章小结	177	7.2.1 创建独立宏	212
5.7 思考与练习	178	7.2.2 独立宏的运行	213
5.7.1 选择题	178	7.2.3 自动运行宏	214
5.7.2 填空题	178	7.2.4 宏的调试	215
5.7.3 简答题	179	7.3 嵌入宏的创建与运行	217
第6章 报表	180	7.3.1 创建嵌入宏	217
6.1 报表概述	182	7.3.2 通过事件触发宏	217
6.1.1 报表的类型	182	7.4 数据宏的创建与运行*	218
6.1.2 报表的组成	183	7.4.1 创建事件驱动的数据宏	218
6.1.3 报表的视图	186	7.4.2 创建已命名的数据宏	220
6.2 创建报表	186	7.4.3 数据宏的运行	220
6.2.1 使用“报表”按钮自动创建		7.5 宏的应用	220
报表	186	7.5.1 自定义菜单简介	220
6.2.2 使用“报表设计”按钮创建		7.5.2 自定义功能区菜单的创建	220
报表	187	7.5.3 自定义快捷菜单的创建	223
6.2.3 使用“空报表”按钮创建		7.6 本章小结	225
报表	189	7.7 思考与练习	225
6.2.4 使用“报表向导”按钮创建		7.7.1 选择题	225
报表	190	7.7.2 填空题	226
6.2.5 创建标签报表	192	7.7.3 简答题	226
6.2.6 创建图表报表	193	第8章 VBA程序设计	227
6.3 报表设计	195	8.1 VBA语言概述	229
6.3.1 设计报表的外观	195	8.1.1 程序设计概述	229
6.3.2 报表的排序、分组和计算	199	8.1.2 VBA编程环境	230
6.4 导出报表	202	8.2 VBA编程基础	235

8.2.1	数据类型	235	9.5	思考与练习	296
8.2.2	常量	237	9.5.1	选择题	296
8.2.3	变量	238	9.5.2	填空题	296
8.2.4	数组	241	9.5.3	简答题	296
8.2.5	运算符和表达式	242	9.5.4	设计题	296
8.2.6	VBA内部函数	245	第10章 数据库应用系统开发	298	
8.3	程序控制结构	249	10.1	软件工程基础	300
8.3.1	VBA基本语句	249	10.1.1	基本概念	300
8.3.2	顺序结构	250	10.1.2	软件测试	302
8.3.3	分支结构	253	10.1.3	程序调试	303
8.3.4	循环结构	260	10.2	大模型赋能应用开发	304
8.3.5	过程与函数	269	10.2.1	什么是大模型	304
8.4	本章小结	271	10.2.2	大模型的发展	305
8.5	思考与练习	272	10.2.3	大模型应用体验	306
8.5.1	选择题	272	10.2.4	大模型辅助开发案例	306
8.5.2	填空题	273	10.3	驾驶人科目一模拟考试系统	312
8.5.3	简答题	274	10.3.1	系统功能简介	313
8.5.4	程序设计题	274	10.3.2	系统VBA源代码简介	316
第9章 ADO数据库编程	276		10.3.3	系统的维护与升级	318
9.1	ADO概述	278	10.4	客户管理系统	318
9.1.1	数据库引擎和接口	278	10.4.1	系统功能简介	318
9.1.2	ADO	279	10.4.2	系统VBA源代码简介	319
9.2	ADO主要对象	280	10.5	采购报销管理系统	323
9.2.1	ADO对象模型	280	10.5.1	系统功能简介	323
9.2.2	Connection对象	281	10.5.2	盟威平台实现系统功能	324
9.2.3	Recordset对象	283	10.6	本章小结	326
9.2.4	Command对象	287	10.7	思考与练习	326
9.3	ADO在Access中的应用	289	10.7.1	选择题	326
9.3.1	ADO编程方法	289	10.7.2	填空题	327
9.3.2	ADO编程实例	290	10.7.3	简答题	327
9.4	本章小结	295	参考文献	328	

第 1 章

数据库系统概述

知识目标

1. 掌握数据库系统的基础知识。
2. 掌握关系数据库的基本原理。
3. 掌握数据库应用系统设计方法。
4. 了解数据管理技术的历史和发展趋势。
5. 了解我国数据库发展历程和国产数据库。

素质目标

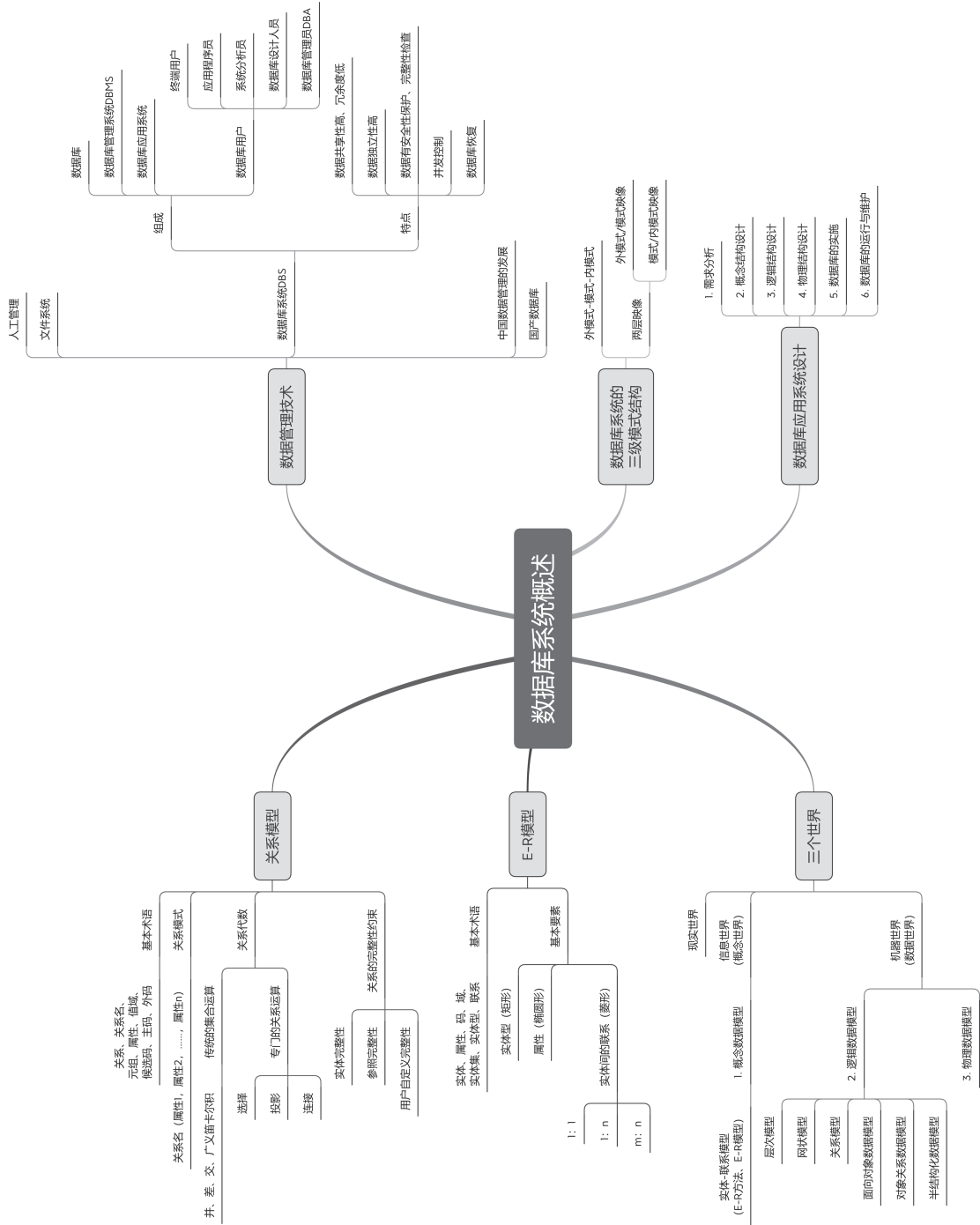
1. 培养学生胸怀祖国，放眼世界的系统思维和全局观念。
2. 培养学生守正创新，踔厉奋发的精神品格和实践才能。

学习指南

本章的知识重点是 1.1.2 节、1.1.3 节和 1.2 节，难点是 1.2 节。

本章内容理论性较强，涵盖的知识面广，涉及的概念多，读者不容易理解掌握。建议初学者循序渐进地学习本章，通过比对、类比等方式识记数据库基本术语和基本原理，依据本章思维导图理清知识脉络，借助人工智能大语言模型工具，深入理解各案例的前提条件和解决方法。

思维导图



1.1 数据库技术

数据库是数据管理的有效技术，是计算机科学的重要分支，是现代大数据管理和分析的基石。数据库技术诞生于20世纪60年代末期，历经数代演变，已造就四位图灵奖获得者：查理士·巴赫曼(1973年)、埃德加·弗兰克·科德(1981年)、詹姆斯·尼古拉·格雷(1998年)和迈克尔·斯通布雷克(2014年)。数据库技术有效地解决了对大量数据进行科学组织、存储、管理和共享的问题。利用数据库技术的理论和设计方法，可实现对数据库数据的有效存取访问、数据处理、数据分析和数据应用。

数据库技术有扎实的理论基础和广泛的应用领域，其目标是在数据库系统中**尽可能减少数据冗余，实现最多的数据共享，并确保数据库系统能安全、高效地检索和处理数据**。农、林、牧、渔业，采矿业，制造业，电力、热力、燃气及水生产和供应业，建筑业，批发和零售业，交通运输、仓储和邮政业，住宿和餐饮业，信息传输、软件和信息技术服务业，金融业，房地产业，租赁和商务服务业，科学研究和技术服务业，教育等国民经济行业都有数据库的应用。人们的日常生活也无时无刻不在享受着数据库应用带来的便利，例如，各种智能终端的安全便捷消费、各类新媒体及其服务、沉浸式体验等。

1.1.1 数据管理技术的产生与发展

数据库技术是应数据管理任务的需要而产生的。数据管理是指对数据进行分类、组织、编码、存储、检索、维护和应用，它是数据处理的中心问题。数据处理是指对各种数据进行采集、存储、检索、加工、传播和应用等一系列活动的总和。随着应用需求的推动和计算机软硬件的发展，**数据管理技术经历了人工管理、文件系统和数据库系统三个阶段**。这三个阶段产生的背景与特点如表 1-1 所示。

表 1-1 数据管理技术三个阶段的比较

背景与特点		三个阶段		
		人工管理阶段	文件系统阶段	数据库系统阶段
背景	应用背景	科学计算	科学计算、数据管理	大规模数据管理
	硬件背景	无直接存取存储设备	磁带、磁鼓、磁盘	大容量磁盘、磁盘阵列
	软件背景	没有操作系统(OS)	有文件系统	有数据库管理系统
	处理方式	批处理	联机实时处理、批处理	联机实时处理、批处理、分布处理
特点	数据的管理者	用户(程序员)	文件系统	数据库管理系统(DBMS)
	数据面向的对象	某一应用程序	某一应用	现实世界(一个部门、企业等)
	数据的共享程度	无共享，冗余度极大	共享性差，冗余度大	共享性高，冗余度小
	数据的独立性	不独立，完全依赖于程序	独立性差	具有高度的物理独立性和一定的逻辑独立性

续表

背景与特点		三个阶段		
		人工管理阶段	文件系统阶段	数据库系统阶段
特	数据的结构化	无结构	记录内有结构，整体无结构	整体结构化，用数据模型描述
点	数据控制能力	应用程序自己控制	应用程序自己控制	由 DBMS 提供数据安全性、完整性、并发控制和恢复能力

1. 人工管理阶段

数据管理技术的人工管理阶段主要指 20 世纪 50 年代中期以前，数据需要由应用程序定义和管理，一个数据集只能对应一个应用程序。**数据无共享，冗余度极大；数据不独立，完全依赖于程序。**这个阶段的计算机很简陋，主要应用于科学计算。计算机硬件状况是：没有直接存取存储设备；软件状况是：没有完整的操作系统，没有管理数据的专门软件；数据处理方式是批处理。批处理方式是对需要处理的数据不做立即处理，待积累到一定程度、一定时间，再成批地进行处理。

赫曼·霍列瑞斯 (1860.02.29—1929.11.17) 被视为现代机器数据处理之父。为解决美国人口普查统计繁难、花费时间长的问题，霍列瑞斯根据自动提花织布机原理，利用穿孔卡片记录美国人口普查信息，发明了穿孔卡片制表机 (punch card tabulating machine)。这种穿孔卡片制表机是机电式计数装置，安装有一组盛满水银的小杯，已穿孔的卡片放置在这些小杯上。卡片上方有几排精心设置的金属探针，探针连接在电路的一端，小杯连接在电路的另一端。只要某根探针遇到卡片上有孔的位置，便会自动掉落下去，与水银接触即接通电路，启动计数装置前进一个刻度。霍列瑞斯的穿孔卡片表达了二进制思想：有孔处能接通电路计数，代表该调查项目为“有”，即 1；无孔处不能接通电路计数，表示该调查项目为“无”，即 0。

1890 年美国人口普查首次选用霍列瑞斯的穿孔卡片制表机，获得巨大成功。在此后的计算机系统里，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。霍列瑞斯发明的制表机和穿孔卡片如图 1-1 和图 1-2 所示，制表机由两部分构成：整理箱 (sorting box) 由制表机控制，分拣机 (sorter) 是后期发展起来的独立机器。霍列瑞斯于 1896 年创办的公司 Tabulating Machine Company 后来成为 IBM(International Business Machines Corporation) 的前身之一。1900—1950 年，穿孔卡片是商务数据存储和检索的基本形式，IBM 是组合和排序穿孔卡片设备，以及利用穿孔卡片数据打印报表的主要供应商。

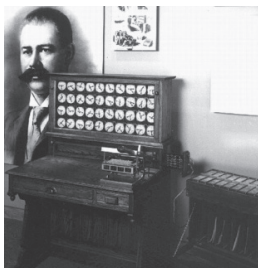


图 1-1 霍列瑞斯和穿孔卡片制表机

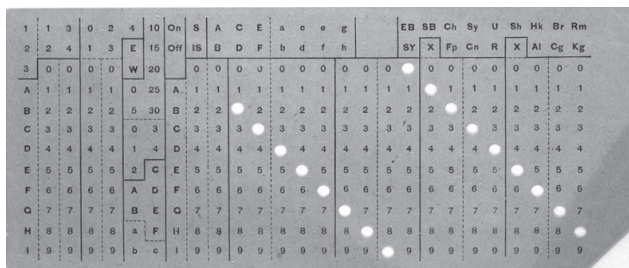


图 1-2 穿孔卡片

2. 文件系统阶段

文件系统阶段主要指 20 世纪 50 年代末到 20 世纪 60 年代中期，这一阶段利用文件系统管理数据。对于一个特定的应用，数据被集中组织存放在多个数据文件或文件组中。为了更好地管理和利用这些数据，会针对该文件组开发特定的应用程序。然而，这种数据管理方式存在一些问题，**数据的共享性差，冗余度大；数据独立性差。**

这个阶段的计算机已有操作系统，在操作系统基础之上建立的文件系统已经成熟并得到广泛应用。计算机除了应用于科学计算，也开始应用于数据管理。数据处理方式不仅有批处理，还有联机实时处理。联机实时处理是指需要对收集到的数据立即进行处理，并及时给出反馈。

在这个阶段，磁带主要被用于数据存储。数据处理是从一个或多个磁带上顺序读取数据，再将数据写回到新的磁带上。数据也可以由一叠穿孔卡片输入，再输出到打印机上。

以美国在 20 世纪 60 年代初制定的阿波罗计划 (Apollo Program/Project Apollo) 为例，阿波罗飞船由约 200 万个零部件组成，它们分散在世界各地制造生产。为了掌握计划进度及协调工程进展，阿波罗计划的主要合约者 Rockwell 公司开发了一个基于磁带的零部件生产计算机管理系统，该系统共使用了 18 盘磁带，虽然可以工作，但效率极低，18 盘磁带中 60% 是冗余数据，维护十分困难。这个系统曾一度成为美国实现阿波罗计划的重大障碍之一。

3. 数据库系统阶段

自 20 世纪 60 年代末期以来，计算机管理的对象规模越来越大，应用范围越来越广，数据量激增，多种语言、多种应用互相覆盖的共享数据集合的需求越来越强烈。这时计算机硬件价格下降，软件价格上升，编制和维护系统软件及应用程序所需的成本相对增加。文件系统作为数据管理手段已经不能满足应用的需求。为解决多用户、多应用共享数据的需求，使数据为尽可能多的应用服务，数据库技术应运而生并不断发展壮大。**从文件系统到数据库系统标志着数据管理技术的飞跃。**数据库系统的详细介绍参见 1.1.2 和 1.1.3 节。

1) 20 世纪 60 年代末至 70 年代

20 世纪 60 年代末，硬盘的广泛使用改变了数据处理的方式，硬盘允许直接对数据进行访问，数据摆脱了磁带和卡片组顺序访问的限制。人们可以创建网状数据库和层次型数据库，可以将表和树这样的数据结构保存在磁盘上。程序员可以构建和操作这些数据结构。

1970 年美国 IBM 公司 San Jose 研究室的埃德加·弗兰克·科德发表了一篇具有里程碑意义的论文 *A Relational Model of Data for Large Shared Data Banks*，定义了关系模型，以及在关系模型中查询数据的非过程化方法，奠定了关系数据库理论的基础。ACM 在 1983 年将该论文列为自 1958 年以来的 25 年中具有里程碑意义的 25 篇研究论文之一。

2) 20 世纪 80 年代

尽管关系模型简单，能够对程序员屏蔽所有实现的细节，在学术界备受重视，但它最初被认为性能不好，没有实际的应用价值。关系数据库在性能上还不能和当时已有的网状数据库和层次型数据库相提并论。这种情况直到 IBM 的 System R 出现才得以改变。与此同时，加州大学伯克利分校 (University of California, Berkeley) 专家迈克尔·斯通布雷克主持开发了 INGRES(interactive graphics and retrieval system) 系统，它后来发展成商品化的

关系数据库系统。

关系数据库凭借其简单易用的优势，最后完全取代了网状数据库和层次型数据库。

在这个时期，人们还对并行数据库和分布式数据库进行了很多探究，对面向对象数据库也有初步的研究。

3) 20 世纪 90 年代

许多数据库厂商推出了并行数据库产品，并开始在数据库中加入对“对象 - 关系”的支持。决策支持和查询再度成为数据库的一个主要应用领域。分析大容量数据的工具有了长足进步。

随着互联网的爆炸式发展，数据库技术有了更加广泛的应用。在这个时期，数据库系统必须支持高速的事务处理，支持对数据的 Web 接口，而且还要有很高的可靠性和 7×24 小时的可用性。

4) 21 世纪第一个十年

万维网联盟 (World Wide Web Consortium, W3C) 于 1998 年 2 月推荐的 XML(eXtensible Markup Language) 兴起，与之相关联的 XQuery 查询语言成为新的数据库技术。虽然 XML 被广泛应用于数据交换和一些复杂数据类型的存储，但关系数据库仍然是构成大多数大型数据库应用系统的核心。为减少系统管理开销，自主计算 / 自动管理技术得到了发展，开源数据库系统 (如 PostgreSQL 和 MySQL) 的应用也显著增长。

用于数据分析的专门的数据库增速惊人，特别是将一张二维表的每列高效地存储为一个单独的数组的列存储，以及为非常大的数据集的分析而设计的高度并行的数据库系统。有几个新颖的分布式数据存储系统被构建出来，以应对庞大的 Web 节点 (如 Amazon、Google、Microsoft 和 Yahoo!) 的数据管理需求。数据挖掘技术被广泛部署和应用，如基于 Web 的产品推荐系统和 Web 页面上的相关广告自动投放。在管理和分析流数据 (如股票市场报价数据、计算机网络监测数据) 方面也取得了重要进展。

