普通高中教科书

信息技术

选择性必修 3

数据管理与分析

人民教育出版社 中国地图出版社

微信公众号：电子课本大全

普 通 高 中 教 科 书

信息技术

选择性必修 3

数据管理与分析

人民教育出版社课程教材研究所信息技术课程教材研究开发中心
中国地图出版社教材出版分社 编著

总主编 祝智庭 樊 磊

人民教育出版社 中国地图出版社
·北京·

微信公众号：电子课本大全

总 主 编：祝智庭　樊　磊
副总主编：郭　芳　高淑印　李　锋

本册主编：黄应会　倪俊杰
编写人员：杜宗飞　杨　俊　赵婕瑜　钱华斌　徐建东

责任编辑：朱从娜　刘利华
美术编辑：李　媛　徐海燕

微信公众号：电子课本大全



前言

同学们，欢迎探索信息技术这个神奇而充满魅力的世界。

在以往的学习、生活中，你们已经积累了许多信息技术方面的知识与技能，例如：在网上查阅资料，用手机与亲朋好友保持联系，使用移动终端、自动柜员机等设备……你们知道这些应用中都包含哪些关键技术，涉及哪些领域吗？怎样有效地利用这些技术帮助我们培养信息意识，提升计算思维，进而通过数字化学习与创新，承担起信息社会责任呢？即将开始的这门课程，会帮助你们对信息技术有更多的认识和思考，获得更丰富的体验和感受。

为了很好地掌握信息技术，希望同学们按以下三个要求去努力。

1. 认真阅读教科书，理解基本概念和原理。信息技术发展非常迅猛、各类信息系统不断涌现，但信息系统的基础和运行体系相对稳定，离不开算法的设计及对数据的利用。只有夯实基础，才能学好本领，跟上时代发展的步伐。

2. 敢于动手，勤于实践。信息技术是一门实践性较强的课程。实践能帮助同学们熟练操作技能，进一步掌握知识。因此，要认真阅读理解每章的主题学习项目，并逐步完成“实践活动”“思考活动”“技术支持”“阅读拓展”等栏目的学习内容，在实践中获取知识和经验。

3. 要有积极探究、锲而不舍的精神。掌握信息技术的知识与技能需要一个过程，不可能一蹴而就。信息技术学科内容非常丰富，各知识点之间联系密切，但名词术语多，有可能令人感到繁杂，甚至产生畏难情绪。学习新知识，首先要知其然，接着通过不断学习，积极动手操作，大胆请教，加深对知识的理解，然后才能知其所以然，在不断的探索过程中取得进步。

本书中涉及的配套资源，可在教科书配套教学资源平台的信息技术栏目中获得。让我们开始一段信息技术新旅程，成长为信息社会中合格的中国公民！

微信公众号：电子课本大全

目录



第1章　数据与数据科学　1

主题学习项目：走近送货机器人　2

1.1　从数据到数据科学　3

1.1.1　数据及其价值　4

1.1.2　大数据及其应用价值　7

1.1.3　数据科学　15

1.2　数据管理与分析简介　20

1.2.1　数据管理的发展　21

1.2.2　大数据存储与管理　24

1.2.3　数据分析及其基本过程　26

1.2.4　数据分析助力科学决策　29

总结评价　32

第2章　需求分析与数据采集　33

主题学习项目：交通数据见发展　34

2.1　业务需求与解决方案　35

2.1.1　认识业务需求分析　36

2.1.2　设计解决方案　37

2.1.3　数据需求分析　38

2.2　数据采集与导入　40

2.2.1　数据采集途径　41

2.2.2　创建CSV数据文件　43



微信公众号：电子课本大全

2.2.3 从网络中采集数据 45
2.2.4 导出CSV文件中的数据 49
2.3 数据结构化与数据清洗 52
2.3.1 不同结构化程度的数据 53
2.3.2 噪声数据的现象与成因 55
2.3.3 数据清洗 56
总结评价 62



第3章 数据管理 63
主题学习项目：数据管理助规划 64
3.1 数据库与数据管理 65
3.1.1 数据库与数据库管理系统 66
3.1.2 确定数据库的基本功能 67
3.1.3 建立概念数据模型 68
3.2 设计逻辑结构与建立数据库 74
3.2.1 概念模型转换为关系模型 75
3.2.2 创建和查看数据库 76
3.2.3 MySQL的数据类型 79
3.2.4 创建和查看数据表 81
3.2.5 修改和删除数据表 82
3.2.6 将数据输入数据表 82
3.3 结构化查询与提取 85
3.3.1 结构化查询语言 86

微信公众号：电子课本大全

3.3.2 数据库的查询方法 87
3.3.3 查询数据的提取 92
3.3.4 编程实现SQL查询 93
3.4 备份和还原数据库 96
3.4.1 数据丢失常见的原因 97
3.4.2 常见的备份方法 98
3.4.3 备份与还原数据库 99
总结评价 104
第4章 数据分析 105
主题学习项目：数据分析知天气 106
4.1 数据分析的工具与方法 107
4.1.1 数据分析的工具 108
4.1.2 常用的数据分析方法 108
4.1.3 数据挖掘 118
4.2 数据可视化与数据报告 121
4.2.1 数据可视化中的图形 122
4.2.2 数据可视化的步骤 122
4.2.3 编程实现数据可视化 123
4.2.4 撰写数据分析报告 127
总结评价 131
项目评价 132



微信公众号：电子课本大全

第 1 章
数据与数据科学

自然界的各种现象，植物的生长、动物的习性、人类的思想行为……都可以用数据的形式存储在各类载体之中。随着大数据、云计算和人工智能技术的发展和应用，数据已经成为信息社会的重要资源，成为支撑科学研究、技术进步和社会发展不可或缺的基础。因此，我们可以从社会生产生活中提取数据，然后利用计算思维、运算方法、算法模型等，研究这些数据的类型、状态、属性以及变化形式和规律，并通过科学的管理和分析，获取有价值的信息，从而构建知识、获得智慧，为社会经济发展提供决策依据。



微信公众号：电子课本大全

01 主题学习项目：走近送货机器人

项目目标

想象一下，你在网上购买了几本书，收到“快递已被机器人揽收”的信息后，通过网购平台“告诉”这位机器人快递员你所在的位置，它就能自动优化路线，然后穿梭在城市道路间，避开障碍物，最终出现在你面前。此时，你可能只需“刷脸”，机器人的货仓就会自动打开，轻松递出你的书。

本章以“走近送货机器人”为主题项目，开展学习活动。

1. 了解送货机器人涉及的数据，理解数据对送货机器人的价值。

2. 了解数据科学与送货机器人研制之间的关系。

3. 了解数据管理与分析对送货机器人的重要作用、对挖掘数据价值的意义以及对科学决策的支持作用，了解该领域的发展前景。

项目准备

为了完成项目，需要做以下准备。

- 组建学习小组。开展学习过程中，小组成员要互相讨论、独立思考、相互协作。
- 查阅送货机器人的研发动态。在网络上查阅资料，了解我国送货机器人的研发和应用情况。
- 学习过程中，要充分利用数字化学习工具，例如使用思维导图软件绘制概念和操作之间的关系。

项目过程

为了保证顺利完成本项目的学习活动，在不同学习阶段，小组长要注意检查组员项目学习的进度，并做好协调互助工作。

细化方案 ①

细化项目学习计划，调研送货机器人的应用情况以及所涉及的数据。 P6

调研学习 ②

了解数据与数据科学对送货机器人研制的作用及其所涉及的学科领域。 P19

完成方案 ③

了解数据管理与分析对送货机器人的作用，畅想其未来发展前景。 P30

项目总结

通过本章的项目学习，加深对数据、大数据、数据科学以及数据存储、数据管理、数据分析等核心概念的理解，认同数据及其价值对机器学习、人工智能等领域发展的重要意义，为进一步应用数据管理与分析技术进行创新学习奠定基础。

微信公众号：电子课本大全

1.1
从数据到数据科学

学习目标 ▶▶▶

- 进一步理解数据的含义及其价值。
- 进一步理解大数据及其应用价值。
- 了解数据、大数据与数据科学的关系。
- 认识数据科学的内涵，感受数据科学研究的重要意义。

体验探索

城市里的数据

提起城市，你的脑海中会闪现什么样的景象？请用几个关键词描绘城市印象。

透过城市的表面现象（图1.1.1），你认为是什么在“暗暗地”支撑着城市的正常运转呢？城市的一般印象往往离不开高楼大厦、宽阔的街道、车水马龙、人来人往、繁忙的地铁、设施齐全的社区……在描述是什么在支持城市运转时，也许你会想到“技术”“管理”“监控”“决策”等词汇。事实上，“数据”既是城市运转与发展的无形资源支持，也是巨大的财富。不仅是城市，其实只要有人生活的地方，人们都在不断地创造数据、产生数据。



图1.1.1　城市景象举例

观察与思考：观察周围环境（学校、社区和街道）并描述其景象；思考：这些景象可能涉及哪些数据？人们的哪些活动在创造和产生数据？

微信公众号：电子课本大全

1.1.1 数据及其价值

前面的体验探索告诉我们：在热闹繁华的城市景象中隐藏着丰富的数据。数据是描述事物的符号记录，是信息的载体。在计算机科学中，数据是计算机识别、存储和加工的对象，例如字符、图像和音频等。

每个人都在创造数据，例如：打电话产生的数据可以用来改进通话网络；乘坐公共交通工具的数据可以用来优化公交网络；购买商品产生的数据可以用来调整生产与进货规模。与此同时，人们在不断通过各种信息工具获取数据，如交通拥堵、空气质量、天气、高校录取率、医院专家出诊、居民收入水平等。数据已成为重要的信息资源。

人们可以从数据中获得对自己有价值的信息，更重要的是可以学习知识、增长智慧。要理解“数据具有价值”这句话，首先要了解数据、信息、知识与智慧的关系。

思考活动

对牙膏瓶盖的思考

赵明生活在北方，冬天比较干燥。他买了一支按盖型牙膏，才用了2天，挤牙膏就很费劲。他发现，瓶口被干硬的牙膏堵住了。他很快意识到：该品牌按盖的密封性不好，北方干燥的气候使牙膏水分被快速蒸发掉。他陷入了思考：该品牌的按盖设计有什么缺陷？其他品牌的按盖又是如何设计的？通过对比后，他决定以后改买拧盖型的牙膏或另一个品牌的按盖型牙膏。

思考：对牙膏瓶盖的缺陷，赵明经历了怎样的一个思维过程？

赵明以上的思维过程，可以按数据、信息、知识、智慧逐层来分析（图1.1.2），他的思考不仅可以改进产品，为厂家提升效益，同时也为自己以后购买牙膏时提供了决策。

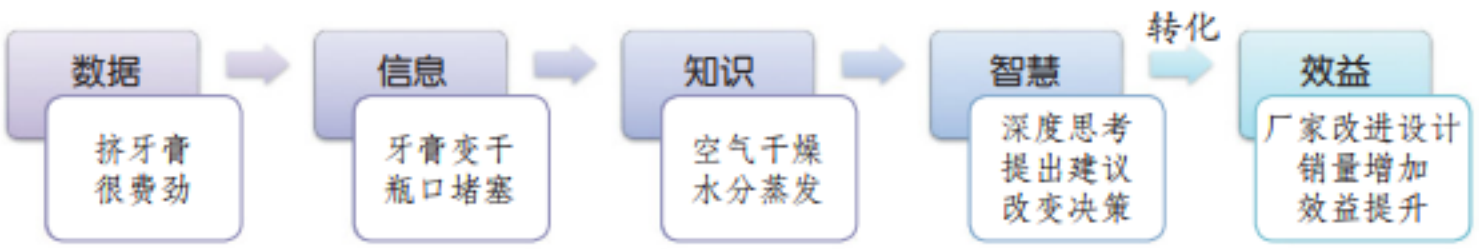


图1.1.2　从数据到信息、知识再到智慧的思考

数据描述了事物客观存在的各种属性，信息是经过加工处理后

微信公众号：电子课本大全

的数据，知识是有组织的、被记忆的信息，智慧是知识的有效应用。通常，人们能相对容易地从数据中获取对自己有用的信息，但要把数据或信息转化为知识、智慧，往往需要深入的分析与挖掘。只有这样，才能发挥数据的价值，让它成为改造社会的智慧工具。

图 1.1.3 所示的 DIKW（data information knowledge wisdom，数据信息知识智慧）金字塔，表明了从“数据”到“智慧”的转变过程，同时也是“从认识部分到理解整体、从描述过去与现在到预测未来”的过程，简明地描绘了数据、信息、知识、智慧的联系。

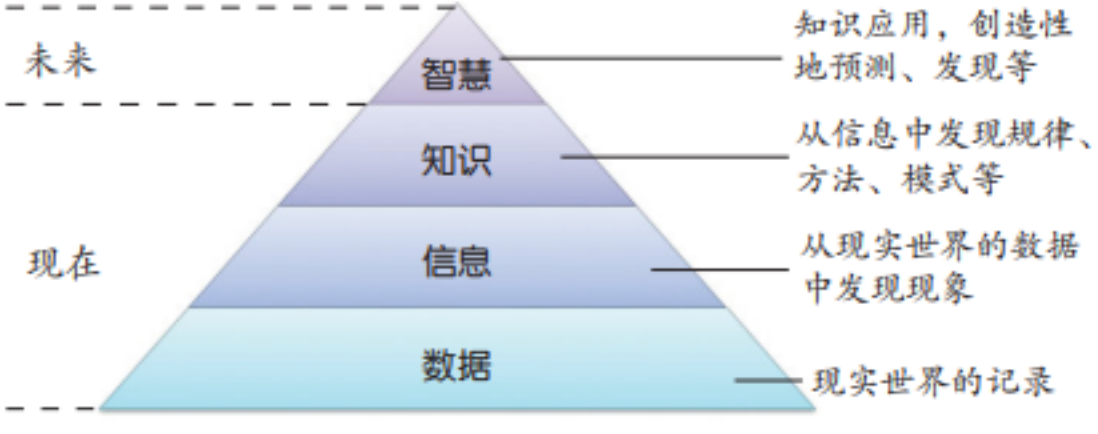


图 1.1.3　DIKW 金字塔

数据的价值在于发现其背后的事实与规律，并通过信息、知识、智慧三个层面体现。对个人来说，只要能从数据中获得有用的信息、支持自己做决策，数据就有价值；而这些信息让你对事物有了新的认知或者构建了新的知识，数据的价值就得到了提升；如果这些信息或知识让你在思考和创新方面取得了进步，就产生了智慧的价值。

思考活动

维修数据与遥控器的改良

陈捷是多家品牌电视机的售后修理人员。他在整理和打印工作清单（包含设备名称、品牌、损坏部位、修理措施、费用等）时，得到这样的信息：遥控器修理中，99% 的用户都是修理开关键和频道切换键，极个别用户修理音量调节键。

思考：你能从以上数据中获得更多信息吗？你认为从陈捷的工作清单中，还可以进行哪些方面的数据调查？你会给厂家提出什么建议？

不难发现，从数据到信息、知识、智慧，其中的任何一个环节都需要对数据进行有效的管理与分析。对个人来说，这些工作很多时候是潜移默化、无意识展开的。事实上，大脑在进行这一系列思维活动时，已经涉及数据分析与呈现的相关知识。

微信公众号：电子课本大全

项目实施

了解送货机器人与数据的关系

小组成员一起细化项目学习计划，调查送货机器人的应用情况，以及送货机器人送货过程中所涉及的数据。

1. 小组成员一起讨论，确定要调研的内容、活动过程和具体实施方法，然后进行任务分工，明确各自的任务。

2. 参考表1.1.1进行调研，了解国内外有哪些公司正在研制送货机器人，这些送货机器人在哪些城市或路段试用。

表1.1.1　送货机器人研制与试用调研表

研制送货机器人的公司	送货机器人功能简介	试用的城市或路段

3. 简述送货机器人在送货过程中涉及的主要数据，以及这些数据对送货机器人所起的决策作用（参考表1.1.2）。

表1.1.2　送货机器人涉及的数据及其作用

数　　据		数据隐含的价值	决策作用
与物流相关的数据			
与路线相关的数据			

微信公众号：电子课本大全

1.1.2 大数据及其应用价值

大数据正在改变着人们的工作、生活与思维模式，进而对文化、技术和学术研究产生深远的影响。

思考活动

大数据与流感趋势预测

新型流感病例的发现到通告，时间上一般会有延迟，从而导致公共卫生机构无法及时应对。在流感高发地区，流感相关知识的搜索趋势与流感的流行趋势及严重程度存在一定的相关性。把这些搜索结果汇总起来，达到足够数量时，就可以建立一个数据系统，用于实时监控流感疫情，预测未来疫情状况。2008年，工程师们曾推出了流感指数的相应产品，用于预测流感疫情。我国相关机构也根据搜索的数据进行了分析，为预测流感提供了决策依据，并为居民提供预警服务。例如，自2015年3月以来，深圳市疾病预防控制中心开始定期提供流感指数预警服务（图1.1.4），提醒人们注意防范。



图1.1.4 流感指数预警

思考：如何从流感程度、流感周期、疾病特征、需求特征、地域特征、人群特征等方面入手，利用网络大数据监测和预防流感疫情？

物联网、移动互联网、人工智能、大数据计算等技术的发展，实现了人与人、人与物、物与物之间的互联，引发了数据规模的爆炸式增长和数据模式的高度复杂化，世界已进入大数据时代。

大数据的内涵

不同领域的专家对大数据有不同的理解，下面从不同角度列举主要的三种。

计算机科学与技术。当数据的量、复杂程度、处理的任务要求等超出了传统数据的存储与计算能力时，就可以称为“大数据”。这是从存储和计算能力的视角来认识的，主要涉及数据存量、数据增量、复杂程度和处理要求等。

微信公众号：电子课本大全

统计学。当能够收集足够的全部或绝大部分个体的数据，且计算能力足够强，可以不用抽样，在总体数据上就可以进行统计分析时，就被称为“大数据”。可见，这一领域认为大数据不是绝对概念，而是相对于总体规模和统计分析方法选择的相对概念。

机器学习。当训练集足够大且计算能力足够强，只需通过对已有的实例进行简单查询即可达到“智能计算的效果”时，这里的数据一般需要大数据的支撑。机器学习就是用数据或以往的经验优化计算机程序的性能标准，这也是大数据应用的典型案例。



阅读拓展

人工智能与数据库

在人工智能系统中，除了先进的硬件、软件系统外，还需要大型数据库的支撑。例如，阿尔法围棋（AlphaGo）不仅记忆超强、计算速度快，还能够自主学习。其核心系统是基于神经网络的深度学习，即模拟人脑的神经网络，通过数据分析，学习了大量的职业棋手棋谱，再通过增强学习方法的自我博弈，寻找比基础棋谱更多的打点来击败人类。

大数据的主要特征

从不同的领域看，大数据表现出多种不同的特征。但通常认为，它具有巨量性、多样性、迅变性、价值性等特征（图1.1.5）。

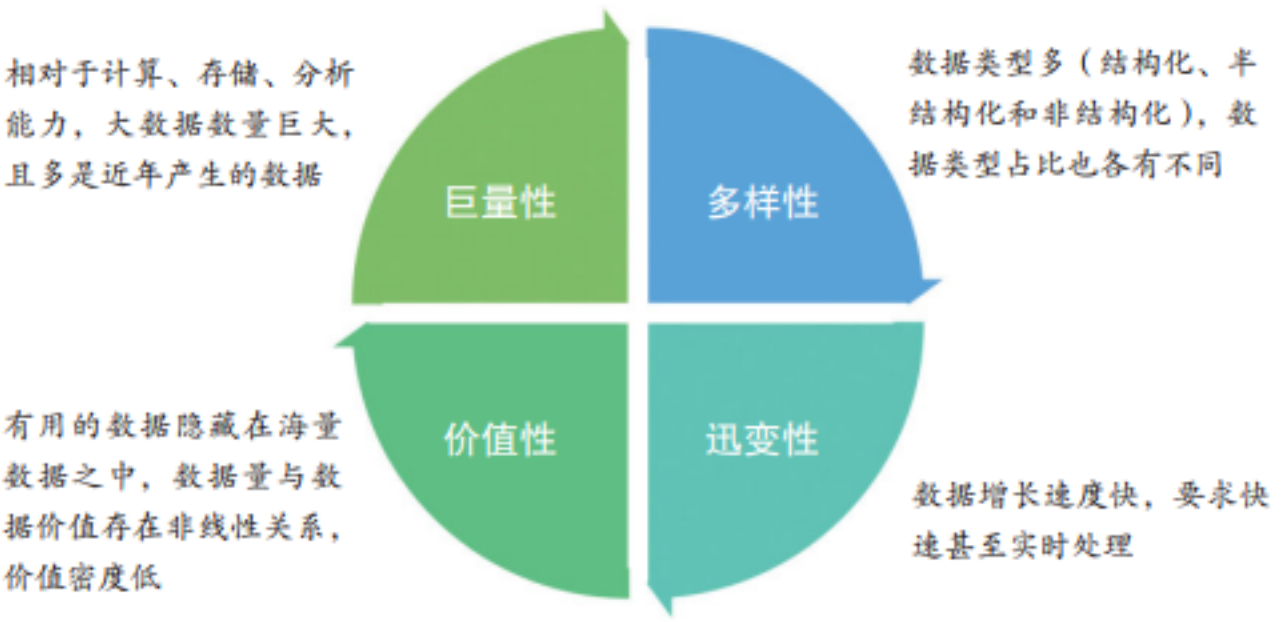


图1.1.5　大数据的主要特征

微信公众号：电子课本大全

巨量性。数据量已经大到无法在可容忍的时间内用传统技术和软硬件工具进行感知、获取、管理、处理和服务。数据规模已从GB到TB再到PB级，甚至开始以EB和ZB来计数。

早在2011年，国际数据集团的研究报告就指出：未来10年全球大数据将增加50倍，管理数据仓库的服务器数量增加10倍。例如，当前一些机构之所以能预测流感的发生，实际上是得益于其拥有大量的数据，他们保存了多年来的搜索记录，而且每天会收到全球超过几十亿条的搜索指令。

1 ZB = 1 024 EB
1 EB = 1 024 PB
1 PB = 1 024 TB
1 TB = 1 024 GB
1 GB = 1 024 MB

阅读拓展

大数据到底有“多大”

2013年左右，一组名为“互联网上一天”的数据显示：一天之中，互联网产生的全部数据可以刻满1.68亿张数字视频光盘；发出的邮件有2 940亿封；发出的社区帖子达200万个……国际商业机器（international business machines, IBM）公司研究称，整个人类文明所获得的全部数据，90%以上是近几年产生的。这样的趋势会持续下去。2017年和2018年，有人用饼图展示了“互联网上一分钟”的数据，都是非常惊人的数据，一年的电子邮件发送数量就接近100万亿。

多样性。大数据技术采集的各种类型数据，既包括传统数据库里结构化的数据，也包括非结构化的数据。在大数据中，目前仅有20%左右属于结构化数据，其余数据属于广泛存在于社交网络、电子商务、物联网等领域的非结构化或半结构化数据。例如，人们网络购物后，通常会对商品和服务进行评价。评定的星级通常属于结构化数据，写的评语、上传的图片或视频则属于非结构化数据，分析处理这类数据需要采用专门的数据处理技术和方法。又如，一个关系数据库管理系统中可能存储着支持呼叫中心的呼叫日志。管理系统将呼叫的特征存储为结构化数据，这些数据具有时间戳、机器类型、问题类型和操作系统等属性。管理系统还可能存储着非结构化数据或半结构化数据，如电子邮件故障单、客户聊天记录、描述问题的通话记录等。

关于数据结构化程度的内容，本书第2章还将进一步介绍。

迅变性。互联网和物联网（图1.1.6）是大数据的主要来源，各类传感器、智能仪表、监控系统和智能终端等，能够实时自动采集和生成数据，使得数据以空前的速度产生。同时，大数据往往以数据流的形式动态产生，数据的状态与价值随时空的变化而发生演变，

微信公众号：电子课本大全

具有很强的时效性（图1.1.7）。只有掌控好数据流，才能有效利用这些数据。



图1.1.6　物物相联



图1.1.7　实时数据流动

随着移动互联网的发展，人们对数据实时应用的需要更加迫切，例如，人们用手机关注天气、交通、物流、医疗等信息时，就要求数据的处理速度要与数据的增长速度相适应。

思考活动

技术促进数据的感知和应用

有些东西无法用眼睛看到（如黑暗中的物体），但人们可以借助手和身体去感知；有些东西无法看到，也无法触摸到，如音乐、对话等，但可以借助耳朵来感知；还有些东西无法用感官直接感知到，如紫外线、红外线、细胞、粒子、电磁波等，但可以利用仪器和工具来感知它们的存在。

思考：技术的进步、工具的使用，对人们感知和应用数据会产生哪些影响？对大数据领域的发展和研究带来什么意义？（例如，传感技术的进步和相关工具的应用。）

我们被淹没在数据的海洋之中，却又在忍受着知识的饥渴。

价值性。虽然数据的价值巨大，但是基于传统思维与技术，人们在实际环境中往往面临着信息泛滥而知识匮乏的窘境，大数据的价值利用密度比较低。有价值的数据往往被隐藏在大量无用的数据之中，只有进行深度分析和挖掘才能发现其中的价值。例如，在一段24小时的不间断监控视频中，有用数据可能仅有几秒，甚至多数时候没有用。因此，如何在大数据中发现有价值的数据并转化为信息、知识，已成为大数据分析与管理的重要研究领域。