2006 年，国务院发布的《国家中长期科学和技术发展规划纲要 (2006—2020 年)》首次提出了“核高基”概念。“核高基”是核心电子器件、高端通用芯片及基础软件产品三者的简称，同时也是与载人航天、探月工程并列的 16 个重大科技专项之一。其中，基础软件包括数据库、操作系统、中间件等。

5) 21 世纪第二个十年

以互联网大数据应用为背景发展起来的分布式非关系型的数据管理系统 NoSQL(Not Only SQL)，融合了 NoSQL 系统和传统数据库事务管理功能的 NewSQL，分析型 NoSQL 技术的主要代表 MapReduce 技术……各类技术的互相借鉴和融合，成为数据管理技术的发展趋势。

大数据已经从概念落到实地，具有体量大、结构多样、时效性强的特点。大数据处理促进了新型计算架构和智能算法的云计算发展，也将 1956 年诞生的人工智能 (Artificial Intelligence, AI) 带入新阶段。我们每个人既是大数据信息的接受者，也是大数据信息的生产者。大数据通常被认为是 PB 或更高数量级的数据 (1YB=1024ZB=1024⁸=2⁸⁰ 字节, 1ZB=1024EB=1024⁷=2⁷⁰ 字节, 1EB=1024PB=1024⁶=2⁶⁰ 字节, 1PB=1024TB=1024⁵=2⁵⁰ 字节, 1TB=1024GB=1024⁴=2⁴⁰ 字节, 1GB=1024MB=1024³=2³⁰ 字节, 1MB=1024KB=1024²=2²⁰ 字节, 1KB=1024Byte=1024¹=2¹⁰ 字节)。

2016年10月美国总统行政办公室(Executive Office of the President)联合美国国家科学技术委员会(National Science and Technology Council, NSTC)共同发布了 *Preparing For The Future of Artificial Intelligence* 研究报告,对人工智能的发展现状、应用领域及目前存在的问题进行了阐述,并向美国政府及相关机构提出了相应的发展建议和对策。

2017年7月20日国务院印发了《**新一代人工智能发展规划**》,这是我国第一部在人工智能领域进行系统部署的规划,将人工智能上升到国家战略层面。该规划提出:到2030年我国人工智能理论、技术与应用总体达到世界领先水平,成为世界主要人工智能创新中心,智能经济、智能社会取得明显成效,为跻身创新型国家前列和经济强国奠定重要基础。

2018年,美国人工智能研究实验室OpenAI发布了基于Transformer架构的GPT(Generative Pre-trained Transformer)模型,它通过大规模数据预训练,能够捕捉复杂的语言结构和语义关系,显著提升了自然语言处理任务的效果。此后,谷歌推出的BERT(Bidirectional Encoder Representations from Transformers)进一步推动了预训练语言模型的发展,通过双向编码器捕捉上下文信息,显著提高了情感分析、命名实体识别等任务的性能。

在大数据、移动互联网、超级计算、传感网、脑科学等新理论新技术的快速发展,以及经济社会发展强烈需求的共同驱动下,人工智能呈现出深度学习、跨界融合、人机协同、群智开放、自主操控等新特征。新一代人工智能相关学科发展、理论建模、技术创新、软硬件升级等整体推进,正在引发链式突破,推动经济社会各领域从数字化、网络化向智能化加速跃升。然而,人工智能的发展存在不确定性。这种影响面广泛的颠覆性技术,可能会冲击现有的法律与社会伦理、侵犯个人隐私、挑战国际关系准则等。

6) 21世纪第三个十年

2020年新型冠状病毒疫情在全球范围内暴发,给人类生命安全带来了严重威胁。大数据、云计算、人工智能等新一代信息技术加速与交通、医疗、教育、金融等领域深度融合,成为战“疫”的强有力武器。

在新冠疫情防控措施的实施下,人们的上网时长大幅增加,线上生活由原先短暂的例外状态转变为常态,由现实世界的补充变成了与现实世界并行的平行世界,人类现实生活开始大规模向虚拟世界迁移。2021年**元宇宙(metaverse)**一词席卷全球,Facebook更名为Meta,取自Metaverse的前缀。元宇宙是人类运用数字技术构建的,由现实世界映射或超越现实世界,可与现实世界交互的虚拟世界。在元宇宙概念刚出现时,我国著名科学家钱学森(1911.12.11—2009.10.31)先生就已经注意到了它,1990年他将虚拟现实技术的元宇宙翻译为“灵境”,并描述了“灵境”之遥触、之遥知、之遥在的场景,以及灵境所展示出的奇妙的境界、虚幻的境界等。2023年8月29日,工业和信息化部、教育部、文化和旅游部、国务院国资委、国家广播电视总局等五部门联合印发政策文件《元宇宙产业创新发展三年行动计划(2023—2025年)》,旨在推动元宇宙产业的高质量发展,加快技术创新与产业融合,助力数字经济成为经济增长的重要引擎。

2023年3月16日,百度正式发布全新一代知识增强大语言模型、生成式AI产品“文心一言”。它基于飞桨深度学习平台和文心知识增强大模型,持续从海量数据和大规模知识中融合学习,具备知识增强、检索增强和对话增强的技术特色,能够与人对话互动,回答问题,协助创作,高效便捷地帮助人们获取信息、知识和灵感。同日,长安汽车官宣,逸达将成为国内首款搭载百度“文心一言”的量产车型。我国还有众多的大模型工具:字节

跳动的豆包、腾讯的混元大模型、阿里巴巴的通义千问、科大讯飞的讯飞星火和昆仑万维的天工 AI 等。

2024 年 12 月 27 日, OpenAI 宣布 2025 年进行公司架构调整, 非营利组织失去控制权。这一转变激起了杰弗里·辛顿等人的担忧, 认为“破坏了 OpenAI 优先考虑公众安全的承诺”。

4. 中国数据管理的发展

中国作为拥有悠久历史的文明古国, 在数千年的发展历程中, 积累了丰富且卓越的数据管理经验。从人口统计到土地丈量, 从赋税征收记录到文物典藏管理, 古人凭借智慧与实践, 构建起一套适应当时社会需求的数据管理体系。

欧洲文艺复兴之后, 西方国家在科学领域占据垄断性优势。近代中国战事连绵不断, 国力羸弱。中华人民共和国成立后, 1956 年, 周恩来(1898.03.05—1976.01.08) 总理领导制定《1956—1967 年科学技术发展远景规划纲要》, 同年 8 月我国成立中国科学院计算技术研究所筹委会, 由 1950 年回国的数学家华罗庚(1910.11.12—1985.06.12) 先生任主任委员。在独立自主、自力更生思想指导下, 1958 年我国第一台电子管计算机 103 机诞生, 1964 年以后我国开始推出一批晶体管计算机, 代表机型有中国科学院计算所的“109 乙”“109 丙”, 15 所和 738 厂的“108 乙”和 320 机, 军事工程学院的 441B 等。

20 世纪 70 年代末, 以萨师煊(1922.12.27—2010.07.11) 先生为代表的老一辈科学家以强烈的责任心和敏锐的学术洞察力, 率先在国内开展数据库技术的教学与研究。萨师煊先生生于福建福州, 是我国数据库学科的奠基人之一, 也是数据库学术活动的积极倡导者和组织者。1978 年萨师煊先生首次在中国人民大学开设数据库课程, 之后, 他将自己的讲稿汇集成《数据库系统简介》和《数据库方法》, 在当时的《电子计算机参考资料》上公开发表供他人免费使用。这是我国最早的数据库学术论文。随后, 他又发表了有关数据模型、数据库设计、数据库管理系统实现和关系数据库理论等诸多方面的论文。1983 年, 他与弟子王珊合作编写了国内第一部系统阐明数据库理论和技术的教材《数据库系统概论》, 2024 年已出版至第六版。

1999 年 8 月, 中国计算机学会数据库专业委员会成立, 标志着中国数据库领域进入了一个组织化、专业化发展的新阶段。

2020 年, 为了更好地推动各行业数据管理工作的开展, 工信部信息技术发展司委托中国电子信息行业联合会牵头负责全国数据管理能力成熟度评估工作体系建设, 根据 GB/T 36073-2018《数据管理能力成熟度评估模型》(Data Capability Maturity Model, DCMM) 国家标准的有关要求, 推动各行业的贯标工作。

2021 年 6 月 7 日, 中国电子信息行业联合会发布《数据管理从业人员能力等级要求》团体标准及编制说明的公告, 将数据管理从业人员分为首席数据官、资深数据管理工程师、数据管理工程师、助理数据管理工程师四个等级。

2021 年 6 月 10 日, 第十三届全国人民代表大会常务委员会第二十九次会议通过《[中华人民共和国数据安全法](#)》, 该法旨在规范数据处理活动, 保障数据安全, 促进数据开发利用, 保护个人、组织的合法权益, 维护国家主权、安全和发展利益。

2022 年 12 月 2 日, 中共中央、国务院发布《关于构建数据基础制度更好发挥数据要素作用的意见》, 以数据产权、流通交易、收益分配、安全治理为重点, 初步提出我国数

据基础制度 20 条政策举措，简称“数据二十条”。它是我国数据基础制度的“四梁八柱”。

2023 年 2 月 17 日，中国电子信息行业联合会举办主题为“贯彻数据二十条，做强做优做大我国数字经济”的首届中国数据治理年会，同时发布了《2022 中国数据管理白皮书》《首席数据官基础和术语》《数据存储安全能力成熟度模型》《数据集成和服务能力评估体系能力要求》《信息技术服务数据安全能力模型》《数据合规管理体系要求》《数据要素市场可信产品评价准则》等标准文件。《2022 中国数据管理白皮书》显示，当前我国数据管理能力呈现 5 个特点：一是我国数据管理水平仍处于初级发展阶段，全国数据管理能力平均水平基线为 2.23，与最高管理能力水平相差 2.77；二是数据管理重实用，轻规范现象突出；三是通信行业数据管理能力水平位居其他行业之首，管理能力平均水平基线高于全国平均基线 1.2；四是数据治理成果日益突出，企业数字化转型加速深化，数据要素市场稳步推进；五是数据治理的共识明显增强，政府进一步强化资金补贴、能力提升、人才培养、行业采信的支持力度，企业贯标积极性快速提升。

2023 年 3 月，中共中央、国务院印发《党和国家机构改革方案》，组建由国家发展和改革委员会管理的国家数据局，负责协调推进数据基础制度建设，统筹数据资源整合共享和开发利用，统筹推进数字中国、数字经济、数字社会规划和建设等。

2024 年 1 月，由国家数据局等 17 个部门联合发布《“数据要素 ×”三年行动计划（2024—2026 年）》。实施“数据要素 ×”行动，就是要发挥我国超大规模市场、海量数据资源、丰富应用场景等多重优势，推动数据要素与劳动力、资本等要素协同，以数据流引领技术流、资金流、人才流、物流，突破传统资源要素约束，提高全要素生产率；促进数据多场景应用、多主体复用，培育基于数据要素的新产品和服务，实现知识扩散、价值倍增，开辟经济增长新空间；加快多元数据融合，以数据规模扩张和数据类型丰富，促进生产工具创新升级，催生新产业、新模式，培育经济发展新动能。

2024 年 7 月，中国通信标准化协会大数据技术标准推进委员会（CCSA TC601）发布《数据库发展研究报告（2024 年）》。

2024 年 8 月 30 日，国务院第 40 次常务会议通过《网络数据安全条例》，自 2025 年 1 月 1 日起施行。

2024 年 10 月 28 日，全国数据标准化技术委员会成立大会暨第一次全体委员会议在京召开；12 月 23 日，国家数据专家咨询委员会成立大会暨第一次全体会议在京召开；12 月 31 日，全国数据标准化技术委员会 2024 年第一次主任办公会在京召开。

1.1.2 基本术语

数据、大数据、数据库、数据库管理系统和数据库系统是与数据库技术密切相关的基本术语。

1. 数据(data)

数据正在以前所未有的速度成为各个领域价值创造的核心驱动力。数据是数据库中存储的基本对象，是描述事物的符号记录。早期的计算机系统主要用于科学计算，处理的数据是整数、实数等数值型数据；现代计算机系统存储和处理的对象十分广泛，除了数值型数据，还可以是非数值型数据，例如文本、音频、视频、图形和图像等。非数值型数据经

过数字化后存入计算机。

2024年12月30日，国家数据局发布第一批数据领域常用名词解释，数据的定义是：数据指任何以电子或其他方式对信息的记录。数据在不同视角下被称为原始数据、衍生数据、数据资源、数据产品和服务、数据资产、数据要素等。

日常生活中，人们通常采用无结构的自然语言描述事物。例如，一个学生的基本情况可描述为：陈榕刚，男，共青团员，2004年3月12日出生，福建生源，专业编号M01等。

日常数据管理中，学生的基本情况通常如表1-2所示。

表 1-2 学生基本情况表

姓名	性别	政治面貌	出生日期	生源地	专业编号
陈榕刚	男	共青团员	2004年3月12日	福建	M01
⋮	⋮	⋮	⋮	⋮	⋮

数据与其语义是不可分的。数据的表现形式需要经过解释才能完全表达其内容，数据的含义（即语义）需要经过解释才能被正确理解。例如，“2004年3月12日”这个数据可能是某人的生日，也可能是纪念日，还可能是某商品的出厂日期等。在表1-2中，其语义已由其所在列的表头栏目名称解释，即出生日期。

记录是计算机中表示和存储数据的一种格式或一种方法。将一个学生的姓名、性别、政治面貌、出生日期、生源地和专业编号等数据组织在一起便构成一条记录，用于描述一个学生的基本情况。这样的数据是有结构的，**表格描述的数据称为结构化数据。**

结构化数据是指一种数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的，并且可以使用关系模型予以有效描述。

数据不能等同于信息(information)。信息论奠基人克劳德·艾尔伍德·香农(1916.04.30—2001.02.24)从通信理论出发，用数学方法定义信息。香农认为：信息是用来消除随机不确定性的东西。控制论创始人诺伯特·维纳(1894.11.26—1964.03.18)认为：信息是人们在适应外部世界，并使这种适应反作用于外部世界的过程中，同外部世界进行互相交换的内容名称。我国信息学专家钟义信教授认为：信息是事物的存在方式或运动状态，以这种方式或状态进行直接或间接的表述。美国信息资源管理专家霍顿认为：信息是为了满足用户决策的需要而经过加工处理的数据。人们从各自研究的角度出发，对信息有不同的理解和定义，随着时间的推移，信息也将被赋予新的含义。

我们可以认为：信息以数据为载体，是具有一定含义的、经过加工处理的数据，是客观事物存在方式和运动状态的反映，对人类决策有帮助和有价值。例如，气象台主要依据事先勘测采集的一组气象数据，如气压、云层、温度、湿度、风力等进行整理加工和综合分析（与经验值比较、统计、运算、推断），从而预报天气情况，如阴、晴、刮风、下雨、下雪等。这里的气压、云层、温度、湿度、风力等称为数据，经过整理加工和综合分析的结果（天气预报）称为信息。

2. 大数据(big data)

依据 GB/T 35295—2017，大数据指具有体量巨大、来源多样、生成极快、多变等特征

并且难以用传统数据体系结构有效处理的包含大量数据集的数据。2014年3月，“大数据”首次被写入《政府工作报告》。

3. 数据库 (database, DB)

数据库从字面上可以理解为存放大量“数据”的“仓库”，只不过这个仓库遵循一定的数学理论创建，位于计算机存储设备中，数据按特定的格式存放。

数据库是长期存储在计算机内，有组织且可共享的大量数据的集合。数据库中的数据按一定的数据模型组织、描述和存储，具有较小的冗余度、较高的数据独立性和易扩展性。数据库中不仅存放数据，还存放数据与数据之间的联系。

随着数据库的发展，出现了数据仓库 (Data Warehouse, DW)。DW 是指在数据准备之后用于永久性存储数据的数据库。数据仓库之父比尔·恩门在经典论著 *Building the Data Warehouse* 第4版中将数据仓库定义为一个面向主题、集成、非易失性和随时间变化的集合，用于支持管理层的决策。

4. 数据库管理系统 (database management system, DBMS)

DBMS 是数据库系统的核心组成部分，是位于用户与操作系统之间的一层数据管理软件，它提供一个可以方便且高效地存取、管理和控制数据库信息的环境。目前商品化的主流数据库管理系统是关系数据库管理系统 (Relational DataBase Management System, RDBMS)。常见的 DBMS 有 Oracle、SQL Server、DB2、Informix、Sybase、PostgreSQL 和 MySQL 等。Access 是一个基于 Windows 的桌面关系数据库管理系统。

DBMS 的主要功能如下。

1) 数据定义、组织、存储和管理功能

DBMS 为用户提供了数据定义语言 (Data Definition Language, DDL)，方便对数据库中数据对象的组成和结构进行定义。DBMS 要确定以何种文件结构和存取方式在存储级别上组织数据，以提高存取效率。

2) 数据操纵功能

DBMS 还为用户提供了数据操纵语言 (Data Manipulation Language, DML) 以操纵数据，实现对数据库的查询、插入、删除和修改等操作。

DDL 和 DML 不是两种分离的语言，它们简单地构成了单一的数据库语言的不同部分，例如，关系数据库的标准语言 SQL (Structured Query Language)。

3) 数据库的建立和维护功能

数据库的建立、运用和维护由 DBMS 统一管理和控制，以保证事务 (transaction) 的正确运行、数据的安全与完整、多用户对数据的并发使用、发生故障后的系统恢复等。

对于用户而言，事务是具有完整逻辑意义的数据库操作序列的集合。对于 DBMS 而言，事务是一个读写操作序列，这些操作要么都执行，要么都不执行，是一个不可分割的逻辑工作单元。为保证事务能安全并发执行，事务具有原子性 (atomic)、一致性 (consistency)、隔离性 (isolation) 和持久性 (durability) 四个特性，统称为 ACID 特性。在关系数据库中，一个事务可以是一条或一组 SQL 语句。

5. 数据库系统 (database system, DBS)

DBS 是由数据库 (DB)、数据库管理系统 (DBMS)、数据库应用系统和用户组成的存储、管理、处理和维持数据的系统。DBS 的组成如图 1-3 所示。

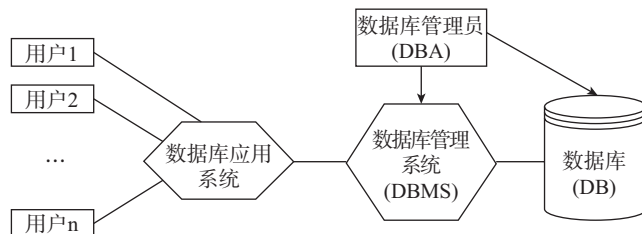


图 1-3 数据库系统的组成

DBS 中的用户有终端用户、应用程序员、系统分析员、数据库设计人员和数据库管理员等多种，分别负责不同的任务。

- **终端用户**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、报表书写和图形显示等。
- **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。
- **系统分析员**负责应用系统的需求分析和规范说明，确定系统的软硬件配置，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、数据库中各级模式的设计。
- **数据库管理员 (DataBase Administrator, DBA)**负责全面管理和控制数据库系统，其主要职责是：决定数据库中的信息内容和结构、决定数据库的存储结构和存取策略、定义数据的安全性要求和完整性约束条件、监控数据库的使用和运行、推动数据库的改进和重组重构等。

DBS 的主要特点如下。

1) 数据的共享性高、冗余度低

共享是指数据库中的相关数据能够被多个不同的用户使用，这些用户可以存取同一种数据并将它用于不同的目的。

数据共享可以大大减少数据冗余，节约存储空间，还能避免数据之间的不相容性和不一致性。

数据冗余是指同一数据在数据库中存储了多个副本，它可能引起如下问题。

- **冗余存储**：相同数据被重复存储，导致浪费大量存储空间。
- **更新异常**：若重复数据的一个副本被修改，则所有副本都必须同时进行同样的修改。因此在更新数据时，为了维护数据库的完整性，系统要付出很大的代价，否则有可能造成数据不一致的后果。
- **删除异常**：删除某些数据时可能丢失其他与之有关联的数据。
- **插入异常**：只有当一些数据事先已存放在数据库中时，另一些数据才能存入该数据库中。

2) 数据独立性高

数据独立性是数据与程序间的互不依赖性，即数据库中的数据独立于应用程序而不依赖于应用程序。

数据独立性一般分为逻辑独立性和物理独立性两种。

逻辑独立性是指用户的应用程序与数据库的逻辑结构是相互独立的。当数据的逻辑结构改变时，用户的应用程序可以不变。

物理独立性是指用户的应用程序与数据库中数据的物理存储是相互独立的。DBMS 负责管理数据库中数据的存储，用户编写应用程序时只需要处理数据的逻辑结构，当数据的物理存储改变时，用户的应用程序不用改变。

3) 数据有安全性保护、完整性检查

数据的安全性是指保护数据以防止非法使用造成的数据泄密和破坏。每个用户只能按规定对某些数据以某些方式进行使用和处理。

数据的完整性是指数据的正确性、有效性和相容性。完整性检查将数据控制在有效的范围内，并保证数据之间满足一定的关系。

4) 并发控制

当多个用户的并发进程同时存取、修改数据库时，DBS 可以对这些用户的并发操作加以控制和协调，以避免相互干扰。

5) 数据库恢复

当计算机系统的硬软件故障、用户的失误或故意破坏造成数据库中的数据错误或丢失时，DBMS 能够将数据库从错误状态恢复到某种已知的正确状态。

1.1.3 数据库系统的结构

可以从多种不同的角度考察数据库系统的结构，以下从两个不同角度进行讨论。

1. 从数据库应用开发者角度

数据库系统通常采用“外模式 - 模式 - 内模式”三级模式结构，如图 1-4 所示。这是数据库系统内部的系统结构。

数据库系统的三级模式是数据的三个抽象级别，它把数据的具体组织交给 DBMS 管理，使用户不必知晓数据在计算机中的具体存储方式和表示方式，就能逻辑地、抽象地处理数据。为了在数据库系统内部实现这三个抽象层次的联系和转换，**DBMS 在三级模式之间提供了两层映像：外模式 / 模式映像和模式 / 内模式映像。**这两层映像保证了数据库系统中的数据能够具有较高的逻辑独立性和物理独立性。

1) 外模式 (external schema)

外模式又称用户模式或子模式，是数据库用户能够看见并使用的局部数据的逻辑结构和特征的描述，是与某一应用有关的数据的逻辑表示。

外模式是保证数据库安全的一个有力措施。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。

一个数据库可以有多个外模式。由于它是各个用户的数据视图，如果不同的用户在使用需求、看待数据的方式、对数据保密的要求等方面存在差异，则其外模式的描述就不

同。模式中的同一个数据在外模式中的结构、类型、长度、保密级别等都可以不同。此外，同一个外模式也可以为某一用户的多个应用系统所使用，但一个应用程序只能使用一个外模式。

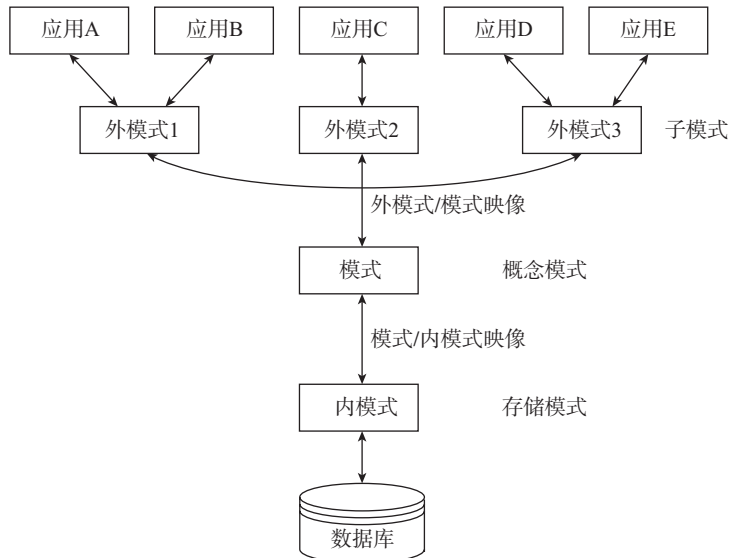


图 1-4 数据库系统的三级模式结构

2) 模式 (schema)

模式又称逻辑模式或概念模式，是数据库中全体数据的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模式。

模式以某一种数据模型为基础，统一综合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。定义模式时不仅要定义数据的逻辑结构，还要定义数据之间的联系，定义与数据有关的安全性和完整性要求。

3) 内模式 (internal schema)

内模式又称存储模式，是数据物理结构和存储方式的描述，是数据在数据库内部的组织方式。一个数据库只有一个内模式。

说明：内模式处于最底层，它反映了数据在计算机物理结构中的实际存储形式；概念模式处于中间层，它反映了设计者的数据全局逻辑要求；外模式处于最外层，它反映了用户对数据的要求。

2. 从终端用户的使用角度

数据库系统的体系结构主要有以下几种。

1) 单用户结构

数据库系统运行于一台计算机中，只支持一个用户访问。

2) 分布式结构

数据库系统运行在分布式计算机系统中，全局数据库的数据可分割存储在系统的多台数据库服务器上，由统一的分布式 DBMS 进行管理，在逻辑上是一个整体。

3) 客户 / 服务器 (Client/Server, C/S) 结构

在 C/S 结构的系统中, 应用程序分为客户端和服务端两大部分。客户端部分为每个用户所专有, 而服务器端部分则由多个用户共享其数据与功能。客户端程序的任务是完成数据预处理、数据表示、管理用户接口和报告请求等; 服务器端计算机安装 DBMS, 接收客户端程序提出的服务请求, 完成 DBMS 的核心功能并将操作结果传递给客户端。这种于 20 世纪 80 年代出现的 C/S 结构比较适合于规模小、用户数少于 100、单一数据库且有安全性和快速性保障的局域网环境下运行。

4) 浏览器 / 服务器 (Browser/Server, B/S) 结构

B/S 结构是随着 Internet 兴起的一种网络应用结构。它主要利用了不断成熟的 Web 浏览器技术, 结合浏览器的多种脚本语言和 ActiveX 技术, 通过通用浏览器实现原来需要复杂专用软件才能实现的强大功能。在这种结构下, 用户工作界面通过 Web 浏览器来实现, 除极少部分的事务逻辑在浏览器端实现外, 大部分事务逻辑在服务器端实现。

也可将 C/S 结构和 B/S 结构结合起来, 形成一种新的结构。在这种结构中, 对于企业外部客户或者一些需要用 Web 处理的、满足大多数访问者请求的功能界面采用 B/S 结构; 对于企业内部少数人使用的功能应用采用 C/S 结构。

1.1.4 国产数据库

1. 国产数据库的发展历程

1) 起步阶段 (1978—2000 年)

20 世纪 70 年代末至 80 年代初: 改革开放初期, 国际上数据库技术已有一定发展, 而中国的数据库技术尚处于萌芽阶段, 科技工作者努力吸收国外先进知识。

20 世纪 80 年代中期至 90 年代初: 国家 863 计划启动, 为数据库技术发展注入了动力, 一批科研人员投身研究, 主要集中在关系型数据库理论和技术上, 清华大学、中国科学院等高校和科研院所成为主要研究基地, 在有限条件下进行了大量自主研发探索。

20 世纪 90 年代中期至 21 世纪初: 随着中国经济快速发展和信息化建设推进, 数据库技术需求日益迫切, 国产数据库产品如东方通信的 TurboBase、神州数码的 PowerBuilder 等面世, 虽与国际巨头产品在性能和功能上有差距, 但这标志着中国已经具备自主开发数据库系统的能力。

2) 跟踪阶段 (2000—2008 年)

国产数据库在企业化过程中遇到了诸多挑战, 如在夹缝中求生存、与国际强手竞争的压力、研发困难等, 但部分企业如人大金仓、南大通用等开始崭露头角, 通过提供更具性价比的解决方案, 在政府、教育等预算有限的行业逐渐赢得部分市场份额。

3) 追赶阶段 (2008—2014 年)

市场需求倒逼国产化替代方案的加速推进, 国产数据库企业在深度兼容、集群策略、行列融合等方面不断创新和发展, 例如, 在民航客票实时交易系统应用中, 国产数据库取得了一定突破, 其技术和产品逐渐成熟, 部分企业开始在特定领域和行业与国际厂商竞争。

4) 并跑阶段(2014年至今)

国产数据库持续迭代,在分布式数据库、云数据库等领域取得显著进展,如 OceanBase 自研的分布式事务处理技术,成功应用于支付宝等大型互联网应用;华为的 GaussDB 采用全分布式架构,实现了多活数据中心的高可用性。国产数据库冲击高端应用,在金融、电信、政府等关键领域广泛应用并站稳脚跟,从“追随者”向“创新者”转变,如达梦数据库。

2021年,工信部印发《“十四五”信息化和工业化深度融合发展规划》,明确要求加速我国数据库产品的研发和应用推广。

2024年,云原生、人工智能等技术的进步,进一步推动了数据库的智能化与高效化演进。在此背景下,国产数据库在技术创新和市场应用方面取得了显著突破,多个国产厂商凭借技术优势和强劲的市场表现,影响力不断增强,为全球数据库生态注入了更多的“中国力量”。墨天轮平台 2024年12月1日公布中国数据库流行度排行榜,共有227个国产数据库参与排名,与2024年1月份(292个)相比,少了65个,数量下降到2022年5月(229个)的规模。排行榜前十名均为关系型数据库,依次是 OceanBase、PolarDB、GoldenDB、GaussDB、TiDB、金仓数据库、GBASE、达梦数据库、openGauss 和 TDSQL。

2. 国产数据库厂商

1) 蚂蚁科技集团股份有限公司

蚂蚁集团起步于2004年诞生的支付宝,2016年12月至2020年7月公司名称为浙江蚂蚁小微金融服务集团股份有限公司,目前是世界领先的互联网开放平台。

中国特色的互联网发展成为促进国产数据库成长的关键力量。数亿人的网购需求,全球第一的移动支付市场,“双十一”的天量交易额,这些西方公司毫无经验的中国式需求,推动中国互联网企业寻求性能更高、成本更低的数据存储和处理方案。然而,囿于技术实力与软件研发的滞后性,越过数据库这座大山并不容易。在我国数据库市场,以 Oracle、IBM 为代表的国外数据库软件长期处于主导地位,软件服务费用居高不下,关键领域存在信息安全隐患。

自2013年开始,蚂蚁集团研发的 OceanBase 应用于蚂蚁集团内部业务,已覆盖支付宝100%核心链路,支撑会员、交易、支付、账务等全部核心业务,并确保了支付宝连续10年“双十一”交易的稳定运行。目前蚂蚁集团、网商银行的全部核心系统都由 OceanBase 支撑。此外, OceanBase 还支撑了工商银行、常熟农商行、红塔银行、中国人民保险、中华保险、中国石化等400余家机构核心系统的稳定运行。

2019年10月2日,据权威机构国际事务处理性能委员会(Transaction Processing Performance Council, TPC)官网披露,中国蚂蚁金服自主研发的金融级分布式关系数据库 OceanBase,在被誉为“数据库领域世界杯”的 TPC-C 基准测试中,打破了美国公司 Oracle 保持了9年之久的世界纪录,成为首个登顶该榜单的中国数据库产品。

2) 华为技术有限公司

华为创立于1987年,是全球领先的 ICT(Information and Communications Technology)基础设施和智能终端提供商。

2020年6月,华为正式宣布开放 openGauss 数据库源代码,并成立 openGauss 开源社

区, 社区官网 (<http://opengauss.org>) 同步上线。openGauss 采用木兰宽松许可证 (Mulan PSL v2), 允许所有社区参与者对代码进行自由修改、使用和引用。openGauss 社区同时成立了技术委员会, 欢迎所有开发者贡献代码和文档。

全国计算机等级考试 (NCRE, <https://ncre.neea.edu.cn/>) 自 2022 年 9 月起, 新增二级“openGauss 数据库程序设计”考试。

openGauss 内核源自 PostgreSQL, 着重在架构、事务、存储引擎、优化器等方向持续构建竞争力特性, 在 ARM 架构的芯片上深度优化, 并兼容 X86 架构。

openGauss 是一款开源关系型数据库管理系统, 具有多核高性能、全链路安全性、智能运维等企业级特性。相比于其他开源数据库, openGauss 主要有多种存储模式、NUMA(Non Uniform Memory Access) 化内核结构、高可用、AI 能力和高安全性等特点。

openGauss 主要应用场景:

- 物联网数据。在工业监控和远程控制、智慧城市的延展、智能家居、车联网等物联网场景下, 传感监控设备多, 采样率高, 数据存储为追加模型, 操作和分析并重。
- 交易型应用。大并发、大数据量、以联机事务处理为主的交易型应用, 如电商、金融、O2O、电信 CRM/ 计费等, 应用可按需选择不同的主备部署模式。

1.2 数据模型

数据库技术的发展沿着数据模型 (data model) 的主线向前推进, 数据模型是数据库系统的基础, 数据库的类型是依据数据模型来划分的。一艘航模舰艇、一组建筑设计沙盘都是具体的模型, 模型是对现实世界中某个对象特征的模拟和抽象。数据模型也是一种模型, 是对现实世界数据特征的抽象, 是用来描述、组织数据并对数据进行操作的。从构成上看, **数据结构、数据操作与数据的约束条件是数据模型的三要素**。其中, 数据结构用于刻画数据、数据语义, 以及数据与数据之间的联系; 数据操作规定了数据的添加、删除、更新、查找、显示、维护、打印、选择、排序等操作; 数据约束是对数据结构和数据操作的一致性、完整性约束, 也称为数据完整性约束。

1.2.1 数据模型的分层

数据模型应满足三方面要求: 一是比较真实地模拟现实世界; 二是容易被人理解; 三是便于在计算机上实现。一种数据模型要同时完美地满足这三方面的要求很困难, 因此, 在数据库系统中要针对不同的使用对象和应用目的, 采用不同的数据模型。

由于计算机不可能直接处理现实世界中的具体事物, 人们必须事先把具体事物转换成计算机能够处理的数据。为了把现实世界中的具体事物抽象、组织为某一个数据库管理系统支持的数据模型, 人们常常**首先将现实世界抽象为信息世界, 再将信息世界转换为机器世界**。

数据的加工是一个逐步转化的过程, 经历了现实世界、信息世界和机器世界这三个不

同的世界。首先将现实世界中的客观对象抽象为某一种信息结构,这种信息结构不依赖于具体的计算机系统,不是某一个 DBMS 支持的数据模型,而是概念级的模型;然后再将概念模型转换为计算机上某一个 DBMS 支持的数据模型。此过程如图 1-5 所示。信息世界又称为概念世界,机器世界又称为数据世界或计算机世界。

从现实世界到概念模型的转换由数据库设计人员完成;从概念模型到逻辑模型的转换可由数据库设计人员完成,也可由数据库设计工具协助设计人员来完成;从逻辑模型到物理模型的转换一般由 DBMS 完成。

根据数据抽象的不同级别,可以将数据模型分为三层:概念数据模型、逻辑数据模型和物理数据模型。

1. 概念数据模型 (Conceptual Data Model, CDM)

概念层次的数据模型称为**概念数据模型**,简称为**概念模型**或**信息模型**,是按用户的观点或认识对现实世界的数据和信息进行建模,主要用于数据库设计。概念模型具有语义表达能力强、易于理解、独立于任何 DBMS、容易向 DBMS 所支持的逻辑数据库模型转换的特点。

常用的概念模型表示方法是**实体-联系模型 (Entity-Relationship model, E-R 模型)**,E-R 模型认为现实世界由一组称为实体的基本对象及这些对象间的联系构成。实体是现实世界中可区别于其他对象的一件“事情”或一个“物体”。例如,大学教务管理系统中的一个学院、一个专业、一门课程、一个学生、一个教师、一条选课记录等都是实体。E-R 模型的详细介绍参见 1.2.2 节。

2. 逻辑数据模型 (Logical Data Model, LDM)

逻辑层是数据抽象的第二层抽象,用于描述数据库数据整体逻辑结构。该层的数据抽象称为**逻辑数据模型**,简称为**逻辑模型**或**数据模型**。在上下文语境没有歧义时,逻辑数据模型常常简称为数据模型。

逻辑模型是用户通过 DBMS 看到的现实世界,是按计算机系统的观点对数据建模,即数据的计算机实现形式,主要用于 DBMS 的实现。因此,**逻辑模型既要考虑用户容易理解,又要考虑便于 DBMS 的实现。**

不同的 DBMS 提供不同的逻辑数据模型,例如层次模型、网状模型、关系模型、面向对象模型、对象关系模型、半结构化模型等,其中**层次模型和网状模型统称为格式化模型。**

格式化模型的数据库系统在 20 世纪 70 年代至 20 世纪 80 年代初十分流行,占据数据库系统产品的主导地位。层次型数据库系统和网状数据库系统在使用和实现上都要涉及数据库物理层的复杂结构,现在除美国和欧洲某些国家有一些早期的层次型数据库系统或网状数据库系统还在继续使用外,其他地方已很少见。

1) 层次模型 (hierarchical model)

层次模型是 DBMS 中最早出现的数据模型,层次型数据库管理系统采用层次模型作为

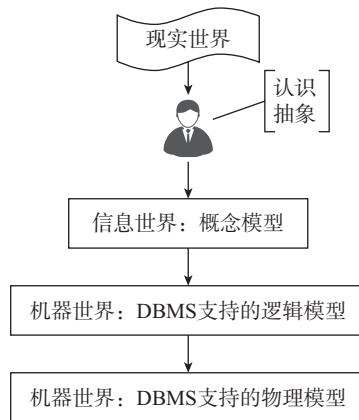


图 1-5 现实世界中客观对象的抽象过程

数据的组织方式，典型代表是 1968 年美国 IBM 公司推出的第一个大型商用数据库管理系统 IMS。

Internet 域名系统 (Domain Name System, DNS) 是一个层状数据库的集合，用于将基于字符的 Internet 域名翻译成用数字表示的 IP(Internet Protocol) 地址。DNS 数据库中的数据由全球成千上万台计算机组成的网络提供，DNS 数据库被称为分布式数据库。

层次模型用树状结构来表示各类实体及实体之间的联系，实体用记录表示，实体间的联系用链接指针表示。层次模型要满足以下条件：

- 有且仅有一个根结点，此根结点没有双亲结点。
- 根结点以外的其他结点有且只有一个双亲结点。

现实世界中许多实体之间的联系本来就呈现出一种层次关系，例如部门组织结构、家族关系等。层次模型的数据结构比较简单清晰，每个结点表示一条记录型，记录型之间的联系用结点之间的连线表示，这种联系是父子之间的一对多联系。但在现实世界中很多联系是非层次的，层次模型表示这类联系时只能通过引入冗余数据（易产生不一致性）或创建非自然的虚拟结点来解决，对插入和删除操作的限制较多，应用程序的编写比较复杂。

2) 网状模型 (network model)

网状数据库管理系统采用网状模型作为数据的组织方式。20 世纪 70 年代，美国数据库系统语言研究会 (CODASYL) 下属的数据库任务组 (DBTG) 提出一个网状数据模型方案，其基本概念、方法和技术具有普遍意义。典型的网状数据库管理系统有 Cullinet Software 公司的 IDMS、Honeywell 公司的 IDS/2、Univac 公司的 DMS1100、HP 公司的 LIMAGE 等。