还有一种观点认为真实性也是大数据的一个特征。真实性主要指数据质量的反映。越接近真实的数据越有助于正确决策，数据规模并不能完全决定能否为决策提供充分依据，但大数据的大样本甚至全样本有利于接近或反映真实性。

微信公众号：电子课本大全

实践活动

进一步认识大数据的特征

除了以上特征外，大数据还有哪些显著特征？请查阅资料并填写表1.1.3，并与小组成员分享你的认识和理解。

表1.1.3　大数据的特征分析

主要特征	举例说明	参考资料

大数据的应用价值

目前，大数据的应用价值主要体现在商业价值、产业价值、科研价值、社会价值等方面。

思考活动

电商数据的价值

某电商的交易分析报告中显示：大额埋单后的重购次单和同店重购次单比例分别为25.0%和16.8%，明显高于普通买单的18.8%和10.7%。由此可知，买家首次订单获得对卖家商品与服务质量的信任后，次单存在加大购买金额的可能。为此，卖家通过跟进服务、适时推荐、坚守质量等，以求获得同类商品大额下单的概率。

思考：在以上情境中，数据背后存在哪些价值？

商业价值。精准预测商业价值是大数据技术发展带来的一种新型能力。在商业领域，客流数据、经营数据、商品数据、浏览人数和点击量等看似简单的数据背后其实隐藏着很大的商机。通过把相关算法运用到数据处理中，就可以获得有价值的产品、服务以及对发展趋势的预测。例如，企业通过分析大量客户的生活方式、行为习惯、网页访问频率、信息搜索记录、商品购买记录等，可以了解客户的爱好、职业、性格等信息，进而分析他们的需求，并预测他们近期的消费行为，从而有针对性地为他们提供服务。此外，大数据能够满足人们不同应用场景的需要，将生活的各个方面融合，让人享受到非常便捷和舒适的信息服务。例如，当你来到一座城市，地图导航软件会及时推送当地的游玩攻略、美食地图、天气信息等。

微信公众号：电子课本大全

阅读拓展

信用卡用户消费记录与商业预测

有的支付公司通过为小银行和商家提供服务，从自己的服务网站调用大量的交易信息和顾客消费信息，收集和分析了信用卡用户的交易记录，用来预测商业发展和客户的消费趋势，然后把分析结果卖给需要的公司。例如，他们预测到某人在下午四点左右给汽车加油，很可能在接下来的1小时内，这个人要去商场购物或去餐馆用餐。商家获取这一预测信息后，就在加油小票的后面附上附近商店或餐馆的优惠券，从而提高商家的销售业绩。

产业价值。大数据是现有产业升级与新产业诞生的重要推动力量。大数据时代的到来，产业界需求与关注点发生了转变。例如，企业关注的重点转向数据，计算机行业从追求计算速度转变为关注大数据处理能力，软件也将从以编程为主转变为以数据为中心。又如，采用大数据处理方法，新材料研制生产的流程会发生革命性的变化，可以通过数据处理能力较强的计算机并行处理，同时进行大批量的仿真比较和筛选，从而提高科研和生产效率。

在药物筛选实验中，大数据技术使大样本或全样本成为可能，大大提高了实验数据的准确性，缩短了实验周期，使实验结果更真实可靠。

科研价值。大数据技术的研发与应用助推了科学技术的快速发展，引发了科技界对科学研究方法的重新审视。最早的科学研究只有实验科学，随后出现了以研究各种定律和定理为特征的理论科学和以模拟仿真为特征的计算科学。大数据的出现催生了一种新的科研模式，图灵奖得主詹姆士·格雷（James Gray）提出了科学研究的第四范式——数据密集型科学。他认为，科研人员只需从数据中直接查找或挖掘所需要的数据信息，甚至无须接触需研究的物理对象。例如，地质学家不再需要每次都亲临地质现场拍照勘察，而是从大数据中发现所需的高清地质影像等数据。第四范式不仅是科研方式的转变，而且是人们思维方式的转变。

社会价值。大数据是与自然资源、人力资源一样重要的战略资源。大数据时代，国家层面的竞争力将部分体现为拥有大数据的规模、活性，以及对数据解释和运用的能力。同时，科学技术发展的最终目的都要落到促进人类社会发展、增进人的幸福感等方面。大数据为我们带来的不仅是便利，还有紧密的生活服务网络。例如，在重大节假日活动中，容易出现因人群过度拥挤而引发的危险，通过大数据分析，可以预测人流情况，从而能及早采取疏散措施（图1.1.8）。



图1.1.8　大数据助力城市管理

微信公众号：电子课本大全

大数据的来源

大数据可以通过不同方式和渠道来获取。物联网、云数据库、移动互联网、车联网、手机、平板计算机、台式计算机以及遍布各个角落的各种传感器，都是数据来源或承载的媒介。

归纳起来，大数据主要来源于以下三方面。

传统数据库。大数据是政府、企业、组织、机构等社会各部门实施科学管理和决策分析的基础，这些部门往往构建了基于网络的事务处理系统和办公自动化系统，用传统的数据库来记录存储事务处理的各种数据。传统数据库中的数据蕴含着更多的潜在价值，对形成科学决策起着关键作用，是大数据的重要来源。

教育领域有很多信息系统，如学校的选课系统、成绩管理系统、在线阅卷系统、高考志愿填报与录取系统、校园一卡通系统等，这些系统的运行都离不开传统数据库的支撑。

互联网数据。互联网上的任何行为都会产生数据并被记录下来。从电子邮件、博客、微信等社交媒体产生的数据，到文本、图片、音频、视频文件的交流与共享，再到在线交易、网上购物、电子商务等，每时每刻都在产生大量数据。目前，这些互联网数据是大数据最有价值的来源。

互联网企业、机构是大数据的主要记录和收集地。早在2011年，据IDG统计，全球创建和复制的数据总量已达1.8 ZB，其中75%来自个人（图片、视频和音乐等），远远超过人类有史以来所有印刷材料的数据总量。

实践活动

互联网海量数据处理

2012年，有专家在《大数据研究的科学价值》一文中提到，有的网站通过大规模集群和相关软件，每月处理的数据量超过400 PB；有的网站每天要处理几十拍字节（PB）的数据；有的网购平台在线商品超过8.8亿种，每天交易数千万笔，产生约20 TB数据，这个平台还通过分析几百个类目的主要商品价格来统计其CPI，从而预测经济走势和行业发展动态。

这么多年过去了，这个数据有了很大的变化，请查阅资料，了解相关互联网公司数据处理量的最新情况。

物联网数据。物联网利用互联网、电信网络等信息承载体，把所有能行使独立功能的普通物体连接起来，形成人员、机器、物体的互联互通，而大数据技术真正把人类带进人、机、物融合的世界。通过物联网可以对设备、人员进行集中管理、控制，也可以对家庭设备、汽车等进行遥控，以及搜索位置，防止物品被盗等。物联网

微信公众号：电子课本大全

（图1.1.9）的发展同时又是大数据应用的又一推动力。目前，各类传感器、智能仪表、视频监控、智能终端等，都在以不同方式实时地采集、生成和传递大量数据。

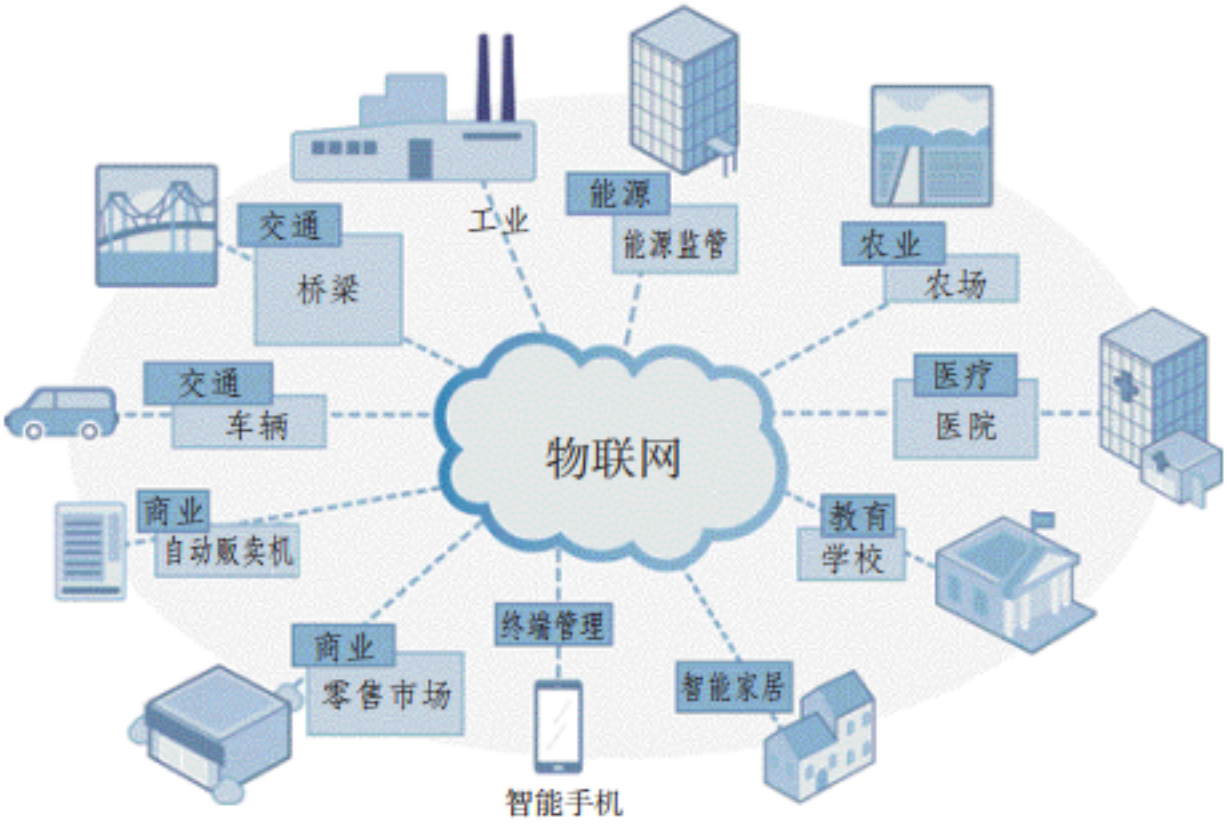


图1.1.9　人、机、物融合的物联网

综合来看，大数据的来源可以粗略地分成两类：一类来自物理世界；另一类来自人类社会。前者多半是科学实验数据或传感数据，后者与人的活动有关，特别是与互联网有关。

思考活动

未来畅想：大数据环境下的学校班级管理

10年后的一天，李紫木老师收到一份学生名单，新学期她将担任50名学生的辅导员。这份名单由学校的大数据管理系统自动生成。

李老师点击了“张超”，除了性别、年龄、家庭住址、联系电话、照片等基本信息外，她还看到了该生初中阶段和小学阶段的学习记录，除了学科成绩，还有很多学习过程的评价和记录，如该生函数方程的学习较好、平面几何的学习较弱等。在“爱好”栏里，记录着该生喜欢玩“宇宙探险”游戏和观看电视节目“科学发现”，建议高中阶段推荐他加入学校空间科学方面的社团。

李老师还了解到该生是过敏性体质，在每年柳絮飘飞时节需要在家学习，并通过学校的在线学习系统参与课堂讨论。李老师继续点击了照片，该生的三维立体图像瞬间生成，犹如看到学生站在自己面前。

交流讨论：大数据对你将来的个人生活和学习会产生怎么样的影响？请展开想象，与同学交流自己的想法。

微信公众号：电子课本大全

1.1.3 数据科学

云计算、物联网、移动计算等新技术的兴起拓展了人们有关数据获取和数据计算的能力，促使大数据时代的到来，同时成为数据科学兴起的必要条件，并进一步推动了数据科学的发展。

数据科学的兴起

1974年，计算机科学家彼得·诺尔（Peter Naur）在自己的一部著作中首次明确提出了数据科学的概念："数据科学是一门基于数据处理的科学"。此后直到2001年贝尔实验室的克利夫兰（Cleveland）发表论文，主张数据科学是统计学的一个重要研究方向，数据科学再度受到统计学领域的关注。2013年，马特曼（Mattmann）和达尔（Dhar）发表论文，从计算机科学与技术视角讨论了数据科学的内涵，使数据科学被纳入该领域的研究范畴。与此同时，数据科学逐渐进入实际应用，如模拟与仿真、集成学习、视频与图像分析、文本分析、语音分析、模型管理、自然语言问答等。

数据科学的内涵

2010年，德鲁·康威（Drew Conway）提出了数据科学的维恩图（图1.1.10），首次明确了数据科学的学科定位：数据科学处于统计学、机器学习和领域实物知识的交叉处，是一门交叉型的新兴学科。图中的"黑客"（Hacker）并不是指"骇客"（Cracker），"黑客精神"是指"大胆创新、喜欢挑战、勇于创新、追求完美和不断进取"的积极精神。