网状模型用网状结构来表示各类实体及实体之间的联系，网状模型是满足以下条件的基本层次联系的集合：

- 允许一个以上的结点无双亲。
- 一个结点可以有多个双亲。

网状模型是一种比层次模型更具有普遍性的结构，能更直观地描述现实世界。与层次模型一样，网状模型中的每个结点也表示一个记录型，每个记录型可包含若干个字段，结点之间的有向连线表示记录型之间的一对多父子联系。网状模型的双亲结点与孩子结点之间的联系不是唯一的，因此要为每个联系命名，并指出与该联系有关的双亲记录和孩子记录。网状模型结构比较复杂，而且随着应用规模的扩大，数据库的结构会变得越来越复杂。此外，网状模型的操作语言也比较复杂。

3) 关系模型 (relational model)

关系数据库管理系统采用关系模型作为数据的组织方式。关系模型的数据之间的联系是通过存取路径（即指针）实现的，应用程序在访问数据时必须选择适当的存取路径，因此编程人员要了解系统结构的细节，这加重了编写应用程序的负担。

关系模型用规范化的二维表来表示各类实体及实体之间的联系，其详细介绍参见 1.2.3 节。

4) 面向对象模型 (object oriented model)

面向对象数据库管理系统采用面向对象模型作为数据的组织方式。面向对象模型也称为 OO 模型，是用面向对象的观点来描述现实世界实体对象的逻辑组织、对象间限制、联

系等的模型。面向对象的方法和技术在计算机各领域，包括程序设计语言、软件工程、信息系统设计、计算机硬件设计等都产生了深远的影响，同时也促进了面向对象数据模型的发展。对象是由一组数据结构，以及在这组数据结构上操作的程序代码封装起来的基本单位。对象通常与实际领域的实体对应，因此，OO模型也可以看成E-R模型在增加了封装、方法和对象标识符等概念后的扩展。

面向对象数据库系统对数据的操纵包括数据查询、增加、删除、修改等，也具有并发控制、故障恢复、存储管理等完整的功能。其不仅能支持传统数据库应用，也能支持非传统领域的应用，例如CAD(Computer Aided Design)、CAM(Computer Aided Manufacturing)、CIMS(Computer/contemporary Integrated Manufacturing System)、GIS(Geographic Information System)，以及图形、图像等多媒体领域、工程领域和数据集成等领域。虽然面向对象数据库拥有众多优势，但在推广过程中仍面临挑战。一方面，其操作语言相对复杂，对广大用户和开发人员的技术要求较高，这限制了其认可度；另一方面，面向对象数据库在发展中试图全面取代关系数据库管理系统，这一策略大幅提升了企业系统升级的成本和难度，致使许多客户对此望而却步。尽管如此，市场上仍存在成功的面向对象数据库产品，如Versant，它在特定专业领域赢得了用户的青睐，通过精准满足特定场景下的数据管理需求，实现了良好的应用落地。

5) 对象关系模型 (object relational model)

对象关系数据库管理系统采用对象关系模型作为数据的组织方式。对象关系数据库系统(Object Relational DataBase System, ORDBS)是关系数据库与面向对象数据库的结合。它保持了关系数据库系统的非过程化数据存取方式和数据独立性，继承了关系数据库系统已有的技术，既支持原有的数据管理，又支持OO模型和对象管理。各数据库厂商都在原来的产品基础上进行了扩展。1999年发布的SQL标准，也称SQL99，增加了SQL/Object Language Binding，提供了面向对象的功能标准。SQL99对ORDBS标准的制定滞后于实际系统的实现。因此，各个ORDBS产品在支持对象模型方面虽然思想一致，但所采用的术语、语言语法、扩展的功能都不尽相同。

6) 半结构化模型 (semi-structured model)

随着互联网的迅速发展，Web上各种半结构化、非结构化数据源已成为重要的信息来源，产生了以XML为代表的半结构化数据模型和非结构化数据模型。通过Internet使用XML元素的形式传递消息以实现共享数据已成为普遍现象，传统关系数据库应用中使用XML的现象也愈加普及。

半结构化数据是指不符合关系型数据库或其他数据表的形式关联起来的数据模型结构，但包含相关标记，用来分隔语义元素，并对记录和字段进行分层。**非结构化数据**是指不具有预定义模型或未以预定义方式组织的数据。

XML是一种描述性的标记语言，被设计用来传输和存储数据，它具有自我描述性，标签没有被预定义，需要用户自行定义标签。XML是纯文本，有能力处理纯文本的软件都可以编辑XML，但是只有能够读懂XML的应用程序可以有针对性地处理XML的标签。标签的功能性意义依赖于应用程序的特性。

XML简化了数据的传输和平台的变更。很多Internet语言是通过XML创建的，例如，

用于描述可用 Web 服务的 WSDL、用于 RSS feed 的语言 RSS、用于描述资源和本体的 RDF 和 OWL 等。

一个 XML 文档由标记和内容组成。XML 中共有 6 种标记：元素 (element)、属性 (attribute)、实体引用 (entity reference)、注释 (comment)、处理指令 (processing instruction) 和 CDATA 区段 (CDATA section)。XML 的语法规则简单，易学易用。下面列举 5 种基本语法规则。

(1) 所有 XML 元素都须有结束标签，属性值须加引号。

例如，`<p>This is a paragraph</p>`，`<p>` 和 `</p>` 要成对出现，`<p>` 是开始标签，`</p>` 是结束标签。

XML 的声明不属于 XML 本身的组成部分，不是 XML 元素，不需要关闭标签。XML 声明形式如下：

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
```

此声明表示 XML 的当前版本是 1.0，使用 UTF-8 字符集，独立文件声明的属性值是 yes。yes 表示所有与文件相关的信息都已经包含在文件中，即文件中没有指定外部的实体，也没有使用外部的模式。

XML 元素指的是从 (且包括) 开始标签直到 (且包括) 结束标签的部分。元素可包含其他元素、文本或者两者的混合物。元素也可以拥有属性。

(2) XML 标签对大小写敏感。

在 XML 中，标签 `<Book>` 与标签 `<book>` 是不同的。

(3) XML 必须正确嵌套。

在 XML 中，所有元素都不能交叉嵌套。例如：

```
<b><i>This text is bold and italic</i></b>
```

由于 `<i>` 元素是在 `` 元素内打开的，它必须在 `` 元素内关闭。

(4) XML 文档必须有根元素。

XML 文档必须有一个元素是所有其他元素的父元素，该元素称为根元素。例如：

```
<root>
<child>
<subchild>J K. Rowling</subchild>
</child>
</root>
```

`<root>` 是根元素。

(5) XML 的注释格式如下：