图1.1.10　德鲁·康威的数据科学维恩图

目前，学术界对数据科学的内涵基本达成共识：数据科学是以数据为中心的科学。朝乐门博士所著的《数据科学》一书中对数据科学有以下阐述。

·将"现实世界"映射到"数据世界"之后，在"数据层次"上研究"现实世界"的问题，并根据"数据世界"的分析结果，对"现实世界"进行预测、洞见、解释或决策的一门新兴科学。

·以"数据"尤其是"大数据"为研究对象，并以数据统计、机器学习、数据可视化等为理论基础，主要研究数据加工、数据管理、数据计算、数据分析和数据产品开发等活动的一门交叉性新兴学科。

从研究目的看，数据科学是将数据转化成信息、知识或智慧的

微信公众号：电子课本大全

过程。这一转变过程是一种从不可预知到可预知的增值过程，即数据通过还原其真实发生的背景成为信息，信息赋予其内在含义之后成为知识，而知识通过理解转化成智慧。

数据科学横跨计算机科学与技术、信息学、数学、社会科学、网络科学、系统科学、心理学、经济学等诸多领域。从计算机科学与技术角度看，数据科学研究主要包括数据加工、数据计算、数据管理、数据分析和数据产品开发等方面以及数据科学的基础理论（见图1.1.11）。



图1.1.11　数据科学的知识体系

数据加工。为了提升数据质量、降低数据计算的复杂度、减少计算量并提升数据处理的精准度，数据科学需要对原始数据进行一定的加工处理，如数据审计、数据清洗、数据变换、数据集成、数据脱敏、数据归约和数据标注等。值得一提的是，与传统数据处理不同，数据科学中的数据加工更强调数据处理中的增值过程，即如何将数据科学研究人员的创造性设计、批判性思考和好奇性提问融入数据的加工活动之中。

数据计算。在数据科学中，计算模式发生了根本性变化——从集中式计算、分布式计算、网格计算等传统计算过渡到云计算。比较有代表性的有GFS、BigTable、MapReduce、Hadoop MapReduce、Spark等。

数据管理。完成数据加工和计算之后，还需要对数据进行管理与维护，以便进行数据分析以及数据的再利用和长久存储。在数据科学中，数据管理方法与技术也发生了重要变革，出现了一些新兴的数据管理技术，如NoSQL、NewSQL技术等。

微信公众号：电子课本大全

数据分析。数据科学中采用的数据分析方法具有较为明显的专业性，通常以开源工具为主。目前，Python语言和R语言已成为使用较为普遍的数据分析工具。

数据产品开发。这是数据科学与其他科学的主要区别。与传统产品开发不同，数据产品开发具有以数据为中心、多样性、层次性和增值性等特征。数据科学的研究目的之一就是提升数据产品的设计与开发能力。

阅读拓展

数据科学和机器学习工具调查

数据科学网站 KDnuggets 发布了 2019 年数据科学和机器学习方面的工具调查结果。数据显示，计算机科学领域使用最多的是 Python 编程语言。R 语言的使用率自 2018 年第一次降到了 50% 后，2019 年继续下降（图 1.1.12，数据来源 KDnuggets）。可以看出，Python 颇受欢迎，RapidMiner 的热度仍然保持在 50% 以上，SQL 和 Excel 略有下降。

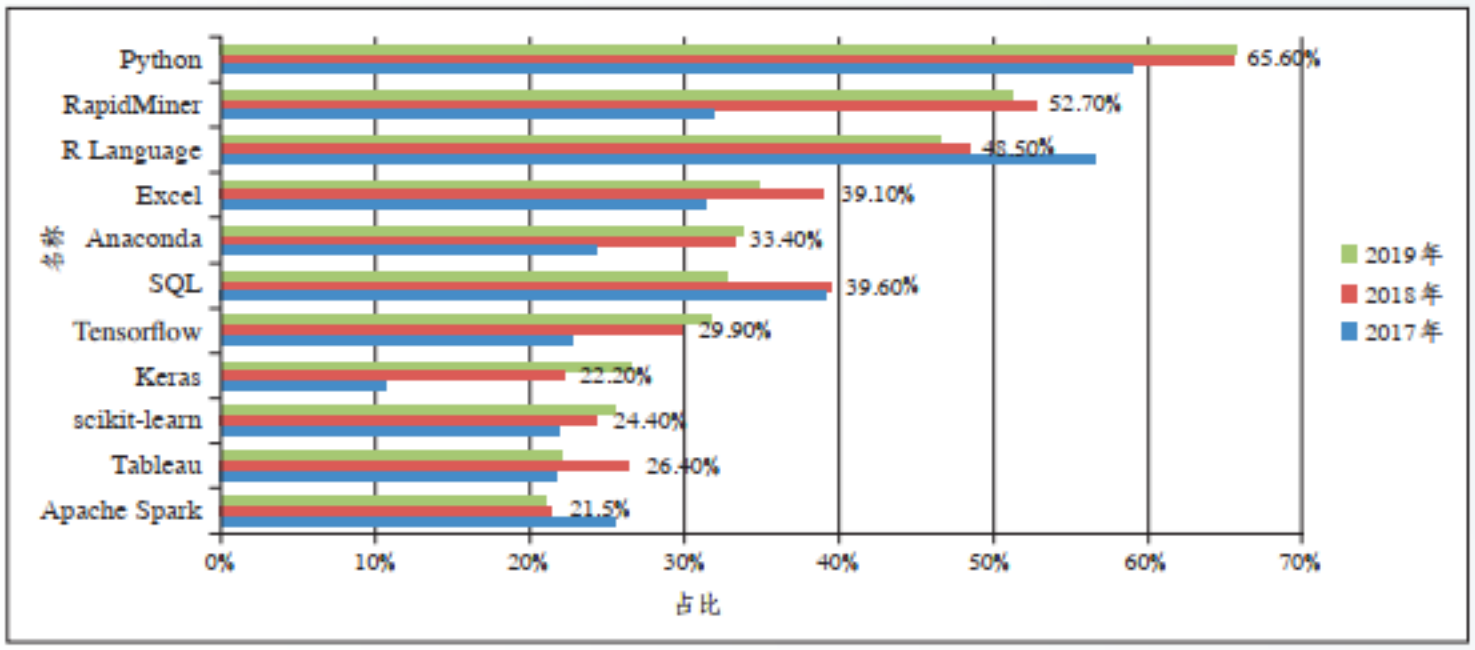


图 1.1.12　2017—2019 年数据科学和机器学习工具使用率调查

另外，该网站的其他数据表明，在深度学习方面，最近关注度较高的深度学习框架 PyTorch 的使用率比去年有所上升，但仍然落后于 TensorFlow 和 Keras。

数据科学作为一门与领域知识和行业实践高度交融的学科，从目前的研究来看，主要包括两个层面：用数据的方法研究科学和用科学的方法研究数据。

用数据的方法研究科学。主要指以数据为中心来开展各学科的

微信公众号：电子课本大全

研究，如基因组学、蛋白组学、天体物理学、脑科学、生物信息学、地球环境学等研究。随着数据科学相关技术的发展，越来越多的科学研究将直接针对数据展开，人类通过认识数据，进一步认识自然和社会。与此同时，这些学科的研究又产生了更多的数据。例如，用电子显微镜重建大脑中的突触网络，1 mm^3 大脑的图像数据就超过 1 PB，处理这些数据需要数据科学相关技术与方法的支持。未来，各个学科领域还将形成相应的数据科学研究理论与方法。

用科学的方法研究数据。主要指选用科学的方法来研究数据的采集、存储、加工、管理、分析、可视化等问题。例如，当要处理的数据量巨大、给计算带来挑战时，需要随机方法或分布式计算来解决问题。当错误或异常数据较多、给数据分析带来困难时，需要有一定修正功能的数学、统计学等模型来进行处理。

思考活动

大数据用于交通监测

近年来，许多城市应用智能交通，利用车载全球卫星导航系统，实时记录正在路上行驶的汽车的位置、方向和速度等数据。当大量数据被融合在一起后，由研究人员对样本数据进行计算、汇总与分析，从而推算出当前道路的平均通行速度和可能的突发事件，并根据道路通行速度的高低，将路况标注为红色、黄色和绿色三个等级，即严重拥堵、出现拥堵和道路畅通（图 1.1.13）。

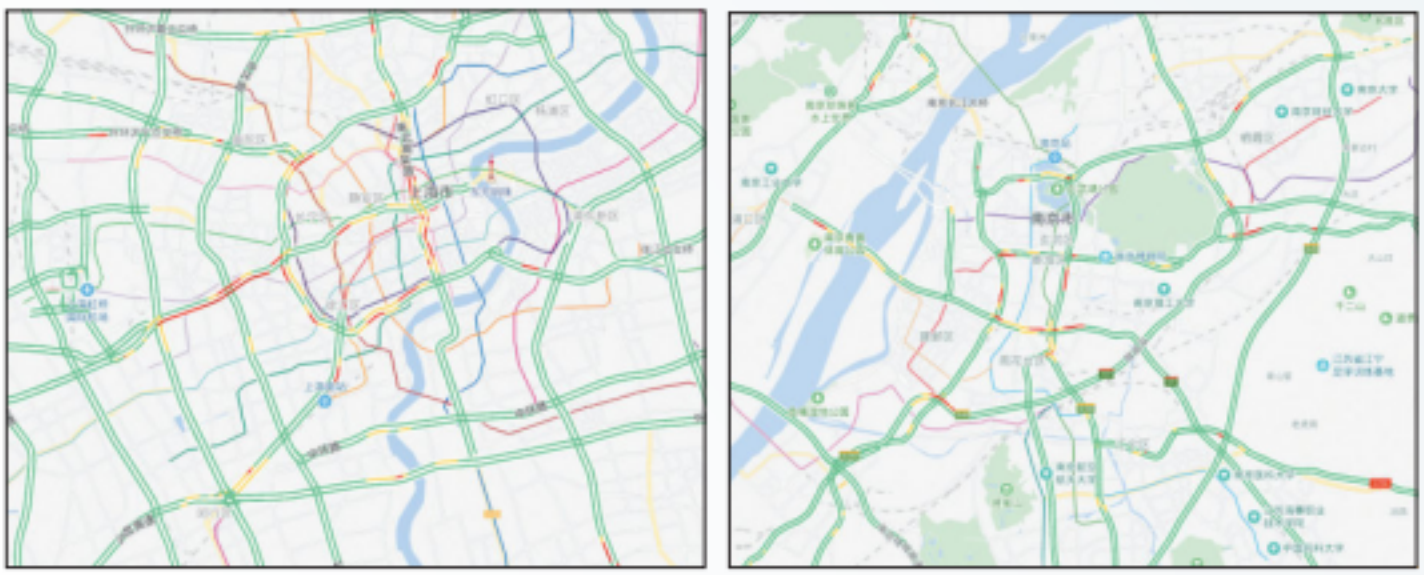


图 1.1.13　电子地图的交通路况监测

讨论、思考以下问题。

1. 交通监测中需要采集汽车定位系统的哪些数据？如何存储和管理这些数据？
2. 大数据在智能交通中起什么作用？如何更有效地挖掘大数据的应用价值？

微信公众号：电子课本大全

项目实施

了解数据科学对送货机器人研制的作用

小组成员一起协作，完成项目的调研与分析，然后进一步认识、了解送货机器人与数据科学的关系。

1. 小组讨论：目前送货机器人有哪些优势与局限性？

2. 网络调研与分析：数据科学给送货机器人的研制带来哪些帮助？制作一个演示文稿，并尝试利用思维导图软件，展示调研与分析成果。

练习提升

1. 结合身边大数据的应用案例，思考大数据及其技术的价值体现。

2. 数据科学涵盖哪些知识体系？哪些学科与数据科学相关？

3. 当前，从事哪些专业领域的工作需要学习数据科学的相关知识？

4. 数据、信息、知识与智慧的转化关系，对你的学习有哪些启示？

5. 智能同步教学平台是目前广受关注的应用工具。阅读以下有关介绍，思考这样的智能平台给学习带来了哪些好处？你所期待的智能教学平台是怎样的呢？

> 智能同步教学平台利用云计算平台，深度挖掘教学大数据潜力，通过语音识别和评测、自然语言处理等人工智能技术，对教师的教学结果和学生的学习行为进行记录、存储、统计、分析和预测，教师可以根据这些数据调整教学思路、教学设计以及教学方法，让课堂教学从“预设性教学”向“生成性教学”转变。同时，在持续反馈和不断调整的过程中，教师可以对学生进行连续测评，分析学生学习行为及学习方式的变化，作为后续教学设计时的依据，使课堂教学管理从经验型向数据型、智能型、科学型转变，最终实现针对学生的个性化教学。