```
<!-- This is a comment -->
```

相对于关系型数据存储模式，通过 XML 存储数据有以下优势：

- XML 标签型的数据格式易于人们理解。
- XML 层次型的数据表达，更能反映出对象和业务的实际层次关系。
- XML 灵活的数据存储方式，更能反映业务的变化，能够存储相对更广泛的数据。

因此，在数据建模的时候，使用 XML 能够保证数据模型的扩展能力。

在数据模型设计时，通常有两种使用 XML 的方式：完全 XML 的数据模型设计和部分 XML 的数据模型设计。

完全 XML 的数据模型设计虽然简化了很多数据模型的工作，但是要求开发人员必须熟悉 XML 的 Xquery 语言，完全抛弃已有的 SQL 规范，这给现有的技术体系的延续性增加了难度。完全 XML 的数据模型设计节省了模型设计的工作时间，但是现有的一些开发工具还不能完全支持 XML 的技术，因此，当需要手工进行一些开发工作时，会增加开发工作量。

关系模型和 XML 模型相结合的部分 XML 的数据模型设计方式，通过关系模型延续现有的体系架构，通过 XML 模型提升现有数据模型的扩展能力，兼顾了关系模型和 XML 模型的优点，发挥了两者的长处，规避了两者的不足，在实际数据模型设计中常常采用。

XML 模型将一个 XML 文档建模为一棵树，把文档中的每一个元素、属性、命名空间、处理指令和注释等内容都建模为这棵树中的一个结点，它是非线性的树结构。

3. 物理数据模型 (Physical Data Model, PDM)

物理层是数据抽象的最底层，用于描述数据的物理存储结构和存取方法。例如，数据的物理记录格式是变长还是定长；数据是压缩还是非压缩；一个数据库中的数据和索引 (index) 存放在相同的还是不同的数据段上等。索引提供了对包含特定值的数据项的快速访问，建立索引是加快查询速度的有效手段。索引占用一定的存储空间，当基本表更新时，索引需要进行相应的维护，这就增加了数据库的负担，因此要根据实际应用的需求有选择地创建索引。

物理层的数据抽象称为物理数据模型，简称为物理模型，它不但由 DBMS 的设计决定，而且与操作系统、计算机硬件密切相关。物理模型的具体实现是 DBMS 的任务，数据库设计人员要了解 and 选择物理模型，一般用户不必考虑物理层的细节。

1.2.2 概念模型的代表方法

概念模型是对信息世界建模，概念模型的代表方法有很多，常用的是 P.P.S.Chen 于 1976 年提出的实体 - 联系方法 (又称 E-R 方法)。该方法用 E-R 图 (E-R diagram) 来描述概念模型，E-R 方法也称 E-R 模型、E-R 图或实体 - 联系图。E-R 图是一种语义模型，是现实世界到信息世界的事物及事物之间关系的抽象表示。

E-R 图是不受任何 DBMS 约束的面向用户的表达方法，能够直观表示现实世界中的客观实体、属性及实体之间的联系。构成 E-R 图的基本要素是：实体型、属性和联系。实体型用矩形表示，矩形框内写实体名；属性用椭圆形表示，椭圆形框内写属性名，并用直线与相应的实体型连接；联系用菱形表示，菱形框内写联系名，并用直线分别与有关实体型相连，同时在直线旁标上联系的类型 (1:1、1:n 或 m:n)。

1. 基本概念

1) 实体 (entity)

实体是客观世界中可区别于其他事物的“事物”或“对象”。实体既可以是无形的实在的事物，例如一个学生、一本书、一张身份证；也可以是抽象的概念上存在的事物，例如一个专业、一门课程、一次选课、一次借书或还书、一次网络教学平台的在线学习等。

2) 实体集 (entity set)

实体集是指具有相同类型及相同性质或属性的实体集合。例如，全体学生的集合可定义为学生实体集，学生实体集中的每个实体具有学号、姓名、性别、出生日期等属性。

3) 实体型 (entity type)

具有相同属性的实体必然具有共同的特征和性质。用实体名及其属性名集合来抽象和刻画同类实体，称为实体型。实体型是实体集中每个实体所具有的共同性质的集合。例如，学生 {学号, 姓名, 性别, ...} 就是一个实体型。实体是实体型的一个实例，在含义明确的情况下，实体与实体型通常互换使用。

4) 属性 (attribute)

实体是通过一组属性来描述的，属性是实体集中的每个实体都具有的特征描述。在一个实体集中，所有实体都具有相同的属性。例如，学生实体集中的每个实体都具有学号、姓名、性别、出生日期等属性，如图 1-6 所示。选课实体集中的每个实体都具有学号、课程编号、课程名称、学分、学期、课程类型、成绩等属性，如图 1-7 所示。

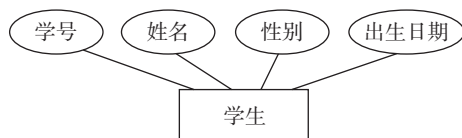


图 1-6 学生实体及其属性

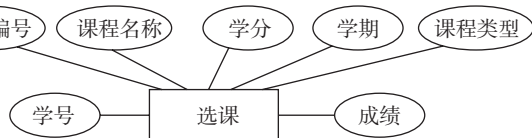


图 1-7 选课实体及其属性

5) 码 (key)

能唯一标识实体的属性或属性集称为码或键。在图 1-6 中，学号可以作为学生实体的码，而学生姓名由于可能重名，不能作为学生实体的码。在图 1-7 中，每个学生每学期可以选修多门课程，因此学号不是选课实体的码，学号和课程编号这两个属性合在一起才能作为选课实体的码。

6) 域 (domain/value set)

每个属性都有自己的取值范围，一个属性所允许的取值范围或集合称为该属性的域，实体的属性值是数据库中存储的主要数据。例如，“姓名”属性的域是字符串集合，“性别”属性的域值是“男”或“女”，“学号”属性的域可以是字母和数字的组合。

2. 实体间的联系 (relationship)

正如现实世界的事物之间存在着联系一样，实体之间也存在着联系。联系是指多个实体之间的相互关联。两个实体间的联系可分为一对一、一对多和多对多 3 种联系类型，如图 1-8 所示。

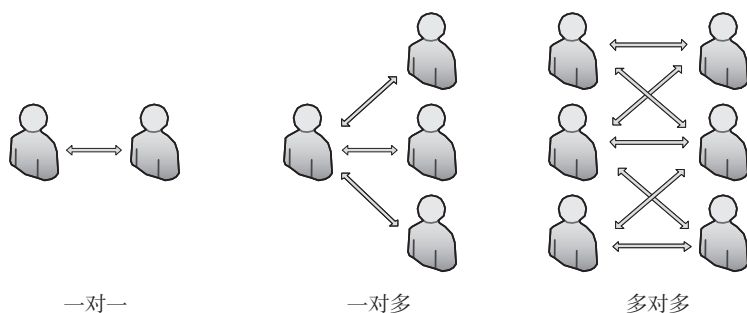


图 1-8 实体间的 3 种联系

1) 一对一联系 (1:1)

若对于实体集 A 中的每一个实体，在实体集 B 中至多有一个（也可以没有）实体与之联系，反之亦然，则称实体集 A 与实体集 B 具有一对一联系，记为 1:1。

例如，学校里一个班级只有一个正班长，而一个正班长只在一个班级里任职，班级与正班长之间具有一对一联系。电影院中座位实体集和观众实体集之间具有一对一联系，因为一个座位最多坐一名观众或没有观众，而一名观众也只能坐在一个座位上。

2) 一对多联系 (1:n)

若对于实体集 A 中的每一个实体，在实体集 B 中有 n 个实体 ($n \geq 2$) 与之联系，反之，对于实体集 B 中的每一个实体，在实体集 A 中至多只有一个实体与之联系，则称实体集 A 与实体集 B 具有一对多联系，记为 1:n。实体集 A 是 1 端，实体集 B 是 n 端。

例如，一所学校有若干名学生，而每个学生只能在一所学校注册，学校与学生之间具有一对多联系，即学生与学校具有多对一联系。如图 1-9 所示，学校是 1 端，学生是 n 端。



图 1-9 一对多联系示例

3) 多对多联系 (m:n)

若对于实体集 A 中的每一个实体，在实体集 B 中有 n 个实体 ($n \geq 2$) 与之联系，反之，对于实体集 B 中的每一个实体，在实体集 A 中也有 m 个实体 ($m \geq 2$) 与之联系，则称实体集 A 与实体集 B 具有多对多联系，记为 m:n。

例如，公司生产的产品与其客户之间是多对多联系，因为一个产品可以被多个客户订购，一个客户也可以订购多个产品；又如，一门课程同时有若干个学生选修，而一个学生可以同时选修多门课程，则课程与学生之间具有多对多联系；一名教师可以讲授多门课程，一门课程可以由多位教师讲授，则教师与课程之间具有多对多联系，如图 1-10 所示。

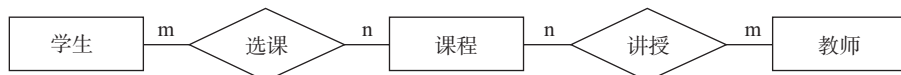


图 1-10 多对多联系示例

【例 1-1】用 E-R 图表示某个工厂物资管理的概念模型。

物资管理涉及以下 5 个实体。

- 仓库：有仓库号、面积、电话号码等属性。
- 零件：有零件号、名称、规格、单价等属性。
- 供应商：有供应商号、姓名、地址、账号、电话号码等属性。
- 项目：有项目号、预算、开工日期等属性。
- 职工：有职工号、姓名、性别、工龄、职务等属性。

这些实体之间有如下联系：

(1) 一个仓库可以存放多种零件，一种零件可以存放在多个仓库中，因此仓库和零件具有多对多联系。用库存量来表示某种零件在某个仓库中的数量。

(2) 一个仓库有多个职工当仓库保管员，一个职工只能在一个仓库工作，因此仓库和职工之间是一对多联系。

(3) 职工之间具有领导与被领导关系，即仓库主任领导若干仓库保管员，因此职工实体之间具有一对多联系。

(4) 一个供应商可以供给若干项目多种零件，每个项目可能使用不同供应商供应的零件，每种零件可由不同供应商供给，因此供应商、项目和零件三者之间具有多对多联系。

工厂物资管理 E-R 图如图 1-11 所示。“职工”包括仓库主任和仓库保管员两类人员，具有一对多联系。

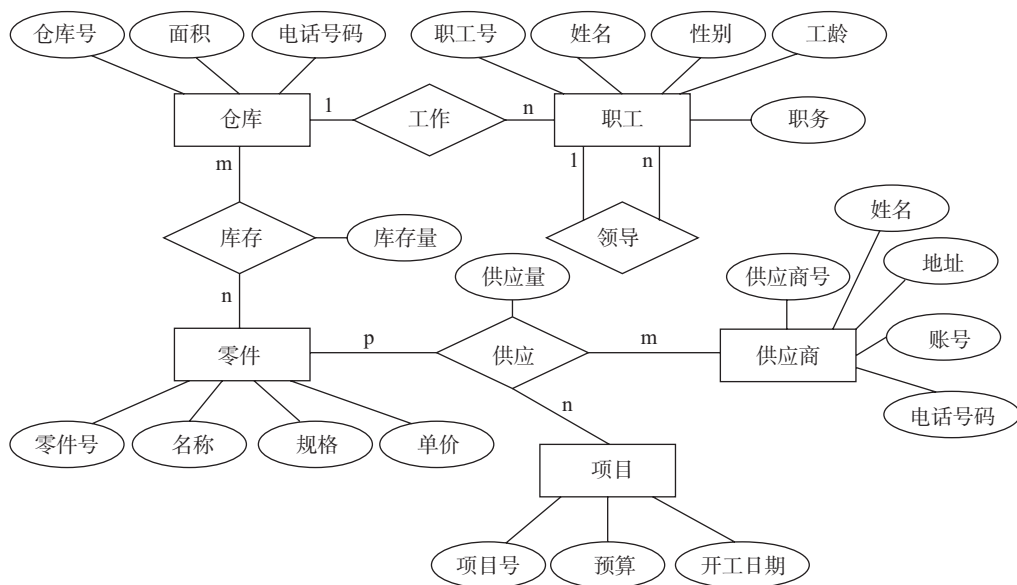


图 1-11 工厂物资管理 E-R 图

3. 三个不同世界的术语对照

现实世界、信息世界 (概念世界) 和机器世界 (数据世界) 的术语对照表如表 1-3 所示。

表 1-3 三个不同世界的术语对照表

现实世界	信息世界 (概念世界)	机器世界 (数据世界)
组织 (事务及其联系)	实体及其联系	数据库
事物类 (总体)	实体集	文件
事物 (对象、个体)	实体	记录
特征	属性	数据项 (字段)

1.2.3 关系模型的基本术语及性质

关系模型是最重要的一种数据模型，关系数据库系统采用关系模型作为数据的组织方式。20 世纪 80 年代以来，计算机厂商新推出的数据库管理系统几乎都支持关系模型，数据库领域当前的研究工作也都是以关系方法为基础。

关系模型是一种用二维表表示实体集，用主码标识实体，用外码表示实体之间联系的数据模型。

1. 关系模型的基本术语

(1) **关系**：对应通常所说的二维表，它由行和列组成，还必须满足一定的规范条件。如图 1-12 中的表 1-4 和表 1-5 就是两个关系。

(2) **关系名**：每个关系的名称。图 1-12 中，表 1-4 的关系名是 **Stu**，表 1-5 的关系名是 **Grade**。

(3) **元组**：二维表中的每一行称为关系的一个元组，它对应于实体集中的一个实体。

(4) **属性**：二维表中的每一列对应于实体的一个属性，每个属性要有一个属性名。

(5) **值域**：每个属性的取值范围。**关系的每个属性都必须对应一个值域**，不同属性的值域可以相同，也可以不同。

例如，用“男”或“女”表示性别的取值范围；用大于或等于 0 且小于或等于 100 的实数可以表示百分制成绩的取值范围，也可以表示其他某种属性的取值范围。

空值用 null 表示，是所有可能的域的一个取值，表示值“未知”或“不存在”或“无意义”。例如，某学生的成绩属性值为空值 null，表示不知道该学生的成绩；或该学生没有参加考试，因而没有获得成绩；或不想让他人知道该学生的成绩等。

(6) **分量**：元组中的一个属性值。

(7) **候选码**：**如果关系中的某一属性组的值能唯一标识一个元组，则称该属性组为候选码。**例如，学生实体的学号和身份证号都可以唯一标识一个元组，学号和身份证号就是候选码。

(8) **主码**：也称主键或关键字。**如果一个关系有多个候选码，则选定其中一个为主码。**例如，学号和身份证号是学生实体的候选码，可以选定学号作为主码。主码也可以是多个属性的组合。**按照关系的完整性规则，主码不能取空值 null。**

(9) **外码**：也称外键或外部关键字。为了实现表与表之间的联系，通常将一个表的主码作为数据之间联系的纽带放到另一个表中，这个起联系作用的属性称为外码。例如，在图 1-12 中的表 1-4(Stu 表) 和表 1-5(Grade 表) 中，利用公共属性“学号”实现这两个表的联系，这个公共属性是 Stu 表的主码、Grade 表的外码。Grade 表的主码是“学号”和“课程编号”这两个属性的组合。



图 1-12 利用公共属性实现两表间的联系

(10) 关系模式：对关系的描述，一般表示为

关系名(属性1, 属性2, …, 属性n)

【例1-2】Stu关系和Grade关系的关系模式如下：

Stu(学号, 姓名, 性别, 是否团员, 出生日期, 生源地, 专业编号)

Grade(学号, 课程编号, 平时成绩, 期末成绩)

其中，Stu 关系模式中带下画线的属性“学号”是主码；Grade 关系模式中带下画线的两个属性“学号”和“课程编号”一起作为主码。

关系模式要求关系必须是规范化的，即要求关系必须满足一定的规范条件，这些规范条件中最基本的一条就是：关系的每一个分量必须是一个不可分的数据项，不允许表中还嵌套有表。

例如，在图 1-13 所示的机动车驾驶证申请条件汇总表中，身体条件和年龄条件是可分的数据项，又分为身高、视力和申请年龄、允许年龄，因此图 1-13 所示的表不符合关系模型要求，这个汇总表不是关系。

	是否 初学	身体条件		年龄条件		增驾条件	可否在暂 住地申请
		身高 (cm)	视力	申请 年龄	允许 年龄	驾驶经历及记分情况	
A1	否	155	5.0	26~50	26~60	B1、B2 五年以上且前三个周期内无满分记录 A2 两年以上且前一个周期内无满分记录 无死亡事故中负主要以上责任的记录	不可
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

图 1-13 机动车驾驶证申请条件汇总表(表中有表)的非规范化实例

关系和现实生活中的表格所使用的术语不同，术语对照如表 1-6 所示。

表 1-6 术语对照

关系术语	生活中的表格术语
关系名	表名
关系模式	表头 (表格的描述)
关系	二维表
元组	记录或行
属性	列
属性名	列名
属性值	列值
分量	一条记录中的一个列值
非规范关系	表中有表 (大表中嵌套小表)

2. 关系模型的性质

关系是一张二维表，但并不是所有的二维表都是关系。关系建立在严格的数学理论基础之上，应具有如下性质。

(1) 元组个数有限性：关系中元组个数是有限的。

(2) 元组的唯一性：关系中每个元组代表一个实体，因此各元组均不相同。

(3) 元组的次序无关性：关系中元组与次序无关，可以任意交换。

(4) 元组分量的原子性：关系中元组的分量是不可分割的基本数据项。关系中的每个属性的值域必须是原子的、不可分解的。若域中的每个值都被看作不可再分的单元，则称域是原子的。例如，表示属性“出生日期”的值域是由所有形如“year/month/day”的值构成，其中 year 是由 4 位数字构成的字符串，表示年份；month 是由两位数字构成的字符串，表示月份；day 是由两位数字构成的字符串，表示日期。将 year、month、day 看作一个整体，则出生日期的值域是原子的。

(5) 属性名唯一性：一个关系中的属性名要各不相同。

(6) 属性的次序无关性：关系中属性与次序无关，可以任意交换。

(7) 分量值域的统一性：关系中各列的属性值取自同一个域，因此一列中的各个分量具有相同性质。

关系模型优化的详细介绍参见 1.3.3 节。

3. 关系模型的优缺点

关系模型具有下列优点。

(1) 关系模型与格式化模型不同，它建立在严格的数学概念的基础之上。

(2) 关系模型的概念单一。无论实体还是实体之间的联系都用关系来表示。对数据的检索和更新结果也是关系，即二维表。关系的结构简单、清晰，用户易懂易用。

(3) 关系模型的存取路径对用户透明，因而具有更高的数据独立性、更好的安全保密性，简化了程序员的工作和数据库开发建立的工作。

当然，关系模型也有缺点，例如，由于存取路径对用户是透明的，查询效率往往不如格式化数据模型。为了提高性能，DBMS 必须对用户的查询请求进行优化，这样就增加了

开发 DBMS 的难度。不过，数据库用户不必考虑这些系统内部的优化技术细节。

1.2.4 关系运算

关系运算是关系数据库的数据操纵，主要用于关系数据库的查询操作。关系模型中常用的关系操作包括查询操作和插入、删除、修改操作两大部分。查询是关系操作中最主要的部分。查询操作可分为并、差、交、笛卡尔积、选择、投影、连接、除等，其中并、差、笛卡尔积、选择、投影是 5 种基本操作，其他操作可以用基本操作来定义和导出，就像乘法可以用加法来定义和导出一样。

关系代数是一种抽象的查询语言，它用对关系的运算来表达查询。关系代数的运算对象是关系，运算结果也是关系。关系代数用到的运算符有集合运算符和专门运算符两类，按照运算符的不同，关系代数的运算可分为传统的集合运算和专门的关系运算两类。

1. 传统的集合运算

传统的集合运算包括并、差、交和笛卡尔积 4 种运算。由于笛卡尔积的元素是元组，这里的笛卡尔积是指广义的笛卡尔积。

设关系 R 和关系 S 具有相同的目 n (即两个关系都有 n 个属性)，且相应的属性取自同一个域，t 是元组变量， $t \in R$ 表示 t 是 R 的一个元组。

1) 并 (union)

关系 R 和关系 S 的并记作： $R \cup S = \{t | t \in R \vee t \in S\}$ 。

其结果仍为 n 目关系，由属于 R 或属于 S 的元组组成。

2) 差 (except/difference)

关系 R 和关系 S 的差记作： $R - S = \{t | t \in R \wedge t \notin S\}$ 。

其结果仍为 n 目关系，由属于 R 但不属于 S 的所有元组组成。

3) 交 (intersection)

关系 R 和关系 S 的交记作： $R \cap S = \{t | t \in R \wedge t \in S\}$ 。

其结果仍为 n 目关系，由既属于 R 又属于 S 的元组组成。关系的交可以用差来表示，即 $R \cap S = R - (R - S)$ 。

4) 广义笛卡尔积 (extended cartesian product)

两个分别为 n 目和 m 目的关系 R 和关系 S 的笛卡尔积是一个 (n+m) 列的元组的集合。元组的前 n 列是关系 R 的一个元组，后 m 列是关系 S 的一个元组。若 R 有 k1 个元组，S 有 k2 个元组，则关系 R 和关系 S 的笛卡尔积有 k1×k2 个元组。记作： $R \times S = \{trts | tr \in R \wedge ts \in S\}$ 。

2. 专门的关系运算

1) 选择 (select)

选择运算是根据给定的条件，从一个关系中选出一个或多个元组 (表中的行)。被选出的元组组成一个新的关系，这个新的关系是原关系的一个子集。例如，表 1-7 就是从图 1-12 中的表 1-4 所示关系中选取性别为“女”的记录而组成的新关系。

表 1-7 选择运算

学号	姓名	性别	是否团员	出生日期	生源地	专业编号
S1701002	张晓兰	女	FALSE	2005/07/11	云南	M01
S1702001	马丽林	女	TRUE	2006/11/06	湖南	M02

2) 投影 (project)

投影运算是从一个关系中选择某些特定的属性 (表中的列) 重新排列组成一个新关系。投影之后属性减少, 新关系中可能有一些行具有相同的值, 若有这种情况, 重复的行将被删除。例如, 表 1-8 就是从表 1-7 所示关系中选取部分属性而得到的新关系。

表 1-8 投影运算

学号	姓名	性别	生源地
S1701002	张晓兰	女	云南
S1702001	马丽林	女	湖南

3) 连接 (join)

连接运算是从两个或多个关系中选取属性间满足一定条件的元组, 组成一个新的关系。等值连接 (equijoin) 和自然连接 (natural join) 是最为重要也最为常用的连接。例如, 表 1-9 就是将图 1-12 中的表 1-4 和表 1-5 按学号进行自然连接而生成的新关系。

在连接运算中, 按照字段值对应相等为条件进行的连接操作称为等值连接。自然连接是去掉重复属性的等值连接。

表 1-9 自然连接运算

学号	姓名	性别	是否团员	出生日期	生源地	专业编号	课程编号	平时成绩	期末成绩