6. 你观看过哪些科幻电影或纪录片？它们呈现了哪些与数据管理和数据分析相关的情境？请与小组同学交流看法。

微信公众号：电子课本大全

1.2
数据管理与分析简介

学习目标 ▶▶▶

- 了解数据管理的发展阶段及数据管理方式。
- 认识大数据的存储与管理。
- 了解数据分析的内涵、工具以及基本过程。
- 感受数据分析对科学决策的作用和意义。

体验探索

让数据变得更有价值

美国亚马逊公司从用户购买行为中获得数据，如用户在页面的停留时间、是否查看评论、搜索的关键词、浏览的商品等。分析人员通过数据分析，发现潜在的购买行为，然后进行专门的方案设计，让数据发挥应有的价值。默克公司创建了制造和分析智能系统，旨在使非技术性业务分析人员能够在可视化软件中直观地浏览和查看数据。专业人士表示，他们花很少的时间进行数据的移动和报告，但会花更多的时间使用数据来获得有意义的成果。我国很多金融机构很早就开始聘请精通数据分析的专家来设计金融产品。国际商业机器（IBM）公司在全球聘请了很多数学家，旨在把他们数据分析的才能应用于石油勘探、医疗健康等各个领域。易贝（eBay）公司通过数据分析，精确计算出广告中的每个关键字，优化广告的投放，大幅降低了产品销售的广告费用。

思考讨论：数据管理与分析技术的发展，对数据价值的发现起到了什么作用？

数据隐含着巨大的社会、经济、科研价值，如果能被有效地管理、分析和利用，将对社会、经济和科学研究产生积极的推动作用，给社会发展带来前所未有的机遇。然而，数据本身并不会自动产生价值，价值是需要通过专业的管理和分析才能被挖掘出来的。因此，数据管理与分析是使数据变得有价值的重要原因。

微信公众号：电子课本大全

1.2.1 数据管理的发展

数据管理主要指对数据进行分类、组织、编码、存储、检索、维护和应用，它是数据处理的核心环节，其目的在于充分挖掘数据的价值并有效地利用。从信息技术应用与发展的角度看，数据管理经历了人工管理、文件系统、数据库系统等发展阶段。

人工管理

20世纪50年代中期以前，计算机主要用于科学计算。硬件方面，计算机的外存只有磁带、卡片、纸带，没有磁盘等直接存取的存储设备，存储量非常小。软件方面，没有操作系统，计算机一次处理一批数据，直到运算完成才能进行另外一批数据的处理，中间不能被打断，原因是此时的外存（如磁带、卡片等）只能顺序输入。这一阶段的数据管理具有以下特点。

没有专门的应用软件来管理数据，而是由调用数据的程序自行管理。数据往往作为程序的组成部分，即程序和数据是一个不可分割的整体，数据和程序同时提供给计算机运算使用。

数据不具有独立性。程序依赖于数据，当数据类型、格式或输入输出方式等发生变化时，就必须修改相应的程序。

数据不能共享。由于数据与程序关系紧密，往往是一组数据对应着指定的一组程序（图1.2.1），因此程序中的数据原则上无法与其他程序共享使用。

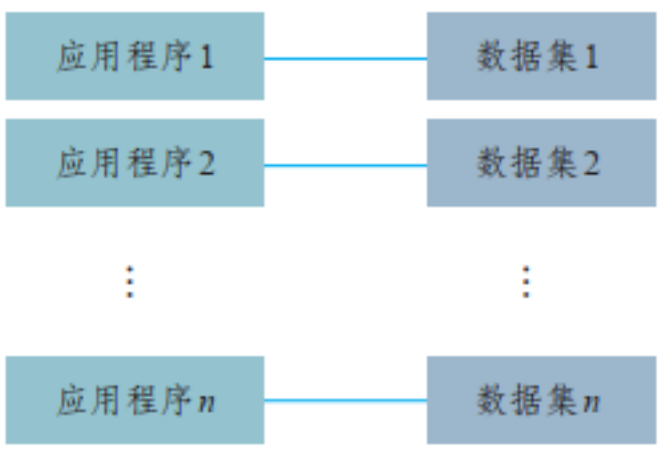


图1.2.1 数据与程序的直接对应关系

思考活动

人工管理数据的主要弊端

在人工管理数据阶段，你认为还存在哪些主要弊端？与周围同学交流自己的看法，并写出几点。

微信公众号：电子课本大全

文件系统

20世纪60年代，随着计算机技术的发展，数据管理发展进入文件系统阶段。此时计算机有了磁盘、磁鼓等直接存取的外部存储设备，操作系统中有了专门管理数据的文件系统。从处理方式上看，能够联机实时处理，即在需要时随时从存储设备中查询、修改、更新数据。

文件系统管理数据具有以下特点。

数据可以长期保存在外部存储介质上并能反复使用。即数据可以进行查询、修改和删除等操作。

数据具有了一定的独立性。但数据文件仍然依赖于指定的程序，一个文件基本对应一个应用程序。

数据共享性差。当不同的应用程序所需的数据有部分相同时，仍需建立各自独立的数据文件，而不能共享这些数据（图1.2.2），致使数据冗余度较大，浪费存储空间。

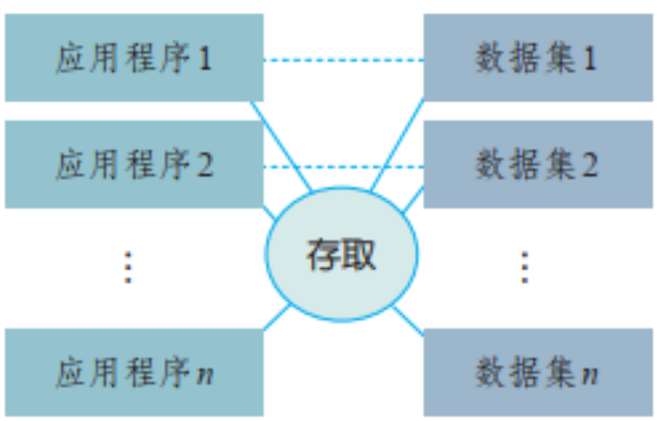


图1.2.2　应用程序与数据之间的存取关系

阅读拓展

Windows系统的文件系统

随着操作系统的发展，Windows系统分别出现了FAT16、FAT32、NTFS等类型的文件系统。不同的存储设备也使用不同的文件系统，如光盘中采用的是CDFS，闪存使用的是exFAT。

其中，NTFS采用了精细的技术，能更有效地管理磁盘空间（图1.2.3）。在NTFS分区上，可以为共享资源、文件夹以及文件设置访问许可权限。与FAT32文件系统对文件夹或文件的访问管理相比，安全性更高。

根据以上提示，同学们可以查看一下自己计算机中操作系统所采用的文件系统类型，从而加深对文件系统的了解。



图1.2.3　个人计算机磁盘的文件系统

微信公众号：电子课本大全

数据库系统

从20世纪60年代后期开始，计算机的应用逐步展开，数据量快速增长。各种应用、不同程序语言互相包容的数据共享要求日益迫切，以文件系统作为数据管理的方式已经无法满足需要。为此，用于数据管理的数据库系统应运而生，而数据库管理系统是其一个组成部分。

利用数据库系统管理数据主要具有以下几个优势。

数据结构化。数据库系统实现整体数据的结构化，这是数据库系统与文件系统的本质区别。

数据易于共享。数据的共享大大降低了数据的冗余度，节约了存储空间，更便于扩展。

数据具有独立性。数据与应用程序相对独立，数据库中数据的物理存储结构与逻辑结构改变时，应用程序不必改变（图1.2.4）。

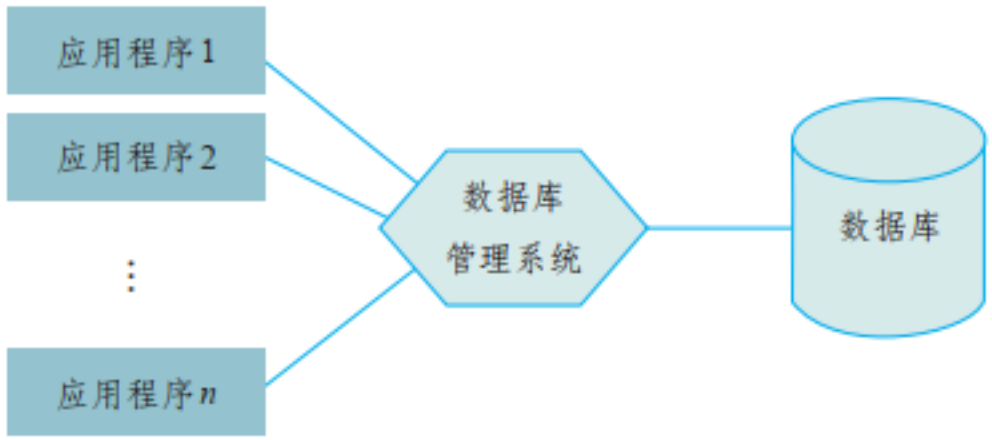


图1.2.4　应用程序、数据库与数据库管理系统

数据的安全性高。由数据库管理系统统一管理和控制数据，能够实现数据的安全性控制、数据的完整性控制及并发控制、数据恢复等功能。

实践活动

探究火车售票系统数据库

学校要组织学生去外地进行社会实践，老师请张晔同学一起组织这个活动。首先是网上购买火车票。张晔在购票过程中对火车售票系统的内部运行原理产生了兴趣。一直在思考：这个系统中这么大量的数据，是如何保存和管理的呢？

请你和张晔同学一起探究以下两个问题。

1. 火车售票系统需要存储和管理哪些数据？请举出一些例子。
2. 如果你想创建一个简单的数据库，需要做哪些基本的工作？

微信公众号：电子课本大全

1.2.2 大数据存储与管理

随着大数据的兴起，数据管理与分析的相关技术也在快速发展，如数据管理过程中的数据采集、存储、加工、转换和传输等技术，数据分析过程中的数据组织、计算、检索、统计等技术。下面介绍大数据的主要存储技术与方式。

对于大量结构化、半结构化和非结构化数据，关系数据库已无法满足存储以及复杂的数据挖掘和分析的要求，目前通常采用分布式文件系统、非关系数据库、云数据库等对这些数据进行存储与管理。

分布式文件系统

分布式文件系统有效解决了大数据时代数据存储和管理的问题，它将固定于某个地点的某个文件系统扩展到任意多个地点或多个文件系统，众多的节点数据块组成文件系统的数据网络，分布在不同地点的多个节点通过网络对数据进行传输。人们使用分布式文件系统时，不需要关心数据存储在哪个节点上或从哪个节点获取，只需像使用本地文件系统一样进行操作。

随着数据规模的增长，系统只需在节点集群中增加更多的数据节点即可，具有很强的可扩展性。数据的分布式存储可以提高大数据量的并行访问能力与计算能力，适应了当前大批量数据存储与管理的需求。分布式文件系统示意图如图 1.2.5 所示。

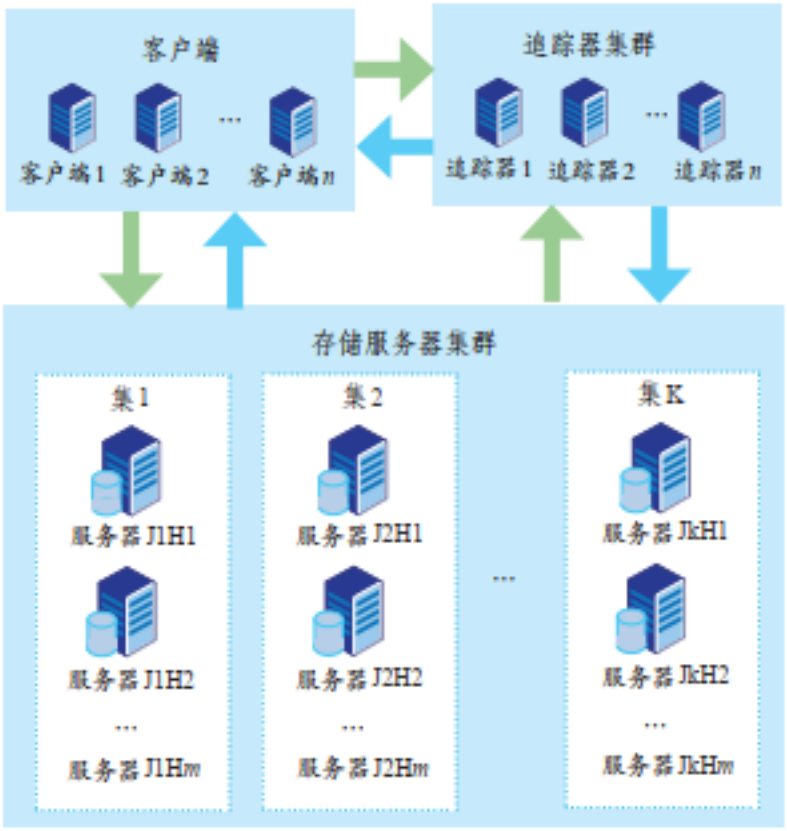


图 1.2.5　分布式文件系统示意图

非关系数据库

非关系数据库是对关系数据库局限性的补充，通过放弃部分复杂处理能力的方式，支持将数据分散存储在不同服务器上，解决了关系数据库在大量数据写入操作上的瓶颈。通过采用缓存技术较好地支持对同一个数据的频繁处理，提高了数据简单处理的效率。同时，遵循“数据在先，模式在后”的设计方式，设计出来的数据模型可以很好地支持网络应用。其常用数据模型主要有四种（表 1.2.1）。