S1701001	陈榕刚	男	TRUE	2004/03/12	福建	M01	C0101	76.00	80.00
S1701001	陈榕刚	男	TRUE	2004/03/12	福建	M01	C0102	82.50	86.00
S1701002	张晓兰	女	FALSE	2005/07/11	云南	M01	C0101	85.00	75.00
S1701002	张晓兰	女	FALSE	2005/07/11	云南	M01	C0102	90.00	93.00

1.2.5 关系的完整性

关系模型的完整性规则是对关系的某种约束条件。实体及其联系要受到现实世界中许多语义要求的约束, 例如, 24 小时制表示的整点时间取值只能在 [0, 23] 区间; 百分制成绩的取值只能在 [0, 100] 区间; 一个学生一个学期可以选修多门课程, 但只能在本学期已开设的课程中进行选修; 学生在选修一门课程所开教学班时, 所有选修该教学班的学生人数之和不能超过该教学班所安排教室的容量等。

为了维护数据库中数据与现实世界的一致性, 关系数据库的数据与更新操作要遵循三类完整性规则: 实体完整性、参照完整性和用户自定义完整性。其中实体完整性和参照完整性是所有关系模型必须满足的数据完整性约束, 被称作关系的两个不变性, 由关系数据库系统自动支持。用户自定义完整性是应用领域需要遵循的数据完整性约束, 体现了具体

应用领域中的数据语义约束。

1. 实体完整性 (entity integrity)

若属性集 (指一个或多个属性) A 是关系 R 的主码, 则 A 不能取空值 $null$ 。

现实世界中的实体都是可区分的, 即它们具有某种唯一性标识; 而一个关系对应于现实世界的一个实体集, 关系中的每一个元组对应于一个实体。因此, 作为唯一区分不同元组的主码属性集不能取空值。若主码的属性取空值, 就说明存在某个不可标识的实体, 即存在不可区分的实体, 这是不允许的。

如果主码是由若干个属性的集合构成, 则要求构成主码的每一个属性的值都不能取空值。例如, 图 1-12 中的表 1-5 所示 $Grade$ 关系中的主码“学号”和“课程编号”都不能取 $null$ 。

【例 1-3】 Stu 关系的主码是“学号”, 因此它在任何时候的取值都不能为空值 $null$, 但其他属性: 姓名、性别、是否团员、出生日期、生源地、专业编号等都可以取空值, 表示当时该属性的值未知或不存在或无意义。如果不知道某个学生的出生日期, 可以将该属性值输入为 $null$, 表示未知; 如果规定学生从大二开始选择专业, 那么新生的专业编号暂时输入为 $null$, 表示不存在, 待学生大二选择专业后再将 $null$ 更新为所选专业的编号。

2. 参照完整性 (referential integrity)

现实世界的实体之间存在各种联系, 而在关系模型中实体及实体间的联系都用关系来描述。因此, 实体间的联系也就对应于关系与关系之间的联系。

若关系 R 的外码 F 参照关系 S 的主码, 则对于关系 R 中的每一个元组在属性 F 上的取值, 要么为空值 $null$, 要么等于关系 S 中某个元组的主码值。

参照完整性反映了“主码”属性与“外码”属性之间的引用规则。

【例 1-4】 $Grade$ 关系和 Stu 关系之间存在着属性之间的引用, 即 $Grade$ 关系引用了 Stu 关系的主码“学号”, 显然, $Grade$ 关系中的外码“学号”属性的取值必须存在于 Stu 关系中。

数据库的修改会导致参照完整性的破坏。当参照完整性约束被违反时, 通常拒绝执行导致完整性被破坏的操作。

3. 用户自定义完整性 (user-defined integrity)

任何关系数据库系统都应该支持实体完整性和参照完整性, 这是关系模型所要求的。除此之外, 不同的关系数据库系统根据其应用环境的不同, 往往还要满足一些特殊的约束条件。用户自定义完整性就是针对不同应用领域的语义, 由用户自己定义的一些完整性约束条件。例如, 课程成绩若是等级制, 可以自定义成绩为优秀、良好、中等、及格和不及格五个等级; 在 Stu 关系中, 若按照应用的要求学生不能没有姓名, 则可以定义学生姓名不能取空值; 学生的出生日期不能晚于当前日期, 需要按标准的年/月/日格式设置。

1.3 数据库设计

数据库设计 (database design) 广义地讲是数据库及其应用系统的设计, 即设计整个数据库应用系统; 本节讨论狭义的数据库设计, 即设计数据库的各级模式并建立数据库, 这是数据库应用系统设计的一部分。数据库设计是指对于一个给定的应用环境, 构造优化的数据库逻辑模式和物理结构, 并据此建立数据库及其应用系统, 使之能够有效地存储和管理数据, 满足各种用户的应用需求, 包括信息管理要求和数据操作要求。信息管理要求是指在数据库中应该存储和管理哪些数据对象; 数据操作要求是指对数据对象需要进行哪些操作, 如查询、添加、删除、修改、统计等操作。

大型数据库的设计和开发是一项庞大的工程, 涉及多学科的综合技术。数据库应用系统从设计、实施到运维的全过程和一般的软件系统设计、开发和运维有许多相似之处, 但更有其自身的一些特点。“三分技术, 七分管理, 十二分基础数据”是数据库设计的特点之一, 因此数据的收集、整理、组织和不断更新是数据库建设中的重要环节。

早期数据库设计主要采用手工和经验相结合的方法, 设计质量与设计人员的经验和水平有直接关系。缺乏科学理论和工程方法支持的数据库设计, 质量难以保证。常常是数据库运行一段时间后会发现各种问题, 需要进行修改甚至重新设计, 增加了系统维护代价。为此, 人们相继提出了各种数据库设计方法。例如, 新奥尔良方法、基于 E-R 模型的设计方法、第三范式设计方法、面向对象的数据库设计方法、统一建模语言 UML(Unified Modeling Language) 方法等。

实践表明, 数据库设计是一项软件工程, 开发过程遵循软件工程的一般原理和方法。数据库设计包括以下 6 个阶段: 需求分析→概念结构设计→逻辑结构设计→物理结构设计→数据库实施→数据库运行和维护。设计一个完善的数据库应用系统不可能一蹴而就, 它往往是这 6 个阶段的不断反复。

在数据库设计过程中, 需求分析和概念结构设计可以独立于任何 DBMS 进行, 逻辑结构设计和物理结构与选用的 DBMS 密切相关。若所设计的数据库应用系统比较复杂, 则可借助数据库设计工具以提高数据库设计质量并减少设计工作量。例如, 支持 60 多种 RDBMS, 并提供 Eclipse 插件的 Power Designer; MySQL 数据库设计专用工具 Navicat for MySQL、MySQL Workbench 等。

1.3.1 数据库系统的需求分析

需求分析是整个设计过程的基础, 是最困难和最耗时的一步。需求分析的结果是否准确反映用户的实际要求, 将直接影响后面各阶段的设计, 并影响到设计结果是否合理和实用。

设计人员要不断深入地与用户交流, 逐步确定用户的实际需求, 与用户达成共识, 然后分析和表达这些需求, 形成需求分析报告, 即需求说明书。需求分析报告必须交给用户

确认，用户认可之后才能开始下阶段的概念结构设计。

数据字典是进行详细的数据收集和数据分析所获得的主要成果。它是关于数据库中数据的描述，即元数据，而不是数据本身。数据字典在需求分析阶段建立，在数据库设计过程中不断修改、充实和完善。

【例 1-5】简要分析大学教务管理系统的功能需求。

大学教务管理系统的设计目标是对高校的学院、专业、课程、教师、学生、学生成绩等进行信息化管理，以方便用户使用并能提高工作效率。该系统的基本要求是采用 Access 关系数据库管理系统对教务信息进行管理，要考虑数据库的完整性要求，保证数据的一致性；要能够方便快捷地查询到相关的教务信息：各学院教务信息、专业信息、课程信息、教师基本信息、学生基本信息、选课成绩等，并且能够对这些信息进行增加、删除、修改、统计分析、打印存档等操作。

1.3.2 概念结构设计

将需求分析得到的用户需求抽象为概念模型的过程就是概念结构设计。

概念模型是各种数据模型的共同基础，它比数据模型更独立于机器、更抽象，从而更加稳定。描述概念模型的常用工具是 E-R 图。

E-R 图是数据库设计中广泛使用的数据建模工具。它所表示的概念模型与具体的 DBMS 所支持的数据模型相独立，是各种数据模型的共同基础。

概念结构设计时，要对各种需求分而治之，即先分别考虑各个用户的需求，形成局部的概念模型，也称局部 E-R 图，再根据实体间联系的类型，将它们综合为一个全局的结构。全局 E-R 图要支持所有局部 E-R 图，能合理地抽象出一个完整的信息世界的结构，即概念模型。

概念模型是对用户需求的客观反映，不涉及具体的计算机软硬件。因此，在概念结构设计阶段只需要关注怎样表达出用户对信息的需求，不需要考虑具体的实现问题。

【例 1-6】按照例 1-5 中的需求分析，设计大学教务管理系统全局 E-R 图。

大学教务管理系统全局 E-R 图如图 1-14 所示，有学院、专业、课程、教师、学生 5 个实体；实体之间通过联系相关联，联系的命名要反映联系的语义，通常采用动词命名。联系本身也可以产生属性，如“选课”联系有“成绩”属性。

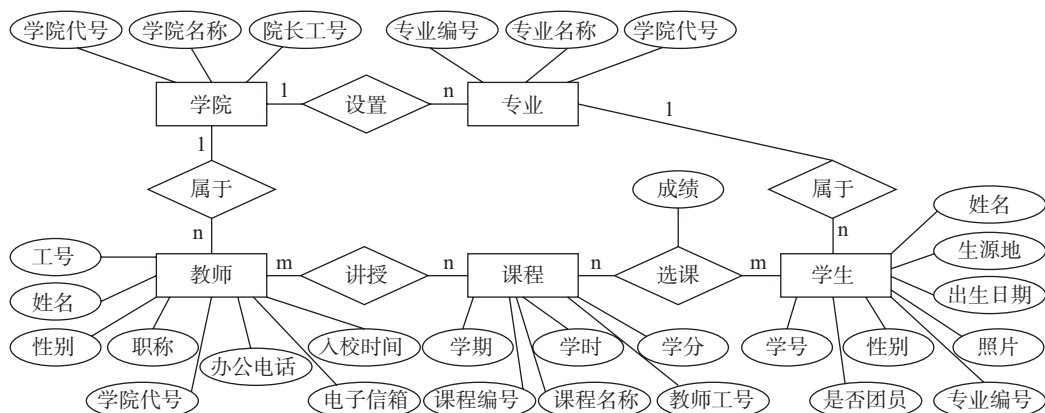


图 1-14 大学教务管理系统全局 E-R 图

1.3.3 逻辑结构设计

概念结构是独立于任何一种数据模型的信息世界的结构，逻辑结构设计任务是把概念结构设计阶段得到的 E-R 图转换为逻辑结构，这个逻辑结构要与选用的 DBMS 产品的数据模型相符合。当前的数据库应用系统大都采用支持关系数据模型的 RDBMS，以下只讨论 E-R 图向关系数据模型的转换原则和方法。

1. E-R 图向关系模型的转换

E-R 图由实体型、实体的属性和实体型之间的联系 3 个要素组成，因此将 E-R 图转换为关系模型要解决的问题是：如何将实体型和实体型之间的联系转换为关系模式，如何确定这些关系模式的属性和码。

转换遵循的原则是：

- 一个实体型转换为一个关系模式，实体的属性就是关系的属性，实体的码就是关系的码。
- 实体型之间不同类型联系的转换规则如下。

(1) 一个 1:1 联系可以转换为一个独立的关系模式，也可以与任意一端对应的关系模式合并。如果转换为一个独立的关系模式，则与该联系相连的各实体的码及联系本身的属性均转换为关系的属性，每个实体的码均是该关系的候选码。如果与某一端实体对应的关系模式合并，则需要在该关系模式的属性中加入另一个关系模式的码和联系本身的属性。

(2) 一个 1:n 联系可以转换为一个独立的关系模式，也可以与 n 端对应的关系模式合并。如果转换为一个独立的关系模式，则与该联系相连的各实体的码及联系本身的属性均转换为关系的属性，而关系的码为 n 端实体的码。

(3) 一个 m:n 联系转换为一个关系模式，与该联系相连的各实体的码及联系本身的属性均转换为关系的属性。各实体的码组成关系的码或关系码的一部分。

(4) 三个或三个以上实体间的一个多元联系可以转换为一个关系模式。与该多元联系相连的各实体的码及联系本身的属性均转换为关系的属性，各实体的码组成关系的码或关系码的一部分。

(5) 具有相同码的关系模式可以合并。

【例 1-7】按照 E-R 图向关系模型的转换规则，将例 1-6 中的大学教务管理系统 E-R 图转换成关系模式，在 Access 中实现。以下关系模式的主码用下划线标出，外码用斜体表示。

学院 (学院代号, 学院名称, 院长工号, ...)

专业 (专业编号, 专业名称, 学院代号, ...)

教师 (工号, 姓名, 性别, 入校时间, 职称, 学院代号, 办公电话, 电子信箱, ...)

学生 (学号, 姓名, 性别, 是否团员, 出生日期, 生源地, 专业编号, 照片, ...)

课程 (课程编号, 课程名称, 学期, 学时, 学分, 课程类型, 教师工号, ...)

成绩 (学号, 课程编号, 平时成绩, 期末成绩, ...)

其中，“成绩”关系由“选课”联系转换，学号和课程编号两个属性组成主码，这两个属性也是外码。

2. 关系模型的优化

数据库逻辑设计的结果不是唯一的。为了进一步提高数据库应用系统的性能，还需要依据应用需要适当地修改、调整数据模型的结构，即进行数据模型的优化。关系数据模型的优化通常以规范化理论为指导。一个“好”的关系模型应该是数据冗余尽可能少，且不会发生插入异常、删除异常和更新异常等问题。

设计关系数据库时，关系模式必须满足一定的规范化要求。在关系数据库理论中，这种规则称为范式 NF(Normal Form)。范式是符合某一种级别的关系模式的集合，目前关系数据库有 6 种范式：第一范式 (1NF)、第二范式 (2NF)、第三范式 (3NF)、Boyce-Codd 范式 (BCNF)、第四范式 (4NF) 和第五范式 (5NF)。满足最低要求的范式是第一范式 (1NF)。在第一范式的基础上进一步满足更多要求的称为第二范式 (2NF)，其余范式以此类推。一般情况下，数据库只需要满足第三范式 (3NF) 即可。

所谓第几范式原本是表示关系的某一种级别，所以常称某一关系模式 R 为第几范式。现在则把范式这个概念理解成符合某一种级别的关系模式的集合，即 R 为第几范式就可以写成 $R \in xNF$ 。

对于各种范式之间的关系有 $5NF \subset 4NF \subset BCNF \subset 3NF \subset 2NF \subset 1NF$ 成立。一个低一级范式的关系模式通过模式分解 (schema decomposition) 可以转换为若干个高一级范式的关系模式的集合，这个过程就叫规范化 (normalization)。

1) 函数依赖

在数据库设计中，除实体型之间存在着联系外，在属性之间还存在着一定的依赖关系。由此引入了属性间的函数依赖概念。

定义 1-1 关系中的主码 X 有一取值，随之确定了关系中的非主属性 Y 的值，则称关系中的非主属性 Y 函数依赖于主码 X，或称属性 X 函数决定属性 Y，记作 $X \rightarrow Y$ 。其中 X 称为决定因素，Y 称为被决定因素。

函数依赖又分为非平凡的函数依赖和平凡的函数依赖；从性质上还可以分为完全函数依赖、部分函数依赖和传递函数依赖。

例如，在设计“学生”表时，一个学生的学号能决定学生的姓名，也可称姓名属性依赖于学号。现实生活中，如果知道一个学生的学号，就一定能知道学生的姓名，这种情况就称姓名依赖于学号，记作：学号 \rightarrow 姓名。

2) 第一范式 (1NF)

定义 1-2 如果一个关系模式 R 的所有属性都是不可分的基本数据项，则称 R 属于第一范式的关系模式，记为 $R \in 1NF$ 。

当一个关系中不存在组合数据项和多值数据项，只存在不可分的数据项时，这个关系是规范化的。在关系数据库中，1NF 是对关系模式的基本要求。

【例 1-8】有一个“选课”关系由“学号”和“课程编号”两个属性组成，每个学生可以选择多门课程，如表 1-10 所示，“课程编号”列中出现了多个值的情况，是非规范化的关系。规范化为 1NF 后，如表 1-11 所示。

表 1-10 不满足 1NF 的关系

学号	课程编号
S1701001	C0101, C0102, C0103
S1701002	C0101, C0102

表 1-11 满足 1NF 的关系

学号	课程编号
S1701001	C0101
S1701001	C0102
S1701001	C0103
S1701002	C0101
S1701002	C0102

3) 第二范式 (2NF)

定义 1-3 如果关系模式 $R \in 1NF$, 且 R 中的每一个非主属性都完全函数依赖于主码, 则称 R 属于第二范式的关系模式, 记为 $R \in 2NF$ 。

2NF 要求关系中的非主属性 (不能用作候选码的属性) 完全依赖于主码。所谓完全依赖, 是指不能存在仅依赖主码一部分的属性。如果存在, 则这个属性和主码的这一部分应该分离出来形成一个新关系, 新关系与原关系之间是一对多的联系。

【例 1-9】 以下选课关系模式在实际应用中存在问题, 请将之规范化为 2NF。

选课 (学号, 课程编号, 成绩, 学分)

此选课关系中, 主码为学号和课程编号组合属性。实际应用中存在数据冗余、更新异常、插入异常、删除异常等问题。

(1) 数据冗余: 若同一门课程有 61 个学生选修, 则这门课相同的学分就重复了 60 次。

(2) 更新异常: 若某门课程的学分变化了, 则相应元组的所有学分值都要更新, 有可能遗漏, 出现同一门课程学分不同的错误。

(3) 插入异常: 若计划开设新课, 由于没有学生选修, 没有学生的学号, 则只能等有人选修时才能把课程编号和学分信息存入。

(4) 删除异常: 若学生已经结业, 要从当前数据库删除选修记录, 而此课程新生尚未选修, 则此课程编号和学分信息无法保存。

解决方法: 将原“选课”关系分解成以下两个新关系模式, 即可满足 2NF 的要求。

课程 (课程编号, 学分)

成绩 (学号, 课程编号, 成绩)

4) 第三范式 (3NF)

定义 1-4 如果关系模式 $R \in 2NF$, 且 R 中的每一个非主属性都不传递函数依赖于任何主码, 则称 R 属于第三范式的关系模式, 记为 $R \in 3NF$ 。

所谓传递函数依赖, 是指如果存在 $A \rightarrow B \rightarrow C$ 的决定关系, 则 C 传递函数依赖于 A 。

【例 1-10】 以下学生关系模式在实际应用中存在问题, 请将之规范化为 3NF。

学生 (学号, 姓名, 专业编号, 专业名称, 专业负责人)

此关系中的专业编号、专业名称和专业负责人等信息会重复存入, 有大量的数据冗余; 插入、删除和更新时也将产生数据异常的情况。这是由于关系中存在传递依赖造成的, 即学号 \rightarrow 专业编号、专业编号 \rightarrow 专业负责人, 但“学号”不直接决定非主属性“专业负责人”, 而是通过“专业编号”传递依赖实现的, 不满足 3NF 的要求。

解决方法: 将原“学生”关系分解成以下两个新关系模式, 即可满足 3NF 的要求。

学生(学号, 姓名, 专业编号)

专业(专业编号, 专业名称, 专业负责人)

由以上分析可知, 部分函数依赖和传递函数依赖是产生数据冗余、异常的两个重要原因, 3NF 消除了大部分冗余、异常, 具有较好的性能。

综上所述, 关系数据模型的优化通常以规范化理论为指导, 按照以下方法:

(1) 确定数据依赖。写出每个数据项之间的数据依赖, 按需求分析阶段得到的语义, 分别写出每个关系模式内部各属性之间的数据依赖, 以及不同关系模式属性之间的数据依赖。

(2) 对于各关系模式之间的数据依赖进行极小化处理, 消除冗余的数据联系。

(3) 按照函数依赖理论对关系模式逐一进行分析, 考察是否存在部分函数依赖、传递函数依赖、多值函数依赖等, 确定各关系模式属于第几范式。

(4) 根据需求分析阶段得到的处理要求, 分析在应用环境中这些模式是否合适, 确定是否要对某些模式进行合并或分解。

必须注意, **并不是规范化程度越高的关系就越优**。例如, 当查询经常涉及两个或多个关系模式的属性时, 系统经常进行连接运算, 而连接运算的代价相当高, 关系模型低效的主要原因之一就是连接运算。这时可以考虑将几个关系合并为一个关系。因此在这种情况下, 2NF 甚至 1NF 也许是合适的。对于一个具体应用, 到底规范化到什么程度, 需要权衡响应时间和潜在问题两者的利弊。

(5) 对关系模式进行必要分解, 提高数据操作效率和存储空间利用率。常用的分解方法是水平分解和垂直分解。

水平分解是把基本关系的元组分为若干个子集合, 定义每个子集合为一个子关系, 以提高系统的效率。垂直分解是把关系模式 R 的属性分解为若干子集合, 形成若干子关系模式。

3. 设计用户子模式

将概念模型转换为全局逻辑模型之后, 还应根据局部应用需求, 结合具体关系数据库管理系统的特特点设计用户的外模式。RDBMS 一般都提供了视图概念, 可以利用这一功能设计更符合局部用户需要的用户外模式。

定义数据库全局模式主要从系统的时间效率、空间效率、易维护等角度出发。从数据库系统的三级模式结构(见图 1-4)可知, 由于用户外模式与模式是相对独立的, 在定义用户外模式时可以考虑用户的使用习惯和使用的便捷性。