微信公众号：电子课本大全

表1.2.1　非关系数据库的四种主要数据模型对比

	键值	列存储数据库	文档型数据库	图形数据库
含义	特定的键与值之间采用散列表等建立映射关系	以列为单位存储，并将同一列数据存放在一起。键指向多个列，这些列由列簇来安排	与“键值”类似，但其中的值指向结构化数据	以图形的方式存储数据
应用场景	大量数据的高访问负载、日志系统	分布式文件系统	万维网应用	社交网络、推荐系统和关系图谱
优点	查找速度快	可扩展性强、易进行分布式扩展	不需要预先定义结构	
不足	数据无结构	功能相对有限	查询性能不高	

云数据库

云数据库是基于云计算技术发展起来的一种共享基础架构的存储方法，主要指被优化或部署到一个虚拟计算环境中的数据库。例如，把一个现有数据库优化到云环境中后，可以使用户按照存储容量和带宽需求付费使用，可以将数据库从一个地方移到另一个地方（云的可移植性），可以实现按需扩展等。

云数据库并非一种全新的数据库技术，而只是以服务的方式提供的数据库存储、计算与管理功能。可为用户提供数据备份与恢复、安全管理、监控与消息通知、故障自动切换等服务支持。大型企业将分散的多个数据库部署到云，还可以在云环境中整合成一个数据库管理系统（图1.2.6），实现存储整合，从而推动数据资源共享。

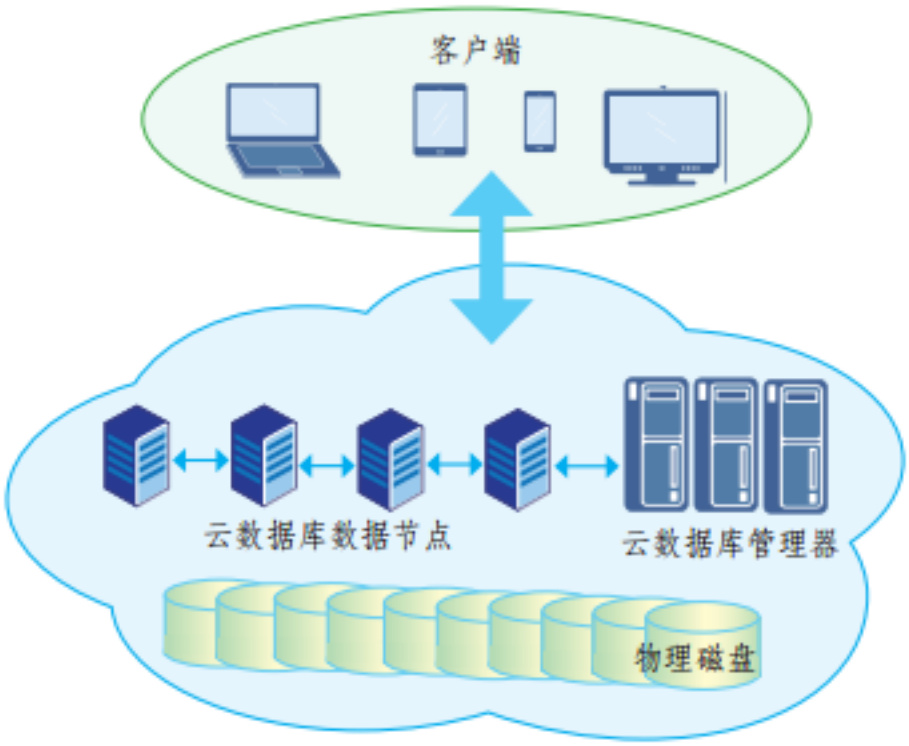


图1.2.6　云数据库管理系统示意图

微信公众号：电子课本大全

1.2.3 数据分析及其基本过程

数据分析主要指用适当的统计分析方法对所获取的数据进行比较、筛选、梳理，提取有用信息，形成结论，并对数据进行深入研究和概括归纳的过程。其目的是把隐藏在看似杂乱分散数据中的信息提炼出来，从而发现研究对象的内在规律。数据分析的结果可帮助人们做出对一些事物的判断或对下一步行为的决策。例如：约翰尼斯·开普勒（Johannes Kepler）通过对观测数据的分析，获得了行星的运动定律；某企业领导通过市场调查、分析获得数据，判定市场动向，并进一步制订合理的生产与销售计划。

数据分析工具有多种。电子表格软件和数据统计软件是人们日常使用的数据分析工具。用电子表格软件分析数据的操作比较简单，绘制图表的功能也很便捷，但能够分析的数据量有限，比较适合数据量较小的场合。SPSS是IBM公司推出的数据统计软件，由一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的功能及相关服务组成。只需掌握一定的操作技能，了解基本的统计分析知识，就可以使用该软件进行数据分析与研究。

此外，还有一些专用的数据分析工具，例如，在商业智能领域常用的Cognos、Style Intelligence、Microstrategy、Brio、Oracle和国产的Yonghong Z-Suite BI套件等。

思考活动

统计图表与数据分析工具

观察你能找到的统计图表，想一想这些图表可能采用了哪些数据分析工具？再调研身边的人或单位，了解他们都使用了哪些数据分析工具来制作统计图表？

数据分析，主要是指对已采集回来的、经过一定处理的现成数据进行分析，但实质上，在数据处理任何环节中都会发生数据分析，无法严格界定。

根据数据科学的方法论，一个数据分析项目的团队在进行数据分析时，基本过程通常可以用图1.2.7来表示。

了解目标问题。首先要充分了解需要解决的问题，包括描述项目问题，提供数据集，确定项目目标等。

数据获取。有些数据可能来自项目内部，但用于分析的数据最好不要局限于项目本身的数据，可以包含来自外部的各种数据（如互联网数据），并尽可能多地获取有用的数据。

数据整理与清洗。规范、严谨的数据整理是后续工作的保障，

微信公众号：电子课本大全

将原始数据清洗并转换为适合机器学习的形式也很关键。同时，清洗转换过程中做好记录日志，以备数据再利用时参考。

数据计算与统计。主要指用统计方法和数据可视化技术来尝试发现数据中潜在的特征和发展趋势。这一过程往往需要对数据进行深入、反复的研究和探索，以免遗漏、错过重要的特征和线索。

数据建模及其应用。主要指选择适于解决问题的机器学习算法，对多种机器学习类型进行测试，从而筛选出适合用于特定应用项目的算法。事实上，一种算法对特定的数据可能最有效，而另一种算法在其他数据上则表现更好，选择最佳算法是数据分析项目实践中最具挑战性的一个环节。

数据拓展及数据可视化。数据分析的最终结果往往是一份具有一定拓展的数据报告，并通过精心设计可视化作品来获得最佳的呈现效果。设计制作可视化作品需要根据数据分析结果，以生动直观的形式将数据所表达的意义呈现出来，因此要求制作者具备一定的创造思维和艺术修养。

解决目标问题。数据经过上述处理后，获得了预设信息，而这些信息需要进一步进行数据挖掘、价值评估，转化为有效的预测和决策，这个过程有时还需要再次进行数据分析。

目前，基于大数据进行数据分析时，分析的不是样本数据，而是所有数据，即采用的不是传统的抽样模式，而是全数据模式。需要采用新的分布式系统架构，把大规模数据变成小规模数据，分配给数台机器进行处理。



图 1.2.7　数据分析的基本过程

阅读拓展

数据分析工具的演进

近年来，数据分析工具出现了一些新的变化，新的分析平台在不断演进。以大数据应用为例，分析平台已从早期的 Java + Hadoop 等逐渐过渡到 R、Python、 Scala、Java + Hadoop 等。同时出现了包括一些流式数据的处理和分析的方案，如 Storm、Kafka、Flume 等工具的应用。在数据库方面，也由传统的关系数据库扩展到了非结构化数据库。数据管理与数据分析的应用更为紧密，往往可以在同一平台上，利用集群快速处理、计算和分析。

微信公众号：电子课本大全

近年来，数据分析技术在两个方面取得了突破。一是对体量庞大的结构化和半结构化数据进行高效率的深度分析，挖掘隐性知识，如从自然语言构成的文本网页中理解和识别语义、情感、意图等；二是对非结构化数据进行分析，将复杂多源的语音、图像和视频数据转化为机器可识别的、具有明确语义的信息，进而提取有用的知识。



阅读拓展

机器学习领域的深度学习

深度学习是近年来机器学习领域的重要研究方向，并在语音识别、图像识别、自然语言处理等方面获得突破。当前许多互联网科技企业纷纷对深度学习进行布局，并取得一些成效。例如，我国腾讯公司研发的Peacock大规模主题模型机器学习系统，能通过并行计算对大规模矩阵进行分解，从海量样本数据中学习10万～100万量级的隐含语义，这对于挖掘用户兴趣、扩展相似用户、进行精准推荐具有重要意义。另外，阿里云机器学习平台的丰富算法功能可以在线使用，数据和计算资源一直处在“在线”状态，降低了人们使用机器学习知识的门槛。



思考活动

大数据分析要解决的首要问题

对大数据的分析，有研究人员认为，目前首先需要解决以下三个问题。

· 可表示问题。当前互联网中非结构化数据的比例大幅增加，专家预计5年内数据会增加8倍，如何有效地表示这些非结构化数据成为一个首要的问题。

· 可处理问题。如今数据规模急剧扩张，远远超越了现有计算机的处理能力。大数据的高效处理已经成为一个核心问题，而数据处理的不同阶段，其处理形式不同，因此需要将计算科学与数学、物理等学科相结合，建立一种新型数据科学方法，以便在数据多样性和不确定性前提下进行数据规律和统计特征的研究。

· 可靠性问题。大数据管理系统在数据采集时往往存在一定的缺陷，容易造成数据中含有各种各样的错误和误差。因此，一方面要通过数据清洗、去冗余等技术提取有价值的数据，实现数据质量高效管理；另一方面要实现对数据的安全访问和隐私保护。这两方面已成为大数据可靠性的关键需求。

思考：结合前面的学习，你如何理解以上这几个问题？你认为还存在哪些需要解决的问题？

微信公众号：电子课本大全

1.2.4 数据分析助力科学决策

阅读拓展

交易记录分析与预测案例

很多大型零售公司的数据库都会记录每个客户的购物清单及消费额，包括购物车中的物品、具体购买时间，甚至购买当日的天气情况。这些公司会对“历史交易记录”这个庞大数据库里的海量交易数据进行分析，挖掘新的业务增长点。例如，一家商品零售公司发现，每当在季节性飓风来临之前，不仅手电筒销量增加了，而且蛋挞的销量也增加了。因此，每当季节性风暴来临，该公司就会把库存的蛋挞放在靠近飓风用品的位置，以节约客户的购物时间。

以上材料让我们感受到数据分析对商业决策的重要作用。事实上，数据分析的目的就是利用数据进行科学决策。人类越来越依赖数据分析进行决策。数据分析对决策的支持主要体现在以下几点。

提升决策的准确性。科学发展观要求按科学发展的规律办事。数据是科学的基础，也是科学的度量标准。数据挖掘、数据分析和可视化技术为快速、准确地做决策提供了数据支撑。随着大数据分析技术的不断成熟，数据分析极大提升了社会各领域决策的能力和决策的准确性，使决策越来越靠近科学发展的规律。例如，在医疗领域，我国部分省市正在实施病历档案的数字化，配合临床医疗数据与病人体征数据的收集分析，可以用于远程诊疗、医疗研发，甚至可以结合保险数据分析用于商业及公共政策制定等。

优化管理决策。管理决策是一个需要不断优化的过程，决策过程需要大量数据的智能辅助。管理者在掌握大量数据和信息后，借助数据分析技术，通过客观、理性的逻辑分析和经验判断，做出决策并在实施过程中不断优化、调整决策。科学的决策优化需要大量的历史数据、即时数据和关联数据，而大数据环境为分析这些数据创造了条件。随着人工智能技术的发展，还可以让智能系统帮助人们完成动态监测、趋势判断、语音咨询、即时翻译乃至医疗诊断等。

对决策结果进行模拟。大数据分析模型还能够对决策的结果进行模拟或仿真效果呈现，帮助决策者有针对性地改进决策的整体方案和细节。例如，交通管理部门往往需要对拥堵路段进行模拟和预判，分析产生拥堵的原因，以便通过调整交通控制系统疏通拥堵路段。我国一些城市安装了自适应交通控制系统，这些系统能够根据路口的车流数据、人流数据、地理位置和监控摄像头传来的数据，

微信公众号：电子课本大全

自动调整红绿灯持续的时间，实现对交通流量的实时配置和控制。

实时反馈数据。决策实施过程中，往往会因为某个影响因素的改变或新要素的加入，导致决策的结果产生偏差，因此需要实时的数据反馈来调整决策，从而及时把握事件发展的趋势，发现新的问题。决策不是孤立的，而是相辅相成的，一个决策实施的数据反馈往往会成为另一个决策的依据。



思考活动

大数据时代的数据分析

现在，人们可以获得更多的数据，甚至是与之相关的所有数据，而不再依赖于采样，从而能更全面地认识事物，并发现样本无法揭示的规律。人脑之所以具有智能、智慧，一个重要原因就在于它能够对周边的数据信息进行全面的收集、逻辑判断和归纳总结，获得有关事物或现象的认识与见解。《大数据时代》一书的作者维克托·迈尔－舍恩伯格（Viktor Mayer-Schönberger）指出："执迷于精确性是信息缺乏时代和模拟时代的产物。只有5%的数据是结构化且能适用于传统数据库的。如果不接受混乱，剩下95%的非结构化数据都无法利用，只有接受不精确性，我们才能打开一扇从未涉足的世界的窗户。"他认为，大数据的出现让人们放弃了对因果关系的渴求，转而关注相关关系，人们只需知道"是什么"，而不用知道"为什么"。

大数据时代带来了数据分析方法和模式的变革，也引发了思维方式的转变。

思考讨论：你如何理解舍恩伯格的观点？在日常学习和生活中，我们需要在哪些方面转变思维方式？需要储备哪些数据分析的知识，才能适应时代的发展？



项目实施

认识数据管理与分析对送货机器人的作用

1. 查阅资料：了解数据分析师这个职业的职责以及所涉及的专业知识。

2. 小组讨论：选取送货机器人中的某一类型的数据（如货物数据、道路数据或客户数据等），通过数据分析，可以从这些数据中获得什么价值？

3. 思考：在数据管理与分析领域以及送货机器人的研发、设计与制造中，存在哪些工作机会？从事这些工作需要具备哪些学科知识？

4. 制作一个思维导图，简要描述数据管理与分析对送货机器人开展工作所起的主要作用。

微信公众号：电子课本大全

练习提升

1. 结合实例和自己的理解，简述数据的价值是如何得以体现的。

2. 简述数据分析的基本过程，其中你觉得最重要的是哪个环节？请说出你的理由。

3. 每年高考录取工作结束后，经常有一些机构会对高考人数、分数、分数线、院校招生人数和招生专业等数据进行分析，并得出很多信息。请查阅相关资料，了解你所感兴趣院校的各项数据，并列出对你有价值的信息。

4. 案例数据分析。

张越是一位高三的学生。很多高校都在网上公示了各个专业规定的选考科目。现在，张越想对这些招生数据进行以下整理和分析：将每所高校的选考科目信息全部汇总到同一张表格中；对自己喜欢的物理、技术和历史三个科目，使用上述数据进行分析汇总，获得适合的大学和专业推荐。

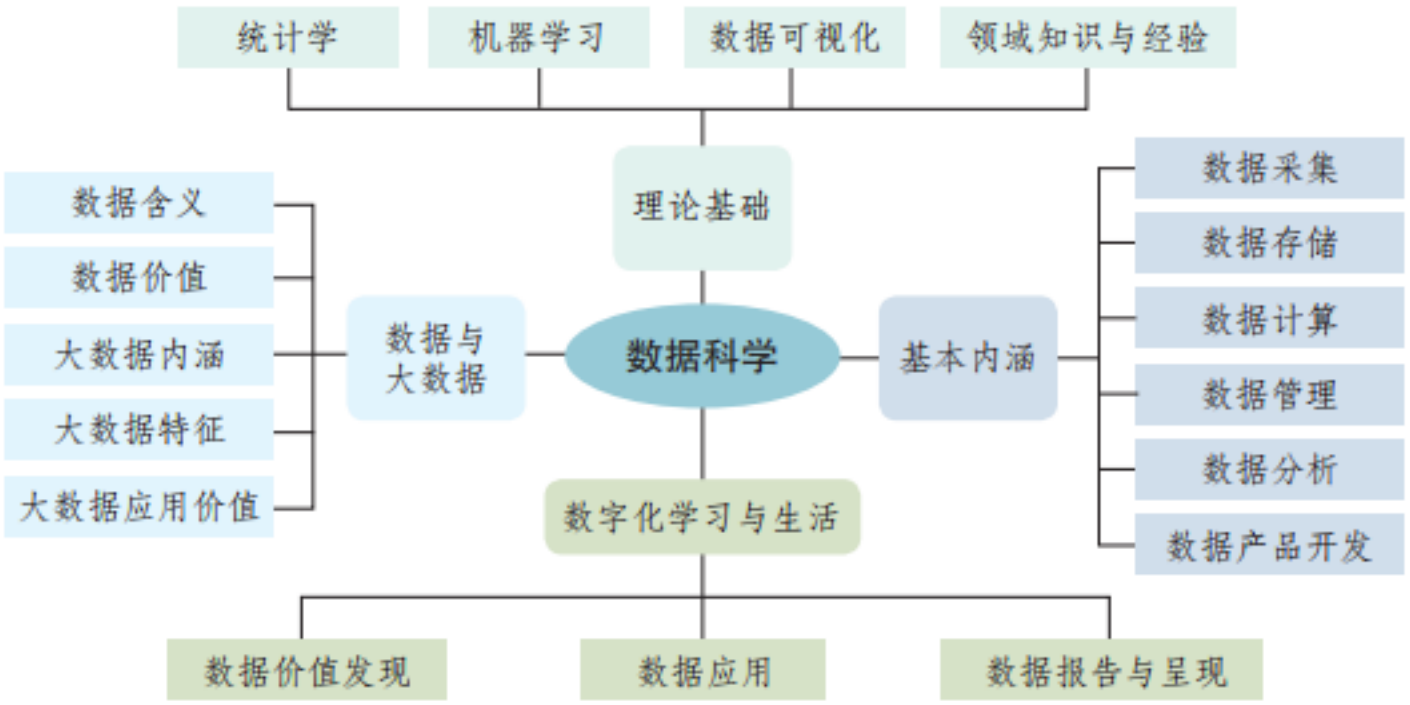
结合张越的这些想法，你觉得数据管理与分析的作用体现在哪些方面？

5. 小宇的团队决定就“更合理地设定收费与用时之间的关系，从而提升共享单车的使用率”这个问题进行调研分析，请帮他们列出一个大体的调研计划和要调研的数据。

微信公众号：电子课本大全

总结评价

1. 下图展示了本章的核心概念与关键能力，请同学们对照图中的内容进行总结。



2. 根据自己的掌握情况填写下表。

学习内容	掌握程度		
数据与大数据的含义	□ 不了解	□ 了解	□ 理解
数据、信息、知识与智慧的关系	□ 不了解	□ 了解	□ 理解
大数据的基本特征及其应用价值	□ 不了解	□ 了解	□ 理解
数据科学的内涵	□ 不了解	□ 了解	□ 理解
数据管理的发展阶段和管理方式	□ 不了解	□ 了解	□ 理解
大数据的存储与管理	□ 不了解	□ 了解	□ 理解
数据分析的基本过程	□ 不了解	□ 了解	□ 理解
数据分析对科学决策的作用	□ 不了解	□ 了解	□ 理解

3. 思考以下问题，完成学习过程的反思。

（1）美国数据科学领域专家希拉里·梅森（Hillary Mason）认为：“数据科学家是懂得获取、清洗、探索、建模、解释数据的人，还要融合入侵技术、统计学和机器学习。他们不仅要处理数据，还要把数据做成一个五星产品。”你如何理解这句话？请与同学交流看法。

（2）在本章的学习过程中，你遇到了什么问题？是如何解决的？你觉得自己有什么样的收获？尝试列举几点与同学分享。

微信公众号：电子课本大全



第 2 章
需求分析与数据采集

数据的表现形式多种多样，要从大量的数据中获取信息，首先需要收集并存储这些数据，在使用数据前对这些数据进行必要的预处理及数据格式化（如把非结构化的数据转换成结构化的数据），然后进一步处理，从而去除冗余及噪声等，最后保存并备份数据以备使用。

那么数据是如何被采集和处理的呢？可以采用哪些方法进行采集和处理呢？本章就一起来探讨这些问题。

微信公众号：电子课本大全

02 主题学习项目：交通数据见发展

项目目标

一个地区的交通运输数据往往能很好地反映该地区的经济发展情况。本章的项目学习，通过采集几个城市的交通运输数据，从数据中了解城市经济发展的情况，认识数据采集的作用与意义。

1. 通过分析交通运输数据采集的需求，制订采集方案，体验剖析问题、分析需求与制订问题解决方案的基本过程。

2. 理解数据采集的重要性，掌握数据采集的思路与方法。

3. 认识到数据预处理的意义，了解数据清洗的方法。

项目准备

为了完成项目，需要做以下准备。

- 4 — 6人组建小组，各组确定一名组长，按小组学习的方式展开项目活动。
- 开展项目时，小组成员应共同收集资料、讨论并确定数据需求，根据实际需要，小组分工协作完成数据采集、存储等事项，每个人应独立撰写数据采集报告的一部分。
- 强调数据保护意识，虽然获取的数据是官方公开的，但也要标明出处。同时强调，合法合规采用网络爬虫获取的数据，仅供学习使用。
- 搭建Python开发环境，简要回顾Python的常用语句以及编程的思路与方法。

项目过程

为了保证顺利完成本项目的学习活动，在不同学习阶段，小组长要注意检查组员项目学习的进度，并做好协调互助工作。

1 确定方案

组员一起分析项目主题，细化要研究的问题，然后制订初步解决方案，并对数据进行需求分析，撰写分析报告。 P39

2 采集数据

研究并选择最佳的采集数据的途径，然后组员分头采集几个城市的交通运输数据，这个过程中要进行必要的讨论交流。 P50

3 清洗数据

每人分析自己所采集数据中可能存在的噪声数据；共同讨论清洗数据的方法，合作编写程序，清洗每个同学采集的数据。 P60

项目总结

通过汇集组员学习心得，并对项目目标进行归纳和提炼，进一步理解数据需求分析在数据采集中的作用，同时感受利用网络爬虫技术获取网络数据的便捷性。最后通过展示和交流，提高数据需求分析和数据采集的能力。

微信公众号：电子课本大全

2.1
业务需求与解决方案

学习目标 ▶▶▶

- 能结合业务特点分析数据需求，并会撰写需求分析文档。
- 根据数据需求，设计合理的数据管理与分析方案。
- 学会通过不断优化，找到解决问题的最佳方案。
- 培养严谨的科学研究态度，提高解决问题的能力。

体验探索

是否可以开设共享汽车业务

某汽车租赁公司为了扩大业务，计划在一些城市开设共享汽车业务（图2.1.1）。在开设业务之前，公司需要进行最基本的需求分析。如果你是业务分析员，尝试根据表2.1.1列出的项目进行需求分析。



图2.1.1　共享汽车服务

表2.1.1　开设共享汽车业务需求分析

项　目	需　求　分　析
公司的目的	
用户的需求	
可行性（来自相关管理部门）	
可行性（来自人力和物力）	

思考：你认为业务需求分析中最关键的一项工作是什么？在进行需求分析时，是否需要采集大量的数据（如城市人口、汽车拥有量、车位数与每年接待旅游人数等）？

微信公众号：电子课本大全

业务从计划到实际开展，一般都要经历图2.1.2所示的几个阶段。



图2.1.2　业务分析的四个阶段

有了某项业务计划或产生某个想法后，最重要的一个阶段就是进行业务需求分析，然后根据分析结果建立解决方案并优化。

2.1.1　认识业务需求分析

业务需求分析就是要回答几个问题：要做什么？要达到什么目的？可行性是什么？

本书所说的“业务”指的是“要解决的问题”，业务需求分析就是针对该问题分析业务目标、用户需求和可行性等。明确需求后，要确定问题的解决方案，根据方案进行深入细致的调研和分析，从而获得数据，对数据进行有效管理和分析，最后获得问题的解决依据（或论证最初提出的观点和设想）。

例如，在前面的“体验探索”中，提出了问题“是否开设共享汽车业务”，那么进行业务需求分析时，可得到表2.1.2所示的分析结果。

表2.1.2　开设共享汽车业务需求分析

问　题	分析结果说明
要做什么（提出问题）	能否在××市开设共享汽车业务
业务目标	为人们提供便利的出行方式；扩大公司业务范围，提高经济效益
用户需求	用户希望该业务能提供什么样的服务，例如：①价格是否合适；②支付方式是否便利；③取车和停车是否方便；④对车型有何要求
可行性	①本市属于著名的旅游城市，游客数量很大，自驾需求潜力巨大；本市实施汽车限购政策，而本地居民有一定的自驾出行需求；共享方式符合很多用户的消费习惯 ②当地政府支持共享经济的发展，支持绿色出行的交通方式 ③公司资金充足、人员分工到位，能保证各项相关业务的开展