(1) 使用更符合用户习惯的别名。用视图机制可以在设计用户视图时重新定义某些属性名, 使其与用户习惯一致, 以方便用户使用。

(2) 可以对不同级别的用户定义不同的视图, 以保证系统的安全性。

(3) 简化用户对系统的使用。如果某些局部应用中经常要用到某些复杂的查询, 则可将这些复杂的查询定义为视图, 用户每次只对定义好的视图查询, 这可以极大简化用户的使用。

1.3.4 物理结构设计

数据库在物理设备上的存储结构与存取方法称为数据库的物理结构, 它依赖于选定的 DBMS。数据库的物理结构设计是为一个给定的逻辑数据模型选取一个最适合应用要求的物理结构的过程。

数据库的物理结构设计通常分为两步。

(1) 确定数据库的物理结构，通常关系数据库物理设计的内容主要包括：为关系模式选取存取方法，以及设计关系、索引等数据库文件的物理存储结构。

(2) 对物理结构进行评价，评价的重点是时间效率、空间效率、维护代价和各种用户要求。

评价物理数据库的方法完全依赖于所选用的 RDBMS，主要从定量估算各种方案的存储空间、存取时间和维护代价入手，对估算结果进行权衡、比较，选择出一个较优的、合理的物理结构。

若评价结果满足原设计要求，则可进入物理实施阶段，否则，需要重新设计或修改物理结构，有时甚至要返回逻辑设计阶段修改数据模型。

1.3.5 数据库的实施

完成数据库的物理结构设计之后，设计人员要用 RDBMS 提供的数据库定义语言和其他实用程序将数据库逻辑设计和物理设计结果严格描述出来，成为 RDBMS 可以接受的源代码，再经过调试产生目标模式，然后组织数据入库，并进行试运行。

数据入库十分费时费力。由于一般数据库系统中数据量很大，而数据来源于不同部门，数据的组织方式、结构和格式往往与新设计的数据库系统有差距。组织数据输入要将各类源数据从各个局部应用中抽取出来，输入计算机，再分类转换，最后综合成符合新设计的数据库结构的形式。

在原有系统的数据有一小部分已输入数据库后，就可以开始对数据库系统进行联合调试，即试运行。试运行阶段要实际运行数据库应用程序，执行对数据库的各种操作，测试应用程序的功能是否满足设计要求，若不满足，则要对应用程序进行修改、调整，直到达到设计要求；还要测试系统的性能指标，分析其是否达到设计目标。一般情况下，设计时的考虑在许多方面只是近似评估，与实际系统运行总有一定的差距，因此必须在试运行阶段实际测量和评价系统性能指标。事实上，有些参数的最佳值往往是经过运行调试后找到的。如果测试的结果与设计目标不符，则要返回物理设计阶段重新调整物理结构，修改系统参数，有时甚至要返回逻辑设计阶段修改逻辑结构。

1.3.6 数据库的运行和维护

数据库应用系统试运行合格后，数据库开发工作基本完成，就可以投入正式运行。但是由于应用环境在不断变化，数据库运行过程中物理存储也会不断变化，对数据库设计进行评价、调整、修改等维护工作是一个长期的任务，也是设计工作的继续和提高。

在数据库运行阶段，对数据库经常性的维护工作主要由 DBA 完成。数据库的维护包括：数据库的转储和恢复，数据库的安全性、完整性控制，数据库性能的监督、分析和改造，数据库的重组织与重构造。

随着时间的推移，数据库应用环境会发生变化，例如，增加了新的应用或新的实体，取消了某些应用，有的实体与实体间的联系也发生了变化等，使原有的数据库设计不能满足新的需求，需要调整数据库的模式和内模式等。有时还需要进行数据库的重新构造，但

数据库的重构也是有限的，只能做部分修改。如果应用变化太大，重构也无济于事，则说明此数据库应用系统的生命周期已结束，应该设计新的数据库系统。

1.4 本章小结

“合抱之木，生于毫末；九层之台，起于累土；千里之行，始于足下。”以科学的态度对待科学，以真理的精神追求真理。本章是学习数据库的基础，也是全书的理论基础。诞生于 20 世纪 60 年代末期的数据库技术有坚实的理论基础、成熟的商业产品和广泛的应用领域。本章引用了近年来我国颁布的一系列国家政策和标准，以及多位中外著名数据库专家的观点，遵循《全国计算机等级考试二级 Access 数据库程序设计考试大纲 (2025 年版)》的要求，对数据库基础理论进行了概述。

本章 1.1 节从数据管理技术经历的三个阶段：人工管理阶段→文件系统阶段→数据库系统阶段，引入了与数据库技术密切相关的基本术语：数据、大数据、数据库、数据库管理系统 DBMS 和数据库系统 DBS，介绍了信息、数据仓库、事务、索引、元宇宙等概念，着重说明了数据库系统的三级模式结构——外模式 - 模式 - 内模式，以及两层映像——外模式 / 模式映像和模式 / 内模式映像。本节对国产数据库的发展历程、两个数据厂商和开源数据库 OceanBase 和 openGauss 进行了简要介绍，并向奋斗在国产数据库研发一线的技术人员和立志创造优秀国产数据库产品的企业致敬。

本章 1.2 节从现实世界→信息世界→机器世界入手，按数据模型的分层，讨论了概念数据模型、逻辑数据模型和物理数据模型。逻辑模型中介绍了层次模型、网状模型、关系模型、面向对象模型、对象关系模型和 XML 模型。重点论述了概念模型的表示方法 E-R 图，关系模型的基本术语、性质和关系的三个完整性规则：实体完整性、参照完整性和用户自定义完整性。

本章 1.3 节从开发者的角度，介绍了数据库应用系统设计的 6 个阶段：需求分析→概念结构设计→逻辑结构设计→物理结构设计→数据库实施→数据库运行和维护。设计一个完善的数据库应用系统不可能一蹴而就，它往往是这 6 个阶段的不断反复。

按照章节顺序可总结出 9 个“三”：①数据管理技术经历三个阶段：人工管理、文件系统和数据库系统。②数据库系统的三级模式结构：外模式（又称用户模式或子模式）、模式（又称概念模式或逻辑模式）和内模式（又称存储模式）。③三个世界：现实世界、信息世界（又称概念世界）和数据世界（又称机器世界或计算机世界）。④数据模型的三要素：数据结构、数据操作和数据完整性约束。⑤数据模型分为三层：按数据抽象的不同级别，分为概念数据模型、逻辑数据模型和物理数据模型。⑥ E-R 模型的三个基本要素：实体型（用矩形表示）、属性（用椭圆形表示）和联系（用菱形表示）。⑦两实体间的联系分为三种：一对一 (1:1)、一对多 (1:n) 和多对多 (m:n)。⑧掌握三种关系运算：选择运算、投影运算和连接运算。⑨关系模型的三个完整性规则：实体完整性、参照完整性和用户自定义完整性。

本章不仅阐述了数据库系统的核心理论，也涉及数据库领域的新成果、新标准和应用的新方向，提到多位中外杰出人物的姓名，读者可以自行查阅资料，进一步了解他（她）们的生平。

拓展阅读

坚持守正创新是进一步全面深化改革必须牢牢把握、始终坚守的重大原则。守正和创新是辩证统一的，只有守正才能保证创新始终沿着正确方向前进，只有持续创新才能更好地守正。

我们要以一往无前的胆魄和勇气，顺应时代发展新趋势、实践发展新要求、人民群众新期待，大力推进理论创新、实践创新、制度创新、文化创新，以及其他各方面创新，为中国式现代化提供强大动力和制度保障。

资料来源：2025年1月16日出版的第2期《求是》杂志，中共中央总书记、国家主席、中央军委主席习近平的重要文章《进一步全面深化改革中的几个重大理论和实践问题》。

1.5 思考与练习

1.5.1 选择题

1. 数据库中存储的是（ ）。

A. 数据库应用程序	B. 数据模型
C. 数据库管理系统	D. 数据及数据之间的关系
2. 在数据库系统的三级模式结构中，描述数据物理结构和存储方式的是（ ）。

A. 外模式	B. 概念模式	C. 内模式	D. 关系模式
--------	---------	--------	---------
3. 按照数据抽象的不同级别，数据模型可分为三种模型，它们是（ ）。

A. 小型、中型和大型模型	B. 网状、环状和链状模型
C. 层次、网状和关系模型	D. 概念、逻辑和物理模型
4. E-R 模型适用于建立数据库的（ ）。

A. 概念模型	B. XML 模型	C. 层次模型	D. 物理模型
---------	-----------	---------	---------
5. 在 E-R 模型中，表示属性的图形是（ ）。

A. 菱形	B. 椭圆形	C. 矩形	D. 直线
-------	--------	-------	-------
6. 用二维表来表示实体及实体间联系的数据模型是（ ）。

A. XML 模型	B. 层次模型	C. 关系模型	D. 网状模型
-----------	---------	---------	---------
7. 在关系模型中，（ ）的值能唯一标识一个元组。

A. 分量	B. 索引	C. 外码	D. 主码
-------	-------	-------	-------

8. 下列实体之间存在多对多联系的是()。
- A. 宿舍与学生 B. 学生与课程 C. 病人与病床 D. 公司与职工
9. 一支球队由一名主教练、一名队医和若干球员组成, 则队医和球员是()联系。
- A. 一对一 B. 一对多 C. 多对一 D. 多对多
10. 下列选项不属于关系运算的是()。
- A. 连接 B. 投影 C. 比较 D. 选择
11. openGauss 是一款开源的()数据库管理系统。
- A. 关系 B. 内存 C. 向量 D. 网状
12. ()是国家基础战略性资源和重要生产要素。
- A. 向量 B. 数据 C. 信息 D. AI

1.5.2 填空题

1. 数据库系统 DBS 的核心组成部分是_____, 其英文缩写是_____。
2. 层次模型、网状模型和关系模型的数据结构依次是_____、_____和_____。
3. 关系数据库系统采用_____作为数据的组织方式。
4. 将 E-R 模型转换为关系模式时, 实体和联系都可以表示为_____。
5. _____是指用户的应用程序与数据库中数据的物理存储是相互独立的。
6. 在数据库运行阶段, 对数据库经常性的维护工作主要由_____完成。
7. 数据独立性是指_____和数据的组织结构相互独立的特性。
8. 在关系数据库的基本操作中, 从表中抽取属性值满足条件列的操作称为_____。
9. 实体与实体之间的联系有_____, _____ 和 _____ 3 种。
10. “顾客”与“商品”两个实体集之间的联系一般是_____。
11. 一个关系表的行称为_____或_____。
12. 自然连接是_____的等值连接。
13. 目前蚂蚁集团、网商银行的全部核心系统都由_____数据库支撑。
14. 2021 年 6 月 10 日, 第十三届全国人民代表大会常务委员会第二十九次会议通过《中华人民共和国_____法》, 规范数据处理活动, 保障数据安全, 促进数据开发利用, 保护个人、组织的合法权益, 维护国家主权、安全和发展利益。
15. 我国著名科学家钱学森在 1990 年将虚拟现实技术的元宇宙翻译为_____。
16. 由 2023 年 3 月组建的_____等 17 个部门联合发布《“数据要素 × ”三年行动计划(2024—2026 年)》。

1.5.3 简答题

1. 请简述数据库系统的组成, 并解释各组成部分的作用。
2. 数据库系统的三级模式结构和两层映像是什么? 这两层映像的作用分别是什么?
3. E-R 模型有什么作用? 构成 E-R 模型的基本要素是什么?
4. “一把钥匙开一把锁”中的钥匙和锁是两个实体, 它们之间存在什么联系?
5. 数据模型在数据库设计中起什么作用?

6. 试述关系模型主要术语的含义。
7. 什么是关系的完整性？试举例说明关系的完整性约束条件。
8. 试述数据库应用系统的设计过程。
9. 我国数据管理经历了怎样的发展历程？
10. openGauss 是何企业研发的数据库管理系统？其资源包可以从哪里获取？
11. 本章中提及了哪些杰出人物？试简述中外两位人物的主要贡献。
12. 试用关系模式的规范化理论分析表 1-12 存在的问题，并将此学生成绩信息表分解成符合范式要求的数据库模式。

表 1-12 学生成绩信息表

学号	姓名	专业	课程成绩						
			课程名	学时	学分	学期	任课教师编号	教师姓名	成绩
S001	刘英	SE	DB	64	4	3	T009	何宾先	90
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

第 2 章

“人工智能+” 数据技术

知识目标

1. 了解人工智能的定义、诞生与发展方向。
2. 熟悉人工智能的主要学派与重要分支。
3. 了解国家数据基础设施的内涵与总体技术架构。
4. 了解我国数据库产业现状。
5. 掌握数据库关键技术发展趋势。

素质目标

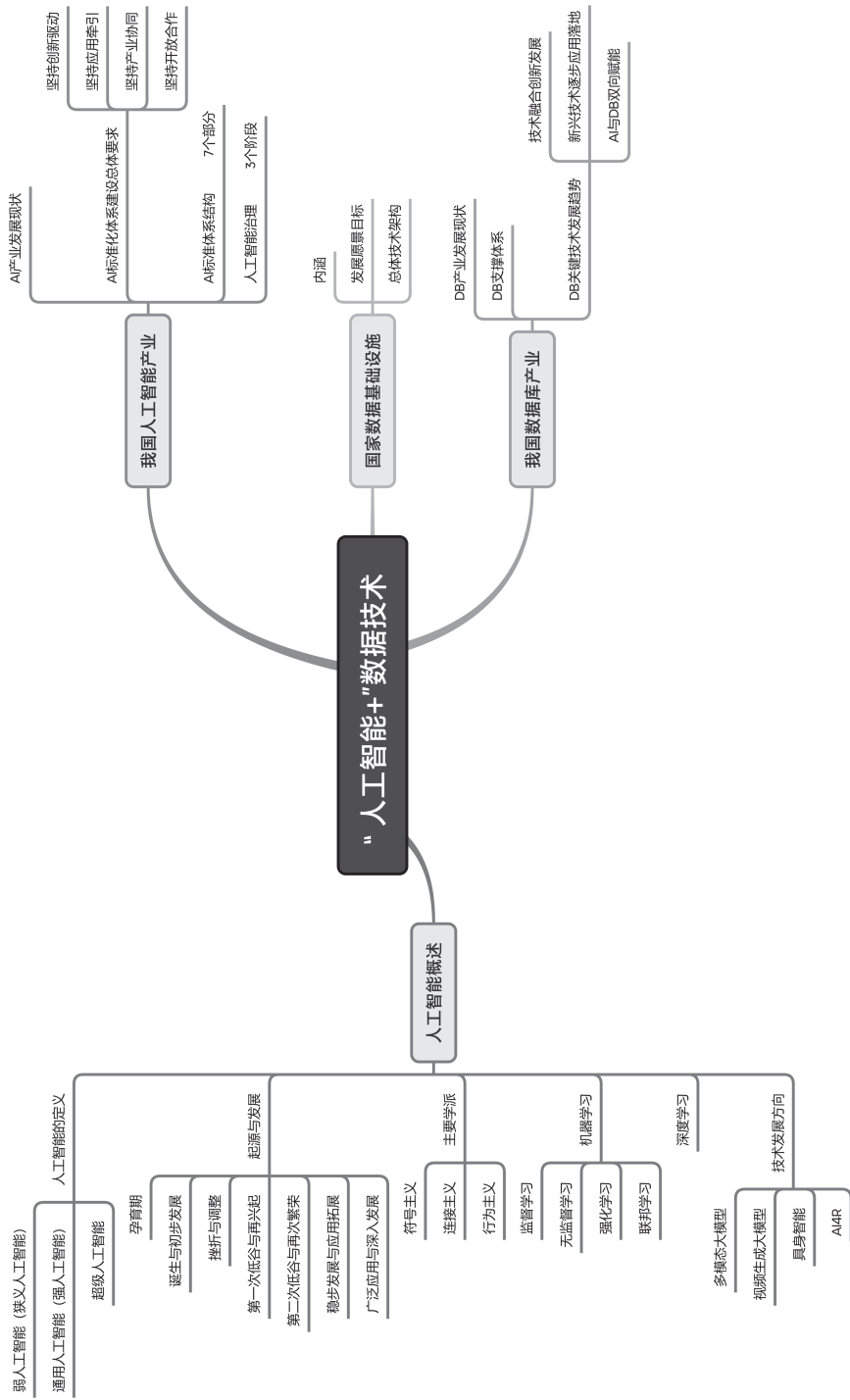
1. 培养学生的全球化视野和社会主义核心价值观。
2. 培养学生的批判性思维和传承中华文化的意识。

学习指南

本章的重点是 2.1 节、2.2 节和 2.4 节，难点是 2.4 节。

本章涉及人工智能和数据库的新技术，新名词较多，建议读者参看思维导图并利用 AI 工具辅助学习，提升学习效率。百度的文心一言、字节跳动的豆包、阿里巴巴的通义千问、华为的盘古、腾讯的混元、科大讯飞的星火认知、智谱 AI 的智谱清言、月之暗面的 Kimi、昆仑万维的天工 AI、360 的纳米搜索等 AI 工具注册后即可使用。本教程所有章节的学习，除了用传统的学习方式，还建议使用以上的国产 AI 工具给学习赋能。

思维导图



为贯彻落实《国务院关于印发新一代人工智能发展规划的通知》(国发〔2017〕35号),2018年教育部制定《高等学校人工智能创新行动计划》,该计划在加强重点领域应用中提及实施“人工智能+”行动。

2024年3月5日,国务院《政府工作报告》明确指出:“深化大数据、人工智能等研发应用,开展‘人工智能+’行动,打造具有国际竞争力的数字产业集群。”这一政策导向为“人工智能+”的发展提供了坚实的基础和广阔的空间。

2024年3月教育部正式启动人工智能赋能教育行动,宣布在国家智慧教育平台(<https://www.smartedu.cn/AIEducation>)上线“AI学习”专栏,启动实施教育系统人工智能大模型应用示范行动。7月,教育部宣布将打造人工智能通识课程体系,赋能理工农医文等各类人才培养。12月,教育部吴岩副部长在2024世界慕课与在线教育大会指出,中国高等教育数字化秉持集成化、智能化、国际化的“3I”理念,坚持应用为王、共建共享、系统集成、能力为重,实现数字技术与教育教学的深度融合,正在推进一场“学习革命”,深化“教学革命”,进而走向“教育革命”,走出一条中国特色高等教育数字化发展道路。2024年末,国家终身教育智慧教育平台(<https://lifelong.smartedu.cn/home>)上线,主页设置人工智能专题。

2025年1月全国科学技术名词审定委员给出“人工智能+”的含义:指将人工智能技术与各行各业深度融合,推动产业转型升级和创新发展的一种理念与实践。它不仅是将人工智能应用于某一特定领域,而是通过技术的集成与创新,实现对传统行业的全面赋能与重构。人工智能作为数字基础设施建设的重要组成部分,是新一轮科技革命和产业革命的核心驱动力。从技术层面来看,“人工智能+”强调人工智能与其他先进技术协同发展,如与物联网、大模型、云计算等结合,形成综合性的技术解决方案。

2.1 人工智能概述

2.1.1 人工智能的定义

人工智能(Artificial Intelligence, AI)自1956年诞生以来,引起了众多学科和不同专业背景的学者、企业家及各国政府的空前重视,已成为一门具有日臻完善的理论基础、日益广泛的应用领域和广泛交叉的前沿学科。像许多新兴学科一样,人工智能至今尚无统一定义。

现行国家标准 GB/T 11457—2006《信息技术 软件工程术语》对人工智能的定义:“计算机科学的一个分支,专门研制执行通常与人的智能有关联的功能(例如,推理、学习和自改进)的数据处理系统;某一设备执行通常与人的智能有关联的功能(例如,推理、学习和自改进)的能力。”GB/T 41867—2022《信息技术 人工智能 术语》中的相关内容:“人工智能系统(artificial intelligence system)是针对人类定义的给定目标,产生诸如内容、预测、

推荐或决策等输出的一类工程系统。”“人工智能计算资源包括中央处理单元 (CPU)、图形处理单元 (GPU)、神经网络处理单元 (NPU)、现场可编程逻辑门阵列 (FPGA)、数字信号处理器 (DSP)、专用集成电路 (ASIC) 等。”

行业标准 YD/T 4994—2024《移动智能终端人工智能应用的个人信息保护技术要求及评估方法》对人工智能的表述：“能以人类智能 (如推理和学习) 相似的方式做出反应的计算机程序, 包括语音识别、图像识别、自然语言处理和专家系统等。”

百度百科的人工智能定义：**研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学**。维基百科的人工智能定义：人工智能是计算机科学的一个分支, 它企图了解智能的实质, 并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能从诞生以来, 理论和技术日益成熟, 应用领域也不断扩大, 可以设想, 未来人工智能带来的科技产品, 将会是人类智慧的“容器”。人工智能能够模拟人的意识和思维过程。人工智能不是人的智能, 但能像人那样思考、也可能超过人的智能。

从宏观层面来看, **人工智能可分为弱人工智能、通用人工智能和超级人工智能三个发展阶段**。

(1) 弱人工智能 (Narrow AI): 也称狭义人工智能, 是指专门针对特定任务的人工智能系统。例如, 语音识别软件、推荐系统、自动驾驶汽车中的某些功能等。这些系统在特定任务上表现出色, 但在其他领域则无法应用。

(2) 通用人工智能 (Artificial General Intelligence, AGI): 也称强人工智能, 是指一种能够像人类一样思考、学习和执行多任务的人工智能系统, 它具有高效的学习和泛化能力, 能根据所处的复杂动态环境自主产生并完成任务, 它具备自主感知、认知、决策、学习、执行和社会协作等能力, 且符合人类情感、伦理和道德观念。全球生成式 AI 领军者 OpenAI 公司将 AGI 写在了自己的企业使命中。

(3) 超级人工智能 (Artificial Superintelligence, ASI): 是指在所有领域都远远超越人类智能水平的人工智能系统。这种系统不仅能够执行人类的所有任务, 还能在这些任务上展现出前所未有的效率和创造力。ASI 目前还属于科幻和理论探讨的范畴。

2.1.2 人工智能的起源与发展

1. 孕育期 (1956 年前)

人类对智能机器和人工智能的梦想和追求由来已久, 可追溯到三千多年前。早在我国西周时期 (公元前 1066—公元前 771 年) 就有关于巧匠偃师献给周穆王一个能歌善舞的机械艺伎的记载。东汉时期 (公元 25 年—220 年) 杜诗发明了以水力为动力的鼓风机“水排”, 用于冶金炉鼓风, 大幅提高了冶炼效率; 同时期张衡发明的指南车, 通过精妙的齿轮传动系统实现定向功能, 被视为最早的机器人雏形。图 2-1 所示为古代指南车复原模型。三国时期 (公元 220 年—280 年) 诸葛亮创造的木牛流马军粮运输装置, 以其独特的机械结构和自主运动特性, 堪称古代陆地军用机器人的杰出代表。这些古代智慧结晶, 展现了中华民族在自动化机械领域的早期探索与卓越成就。

20 世纪初, 乔治·布尔的《思维规律的研究》、德里希·弗雷格的《概念文字》、伯特兰·罗素和阿尔弗雷德·诺斯·怀特海的《数学原理》等著作在数理逻辑研究上取得重大

突破，为人工智能的逻辑推理和符号处理奠定了理论基础。1943年，沃伦·麦卡洛克和沃尔特·皮茨提出人工神经网络的概念并构建人工神经元的MP模型，开创了人工神经网络研究时代。1949年，唐纳德·赫布出版《行为的组织》，提出Hebb学习规则，为机器学习中的神经网络的学习算法奠定基础。1950年，阿兰·麦席森·图灵发表《计算机器与智能》(Computing Machinery and Intelligence)论文，提出“图灵测试”，为判断机器是否具有智能提供了一种方法，被广泛认为是人工智能的开端，论文发表页面如图2-2所示。图灵被誉为人工智能之父。

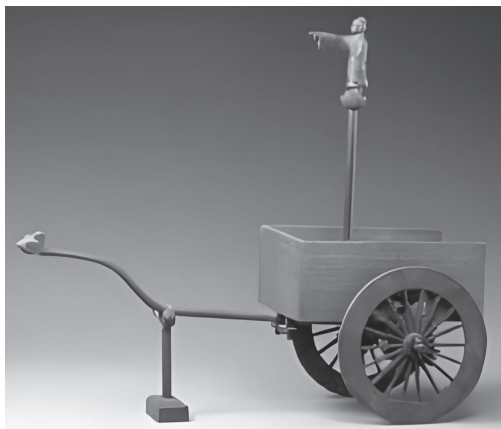


图 2-1 指南车复原模型

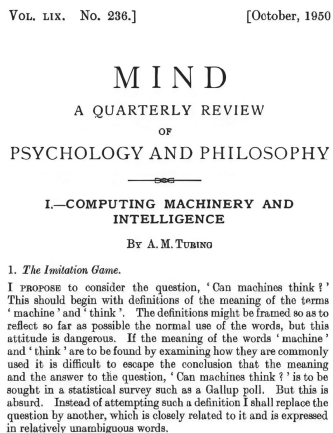


图 2-2 阿兰·图灵的论文《计算机器与智能》

2. 诞生与初步发展 (1956年至20世纪60年代末)

1956年夏，年轻的数学助教约翰·麦卡锡和他的三位朋友马文·明斯基、纳撒尼尔·罗切斯特和克劳德·香农，邀请艾伦·纽厄尔和赫伯特·西蒙等科学家在美国的达特茅斯(Dartmouth)学院组织了一个夏季学术讨论班，历时两个月。参加会议的是在数学、神经生理学、心理学和计算机科学等领域从事教学和研究工作的学者，在会上第一次正式使用了人工智能这一术语，这标志着人工智能学科正式诞生。约翰·麦卡锡通常被认为是人工智能之父，因他首次提出“人工智能”概念，认为人工智能就是为了让机器的行为看起来像是人类所表现出的智能行为一样；他还提出了“通用问题求解器”的概念，希望通过构建一种通用的算法，让计算机能够解决各种类型的问题。此外，他还开发了LISP编程语言，这是世界上最早的用于人工智能研究的编程语言之一，其灵活性使得它非常适合表达复杂的符号和操作，被广泛应用于AI研究，是人工智能发展的基石之一。

亚瑟·塞缪尔研制出具有自学能力的“跳棋程序”，能积累下棋经验、提高棋艺，开拓了“机器博弈”“机器学习”方面的研究。

3. 挫折与调整 (20世纪60年代末至70年代)

人们发现当时的计算机有限的内存和处理速度不足以解决实际的人工智能问题，也难以建立庞大的数据库帮助程序学习。由于研究进展未达预期，英国政府、美国国防部高级研究计划局等机构逐渐停止了对人工智能研究的资助。

尽管如此，这一时期仍有一些重要成果出现，如由斯坦福大学开发的第一个专家系统

DENDRAL 研制成功, 标志着人工智能学科的新分支“专家系统”诞生; 1972 年开始研制的医疗专家系统 MYCIN, 主要用于辅助诊断和治疗细菌感染性疾病, 它包含知识库(存储医学知识, 如病原菌特性、药物抗菌谱等)、推理机(用于根据输入的患者症状等信息进行推理判断)等专家系统的典型组件该系统为其他专家系统的研究与开发提供了范例和经验。

4. 第一次低谷与再次兴起(20 世纪 70 年代至 80 年代末)

20 世纪 70 年代初开始, 人工智能研究进入第一次寒冬, 科学活动和商业活动衰退, 持续近 20 年。

20 世纪 80 年代, 卡耐基梅隆大学制造出可应用于工业领域的专家系统, 企业和大学纷纷参与开发, 世界 500 强企业中近一半都研制或使用了专家系统。同时, 人工智能数学模型方面取得重大突破, 1986 年的多层神经网络和 BP 反向传播算法, 推动了神经网络技术的复兴。

5. 第二次低谷与再次繁荣(20 世纪 80 年代末至 90 年代末)

1987 年至 1993 年, 苹果公司、IBM 推广的第一代台式机费用远低于专家系统的软硬件开销, 专家系统实用性局限于特定情景, 美国国防部高级研究计划局调整拨款方向, 人工智能研究进入第二次寒冬。

1997 年, IBM 公司的“深蓝”电脑与国际象棋世界冠军卡斯帕罗夫对战并获得胜利, 这是首个电脑系统在标准比赛时限内击败国际象棋世界冠军的事件, 展示了人工智能在复杂任务处理上的能力。

6. 稳步发展与应用拓展(2000 年至 2022 年)

21 世纪初, 卷积神经网络(Convolutional Neural Network, CNN)在图像识别任务上表现出色, 推动了深度学习的兴起, 为自动驾驶、医疗影像等应用奠定了基础。2011 年, IBM 公司的 Watson 在美国智力问答节目上击败两位人类冠军, 展现了人工智能在自然语言处理和知识问答方面的能力。

人工智能在一些特定领域的应用逐渐成熟, 如机器人技术的发展。美国 iRobot 公司推出了能避开障碍并自动设计行进路线的吸尘器机器人 Roomba。

2012 年 AlexNet 深度神经网络在 ImageNet 图像分类挑战赛上取得巨大成功, 展示了神经网络在大型数据集上的强大表现, 开启了深度学习的快速发展时代。2013 年, Facebook(现 Meta)、Google、百度等科技公司纷纷成立人工智能实验室或收购相关公司, 加大在深度学习领域的投入。

2014 年聊天程序“尤金·古斯特曼”首次通过图灵测试。2015 年 Google 开源了第二代机器学习平台 TensorFlow。2016 年 Google 人工智能机器 AlphaGo 与世界围棋冠军李世石对弈, 以 4:1 的总比分取得胜利, 之后 AlphaGo 升级款又战胜了我国棋手柯洁, 进一步证明了人工智能在复杂博弈领域的能力。

2017 年 Google 推出 Transformer 模型, 为自然语言处理等领域带来重大变革。2018 年 OpenAI 公司开始推出以海量参数和强大生成能力著称的 GPT 系列模型, 2020 年的 GPT-3 有 1750 亿参数, 在自然语言理解、生成和总结等方面的表现接近人类水平。2022 年生成式人工智能(Generative AI)崛起, 国内外掀起了一场大语言模型(Large Language

Model, LLM) 浪潮，大语言模型的介绍与应用参见本书第 10 章。

7. 广泛应用与深入发展 (2023 年至今)

2023 年大模型正式进入开源商用阶段，生成式预训练模型的应用成为人们日常生活中的热门工具，人工智能大规模应用元年到来。2024 年几乎所有重要的模型供应商都发布了多模态模型，人工智能正处于广泛应用与深入发展阶段，同时也在面临伦理、法律和社会影响等方面的挑战。未来，AI 的发展可能会更加注重可解释性、安全性，以及与人类的和谐共生。

2.1.3 人工智能的主要学派

目前人工智能的主要学术流派包括符号主义、连接主义和行为主义三种。这些学派各自拥有独特的理论基础、应用场景和技术手段，共同推动了人工智能技术的发展。三大主流学派的比较如表 2-1 所示。

表 2-1 人工智能三大主流学派的比较

比较内容	符号主义	连接主义	行为主义
别称	逻辑主义、心理学派或计算机学派	仿生学派、生理学派	进化主义、控制论学派
核心思想	模拟人脑思维的功能	模拟大脑神经网络的结构	模拟智能系统的行为
实现机制	模拟人脑思维流程，用符号表达方式来研究智能，实现搜索、推理、学习等能力	仿造人的神经系统，把人的神经系统的模型用计算的方式呈现，用它来仿造智能	基于自适应、自学习、自组织的感知-控制模型，模拟人在控制过程中的智能行为
面临的问题	知识瓶颈	复杂性、算法黑箱	浅层智能
典型代表	机器定理证明	神经网络	波士顿后空翻机器人

1. 符号主义 (Symbolism)

符号主义又称逻辑主义、心理学派或计算机学派。该学派认为人工智能源于数理逻辑，人的认知基元是符号，而且认知过程即符号操作过程。知识是信息的一种形式，是构成智能的基础，人工智能的核心问题是知识表示、知识推理和知识运用。

符号主义主要采用逻辑推理的方法来实现人工智能。例如，通过构建专家系统来模拟人类专家的知识和推理过程。专家系统中有一个知识库，里面存储了大量的规则和事实，以符号形式表示。当给定一个问题时，系统会根据这些规则和事实进行逻辑推理，从而得出结论。

符号主义在早期人工智能的发展中起到了关键作用。它推动了知识工程等领域的发展，使得许多基于规则的人工智能系统得以建立，为人工智能从理论走向实际应用迈出了重要的一步。在自然语言处理的早期阶段，符号主义方法也被用于语法分析和语义理解等方面。

2. 连接主义 (Connectionism)

连接主义又称仿生学派或生理学派，其主要观点是人工智能可以通过模拟人脑的神经系统结构来实现。大脑是由神经元组成的复杂网络，神经元之间通过突触相互连接，智能活动是由大量神经元的集体活动所产生的。

连接主义的核心技术是人工神经网络。神经网络由大量的神经元节点组成，这些节点之间通过权重连接。在训练过程中，通过调整权重来学习输入和输出之间的关系。例如，在图像识别中，将大量的图像数据输入到神经网络中，神经网络通过反向传播算法等不断调整权重，使得对于不同的图像特征能够产生正确的分类输出。典型的神经网络结构卷积神经网络 CNN 在图像识别领域取得了巨大成功。

连接主义的发展推动了深度学习的兴起。随着计算能力的不断提升和大量数据的出现，神经网络能够学习到非常复杂的模式，在语音识别、图像识别、自然语言处理等诸多领域取得了突破性的成果。它改变了人工智能的发展方向，使得人工智能系统能够自动从数据中学习，而不是完全依赖人工编写的规则。

3. 行为主义 (Behaviourism)

行为主义又称进化主义或控制论学派，该学派认为智能行为可以在与环境的交互作用中不断进化和学习得到，强调智能是在感知环境和做出行动的循环过程中涌现出来的，而不是通过内在的符号表示或复杂的神经网络连接。

行为主义主要采用强化学习的方法。在强化学习中，智能体 (Agent) 在环境中采取行动，环境会根据智能体的行动给予奖励或惩罚，智能体通过不断尝试和调整自己的行为策略，以最大化长期累积奖励。例如，AlphaGo 通过强化学习与自我对弈的方式，不断学习围棋策略，根据胜负的反馈来调整下棋的策略，最终成为一个强大的围棋程序。

行为主义为机器人学和自适应控制系统的发展提供了重要的理论和方法支持。它使得智能体能够在复杂多变的环境中自主学习和适应，在自动驾驶、机器人导航等领域发挥着重要作用。通过强化学习方法开发的智能系统具有很强的适应性和灵活性，能够根据不同的环境和任务要求进行自我调整。

2.1.4 机器学习和深度学习

1. 机器学习 (Machine Learning)

机器学习是人工智能的一个重要分支，是一种能够根据输入数据训练模型的系统。它的主要目标是让计算机系统通过对模型进行训练，使计算机能够从新的或以前未见过的数据中得出有用的预测。

机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。在计算机系统中，“经验”通常以“数据”形式存在。因此，机器学习所研究的主要内容是关于在计算机上从数据中产生“模型” (model) 的算法，即“学习算法” (learning algorithm)。有了学习算法，再把经验数据提供给它，它就能基于这些数据产生模型；在面对新的情况时，模型会提供相应的判断。

2. 深度学习 (Deep Learning)

深度学习是机器学习的一个子领域，其核心在于使用人工神经网络模仿人脑处理信息的方式，通过层次化的方法提取和表示数据的特征。虽然单层神经网络就可以做出近似预测，但是添加更多的隐藏层可以优化预测的精度和准确性。神经网络由许多基本的计算和存储单元组成，这些单元被称为神经元。神经元通过层层连接来处理数据，并且深度学习

模型通常有很多层，因此被称为“深度”学习。深度学习模型能够学习和表示大量复杂的模式，在图像识别、语音识别和自然语言处理等任务中非常有效。

2.1.5 机器学习的类型

机器学习主要分为监督学习、无监督学习、强化学习和联邦学习等类型。图 2-3 所示为机器学习类型示意图。其中，监督学习就像一本有答案的教科书，模型可以从标记的数据中学习，即它有答案可以参考学习；无监督学习则更像一本无答案的谜题，模型需要自己在数据中找出结构和关系；介于两者之间的方法称为强化学习，其模型通过经验学习执行动作。

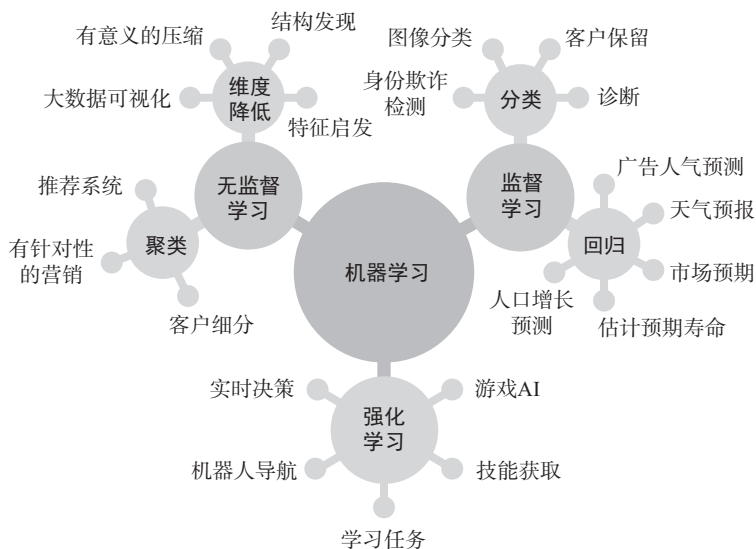


图 2-3 机器学习类型示意图

1. 监督学习 (Supervised Learning)

监督学习也称有导师学习或有监督学习，它使用已标记的训练数据，即数据集中的每个样本都包含输入特征和对应的输出标签（目标值）。模型的任务是学习输入和输出之间的映射关系，从而能够对新的输入数据进行准确的预测。

监督学习的主要类型是分类和回归。

在分类中，机器被训练成将一个组划分为特定的类，例如，邮件中的垃圾邮件过滤器分析用户以前标记为垃圾邮件的电子邮件，并将它们与新邮件进行比较，如果它们有一定的百分比匹配，则这些新邮件将被标记为垃圾邮件并发送到适当的文件夹中。又如，用很多猫猫狗狗的照片和照片对应的“猫”“狗”标签进行训练，然后让模型根据没见过的照片判断照片中是猫还是狗。

在回归中，机器使用先前已标记的数据来预测未来。例如，使用气象事件的历史数据，如平均气温、湿度和降水量等，对未来天气进行预测。又如，使用一些房屋的特征数据，如面积、房间数量、房龄等和相应的房价作为标签进行训练，然后让模型根据没见过的房

屋的特征预测房价。

2. 无监督学习(Unsupervised Learning)

无监督学习也称无导师学习或归纳性学习，它是利用未标记的数据进行学习的机器学习方法。在无监督学习中，数据没有预先定义的标签或目标值，模型需要自己发现数据中的结构、模式和规律。例如，给定一组用户的购物行为数据，无监督学习模型可以发现具有相似购物模式的用户群体，而不需要预先知道这些群体的定义。

无监督学习的主要类型是聚类、数据降维和异常检测。

聚类用于根据属性和行为对象进行分组。它与分类不同，因为这些组不是用户提供的。例如，在客户细分中，将具有相似特征的客户分为一组，以便企业进行针对性的营销；在生物学中，对基因序列进行聚类，以发现基因的相似性和功能类别。

数据降维通过找到共同点来减少数据集的变量。大多数大数据可视化技术使用降维来识别趋势和规则。在高维数据可视化中，将高维数据投影到低维空间，便于观察数据的分布和结构；在特征提取中，减少数据的维度，同时保留主要信息，能够提高后续机器学习模型的效率。

异常检测：在网络安全中，检测网络流量中的异常模式；在工业生产中，发现运行数据中的异常点，提示可能的故障。

3. 强化学习 (Reinforcement Learning)

强化学习也称再励学习、评价学习或增强学习，它让模型在环境里采取行动，获得结果反馈。模型从反馈里学习，从而能在给定情况下采取最佳行动来最大化奖励或是最小化损失。例如，在一个机器人控制的强化学习场景中，机器人(智能体 agent)在房间(环境 environment)中移动，它的每一个动作(action)，如向前走、转弯等，都会得到一个奖励(reward)信号，如到达目标位置得到正奖励，碰撞障碍物得到负奖励，机器人通过不断尝试来学习最优的行动策略，以最快的速度到达目标位置并避免碰撞。又如，在一个游戏环境中，游戏的画面和游戏角色的状态构成环境状态，游戏角色的操作，如跳跃、攻击等，是动作空间，完成任务或获得高分得到正奖励，失败或扣分得到负奖励。

强化学习的主要应用场景是机器人控制、游戏行业和资源管理。

机器人控制：包括工业机器人的路径规划、无人机的自主飞行等，使机器人能够在复杂的环境中完成任务并避免危险。

游戏行业：训练游戏中的智能角色，如让游戏中的 NPC(Non-Player Character, 非玩家角色)能够根据环境和玩家的行为做出合理的反应；在棋类游戏中，训练智能体学会下棋策略。

资源管理：如数据中心的能源管理，通过调整服务器的工作状态来优化能源消耗，同时保证服务质量。

4. 联邦学习 (Federated Learning)

联邦学习由 Google AI 团队在 2016 年提出，主要是为了解决移动设备的模型训练问题，在保护用户隐私的同时进行模型协同训练。

联邦学习是指一种多个参与方在保证各自原始私有数据不出数据方定义的可信域的前提下，以保护隐私数据的方式交换中间计算结果，从而协作完成某项机器学习任务的模式。

例如，多个医院参与联邦学习，每个医院使用自己的患者数据在本地训练模型，然后将训练后的参数加密发送给一个中心服务器，中心服务器聚合这些参数得到一个更准确的疾病诊断模型，而在这个过程中患者的数据始终没有离开医院。

联邦学习的主要应用场景是隐私敏感和物联网领域。

隐私敏感领域：如金融行业，银行之间可以在不共享客户财务数据的情况下，共同训练风险评估模型；医疗行业，多家医疗机构可以联合训练疾病诊断或药物研发模型。

物联网 (Internet of Things, IoT) 领域：物联网是指通过各种信息传感器、射频识别技术、全球定位系统、红外感应器、激光扫描器等各种装置与技术，实时采集任何需要监控、连接、互动的物体或过程，采集其声、光、热、电、力学、化学、生物、位置等各种需要的信息，通过各类可能的网络接入，实现物与物、人与物之间的泛在连接，实现对物品和过程的智能化感知、识别和管理。大量的物联网设备，如智能家居设备、工业传感器等，可以通过联邦学习共同训练模型，用于设备故障预测、能源管理等，同时保护用户的隐私和设备数据安全。

2.1.6 AI 的技术发展方向

2024 年 4 月，中国工程院孙凝晖院士在十四届全国人大常委会举行的第十讲专题讲座《人工智能与智能计算的发展》中预测，人工智能的技术前沿将朝着以下 4 个方向发展。

1. 多模态大模型

从人类视角出发，人类智能是天然多模态的，人拥有眼、耳、鼻、舌、身、嘴（语言），从 AI 视角出发，视觉，听觉等也都可以建模为 token 的序列，可采取与大语言模型相同的方法进行学习，并进一步与语言中的语义进行对齐，实现多模态对齐的智能能力。

2. 视频生成大模型

OpenAI 公司于 2024 年 2 月发布文生视频模型 Sora，将视频生成时长从几秒钟大幅提升到一分钟，且在分辨率、画面真实度、时序一致性等方面都有显著提升。Sora 的最大意义是它具备了世界模型 (World Models) 的基本特征，即人类观察世界并进一步预测世界的能力。世界模型是建立在理解世界的基本物理常识之上，观察并预测下一秒将要发生什么事件。虽然 Sora 要成为世界模型仍然存在很多问题，但可以认为 Sora 具备了画面想象力和分钟级未来预测能力，这是世界模型的基础特征。

3. 具身智能

具身智能指有身体并支持与物理世界进行交互的智能体，如机器人、无人车等，通过多模态大模型处理多种传感数据输入，由大模型生成运动指令对智能体进行驱动，替代传统基于规则或者数学公式的运动驱动方式，实现虚拟和现实的深度融合。因此，具有具身智能的机器人，可以聚集人工智能的三大流派：以神经网络为代表的连接主义，以知识工程为代表的符号主义和控制论相关的行为主义，三大流派可以同时作用在一个智能体，预期会带来新的技术突破。

4. AI4R

AI4R (AI for Research) 已成为科学发现与技术发明的主要范式。由于人工智能大模型

具有全量数据，具备上帝视角，借助深度学习的能力，人工智能大模型能够比人向前看更多步数。如果能实现从推断到推理的跃升，人工智能模型就有潜力具备像爱因斯坦那样的想象力和科学猜想能力，这将极大地提升人类科学发现的效率，打破人类的认知边界。

2.2 我国人工智能产业

为贯彻落实《国家标准化发展纲要》和《全球人工智能治理倡议》，进一步加强人工智能标准化工作系统谋划，加快构建满足人工智能产业高质量发展和“人工智能+”高水平赋能需求的标准体系，夯实标准对推动技术进步、促进企业发展、引领产业升级、保障产业安全的支撑作用，更好推进人工智能赋能新型工业化，2024年6月，工业和信息化部、中央网络安全和信息化委员会办公室、国家发展和改革委员会、国家标准化管理委员会等四部门印发了《国家人工智能产业综合标准化体系建设指南(2024版)》。本节内容主要来源于此指南。

2.2.1 AI 产业发展现状

人工智能正成为发展新质生产力的重要引擎，加速与实体经济的深度融合，全面赋能新型工业化，深刻改变工业生产模式和经济发展形态，将对我国加快建设制造强国、质量强国、网络强国和数字中国发挥重要的支撑作用。

人工智能产业链包括基础层、框架层、模型层和应用层等4个部分。其中，基础层主要包括算力、算法和数据；框架层主要指用于模型开发的深度学习框架和工具；模型层主要指大模型等技术和产品；应用层主要指人工智能技术在行业场景的应用。

目前，我国人工智能产业在技术创新、产品创造和行业应用等方面实现了快速发展，形成庞大的市场规模。伴随以大模型为代表的新技术加速迭代，人工智能产业呈现出创新技术群体突破、行业应用融合发展、国际合作深度协同等新特点，亟需完善人工智能产业标准体系。

2.2.2 AI 标准化体系建设总体要求

1. 坚持创新驱动

优化产业科技创新与标准化联动机制，加快人工智能领域关键共性技术研究，推动先进适用的科技创新成果高效转化成标准。

2. 坚持应用牵引

坚持企业主体、市场导向，面向行业应用需求，强化创新成果迭代和应用场景构建，协同推进人工智能与重点行业融合应用。

3. 坚持产业协同

加强人工智能全产业链标准化工作协同，加强跨行业、跨领域标准化技术组织的协作，