实践活动

对“餐厅提供地方特色小吃”进行业务需求分析

小李是某所大学的学生，他和几个同学都认为，学校应该有一个餐厅提供一些地方特色小吃，让学生在饮食上获得更多的选择，也能在一定程度上丰富对中国饮食文化的体验。于是，小李他们决定就这个问题展开调研，了解大部分同学是否存在这个需求。请仿照表2.1.2，帮助他们进行业务需求分析。

微信公众号：电子课本大全

从以上阐述中可以看出，在业务需求分析中，用户需求和可行性分析一般都需要进行数据调研。实际上，问题的解决离不开数据作为依据。数据的采集、管理和分析是解决问题的关键。

2.1.2 设计解决方案

通过业务需求分析，明确了用户的需求，就可以设计问题的解决方案了。这个方案涵盖的具体工作包括图2.1.3所示的几个方面。

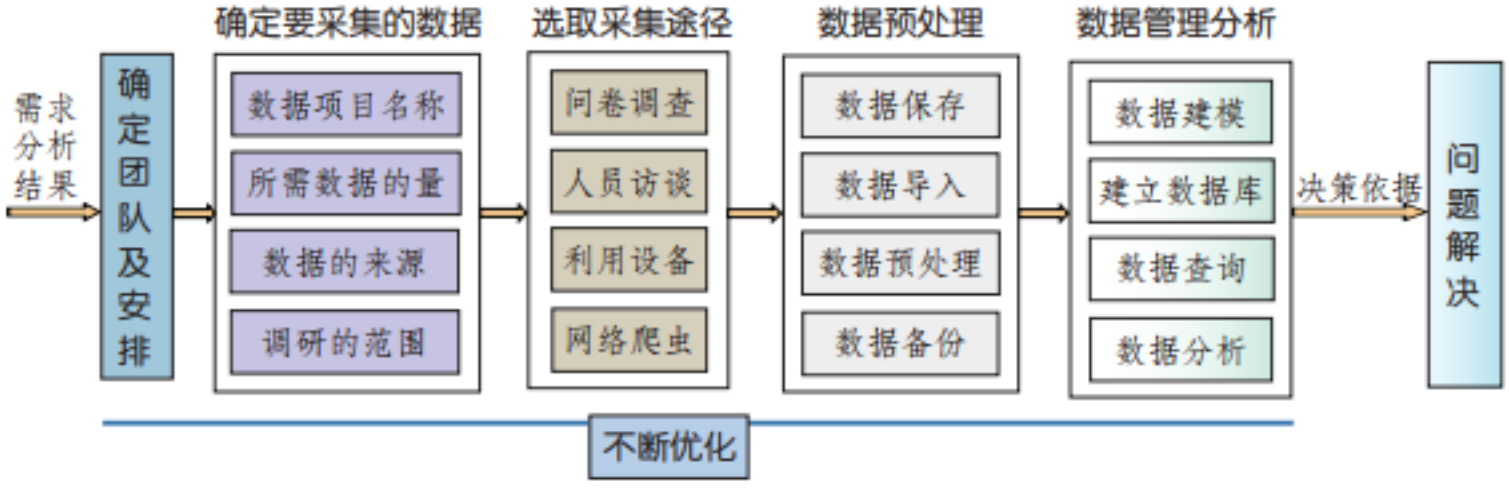


图2.1.3 总体解决方案

例如，在关于“是否开设共享汽车业务”的研究中，公司相关部门制订了以下方案（展示的是方案的部分内容）。

关于“是否开设共享汽车业务”的研究方案

一、确定团队和分工

由李晓峰、王宏博、杜宇轩、宋文杰与林宇等15人组成。李晓峰任组长，负责整体工作的统筹；他和王宏博负责组织人员编制调研数据；杜宇轩和宋文杰负责组织人员把数据导入计算机中并建立数据库；林宇负责组织人员进行数据统计和分析。

二、确定要调查的问题或数据

本方案中确定要获取的数据主要分为“用户需求”和“公司条件”两方面，说明如下。

用户需求	本市每年自由行游客的数量、目前汽车租赁的相关数据、常住人口数量、人均汽车拥有量、车位数量、支付方式、价格……
公司条件	投入资金额度、各岗位人员配备数量、预计车辆投入数量……

三、问卷设计与调查途径

1. 设计一份调查问卷，通过网络途径调查用户的需求。

2. 设计一份调查卷，用于与城市相关管理部门工作人员进行面对面采访，从管理部门的角度了解该项业务的需求。

3. 设计一份调查问卷，调查公司人力、物力等方面的数据。

四、获取官方网络数据

获取官方统计数据，了解城市人口、人均汽车拥有量、旅游出行等数据。

（后面的内容略）

微信公众号：电子课本大全

实践活动

细化方案

小李和同学就“餐厅是否要提供特色小吃”的问题，制订了一个解决方案，并初步确定了一些调研数据。请参考下面的提示，帮他们细化采集、处理和分析的方案。

用户需求	生源地数据（多少个省市）、希望提供小吃服务的人数、接受的价格、供应时间段……
餐厅条件	厨师数量、价格定位、正常就餐人数、最大接待能力……

2.1.3 数据需求分析

数据需求分析要回答以下几个问题：要获得什么数据？数据类型是什么？在什么范围内获取数据？用什么途径获取数据？

从以上方案可以看出，数据是解决问题的核心要素。无论什么研究项目（业务、问题），细化它的方案后都能发现，最终的工作都将围绕数据的采集、管理与分析展开。例如：要研究北京空气质量变化的趋势，就需要收集北京最近几年的空气质量数据、天气数据，甚至工厂数据、气体排放数据与重要会议日程数据等；要分析影响公司销售的关键因素，就需要调用公司的历史销售数据、客户数据与广告投放数据等，这就需要进行数据需求分析。在“是否开设共享汽车业务”的解决方案中，“确定要采集的数据”这个环节属于数据需求分析的内容，但需要进一步细化，最终得到类似表2.1.3所示的数据需求分析结果。

表2.1.3 数据需求分析结果

数据项名称	类　型	调研范围	获取途径
每户拥有汽车数量	数值	全市范围最新官方统计	官方统计网站
是否需要共享服务	文本	抽样调查，约30 000人	网络问卷
……	……	……	……

数据需求细化分析之后，需要写一份报告，将方案和分析结果汇总并在一定范围内交流，从而细化和优化方案。报告中除了把业务需求分析结果、初步解决方案和数据需求分析结果汇总在一起外，还应该增加一项内容，即描述整个过程中遇到的困难以及解决方法和尚未解决的问题，为下一阶段的工作做一些铺垫。

撰写需求分析往往可以用文字处理软件、演示文稿制作软件和思维导图制作软件等。

微信公众号：电子课本大全

项目实施

为“交通数据见发展”项目做方案

本项目研究的任务是“获取城市交通运输数据，了解城市经济发展”。下面就该问题展开相关的研究活动。

一、项目活动

1. 小组讨论，每人选定一个城市，作为数据调研的对象。

提示：后面的需求分析和方案设计，每个城市都相同。这里要求每人选择不同的城市作为研究对象，只是为了扩大样本数量，从而丰富小组内的数据量。

2. 确定问题的初步解决方案，并通过交流不断修改完善方案。

3. 进行数据需求分析，并用表格列出所需的数据以及采集的途径。

二、项目检查

1. 小组提交一份分析报告，其中要列出成员分工情况、业务需求分析表和数据需求分析表，并用思维导图呈现问题的最终解决方案。

2. 每人总结自己在项目中进行需求分析的表现、与其他成员合作与交流的表现。

练习提升

1. 简述业务需求分析与数据需求分析的联系。

2. 说出数据在问题解决中的重要性。

3. 仿照表2.1.2，为小李同学关于餐厅提供地方特色小吃服务的研究项目，制作一个业务需求分析表。

4. 简要阐述：数据需求分析报告主要包含哪些方面的内容。

5. 选定一个研究问题，对它进行数据需求分析，并总结这个过程中遇到的问题。

微信公众号：电子课本大全

2.2
数据采集与导入

学习目标 ▶▶▶

- 通过数据采集的实践，进一步了解数据采集的方法和途径。
- 能根据要求采集数据并保存为CSV文件。
- 掌握不同格式数据文件之间的转换，能熟练调用不同格式文件中的数据。

体验探索

导航地图数据的获取

现在人们用手机就能查到精度很高的地图数据（图2.2.1）。那么，这些数据是通过什么方式获得的呢？

首先，通过采集车（图2.2.2）进行实地采集。采集车车顶安装有多个像素很高的摄像头。副驾驶位置有数据显示屏，放在后备箱的机箱用于存储和处理数据。其次，物流车和半社会化的商用车提供的数据，以及个性化用户的原创数据也会成为高精度地图数据的来源。此外，车辆上的传感器，包括激光雷达、毫米波雷达、摄像头、陀螺仪以及雨水传感器等，都能够回传道路状况和天气状况等高精度地图所需的信息。



图2.2.1　地图数据



图2.2.2　地图数据采集车

实践和观察：使用一个导航软件中的电子地图，体验它的各项功能，观察地图清晰度，感受数据采集的作用。

微信公众号：电子课本大全

2.2.1 数据采集途径

在很多有关数据处理的业务中，数据采集是一项基本的工作，后期所有的处理和分析都要以采集到的数据为依据。

从前，人们主要采用观察法、访问法以及查阅法并通过人工纸笔记录的形式来采集数据。例如，我国进行人口普查时，全国仍有部分人口的抽样调查是通过走访社区、入户调查并用纸和笔记录实现的。但如今，更普遍的是使用各种采集设备进行自动采集，或者人工和自动化相结合的途径进行采集。采集设备有照相机、摄像机、录音设备和传感器等。数据采集正在经历着从人工向自动化设备演变、不同方式相互包容和不断丰富的发展过程。

思考活动

智能手机中有多少种传感器

手机中往往配置了很多种传感器，以下介绍几种。

· 加速度传感器，用于测量手机的加速度，以及手机在三个方向上的角度。

· 陀螺仪，提供高精度的角度信息。

· 磁力传感器，能够检测磁场，属于手机中的指南针类应用。

· 距离传感器，能让系统知道用户在通电话，然后会让系统关闭显示屏，防止用户因误操作影响通话。

· 光线传感器，能检测环境的亮度，当环境亮度高时，会相应调高显示屏的亮度；当环境亮度低时，会相应调低显示屏的亮度。

· 温度传感器，能测量手机自身的温度以及感应周围环境的温度。

· 指纹传感器，通常被用作手机的一种安全措施。

观察并思考：你的手机中有哪些传感器？它们的作用是什么？

无论采集数据的设备是什么、数据的来源是什么，但从用户解决问题的角度看，数据采集主要分为以下两种方式。

现场实测或调研访谈

人员到数据源现场采集或与数据源直接交流获得数据。例如，野外实地勘测、量算数据，监测台（站）观测、记录数据，利用物理和化学方法测定数据，社会调查数据，等等。中学阶段，在很多社会实践的调研项目中，这种数据采集途径应用得比较普遍。例如，在调查

微信公众号：电子课本大全

学生近视情况、课外时间安排情况以及市民垃圾分类意识等项目中，就需要利用问卷与调研对象面谈或网络交流的方式来采集数据。另外，通过现场采集或实测方式获取数据，往往要用到一些采集工具。

实践活动

总结一些数据采集工具

你用过或了解哪些数据采集工具？它们可以采集什么数据？请填写表2.2.1。

表2.2.1　数据采集工具分析

工具名称	采集的数据
智能手环	
温度计	
视频监控器	

思考活动

如何获取大学招生的相关数据

已经上高中的小明，对几所重点大学很向往，他现在就想了解这些大学近五年在本省的招生专业、招生人数和分数线。获取这些数据的好办法就是通过互联网查找。

思考：小明应该掌握哪些网络数据采集的方式？

获取他人调查的数据

这种方式是获取他人的数据（可公开使用），然后将这些数据进行适当的处理，为自己分析解决问题所用。通过这种途径采集的数据称为第二手数据或间接数据。这种数据采集方式还可以分为以下几种。

系统日志采集。很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如Hadoop的Chukwa和Facebook的Scribe等，这些工具均采用分布式架构，能满足每秒数百兆字节的日志数据采集和传输需求。

网络数据采集。网络数据采集是通过网络爬虫或网站公开的应用程序编程接口等方式，从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为本地数据文件，并以结

微信公众号：电子课本大全

构化的方式存储。例如，本书的一些项目学习中，就可以通过网络爬虫技术从国家或地方统计局网站上采集统计数据。

从特定接口采集数据。对于企业生产经营数据或者学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等方式来采集。

当前，数据的来源非常丰富，无论是采取哪种获取方式，都可以从恰当的数据来源中采集到需要的数据（参考表2.2.2）。

表2.2.2　数据的来源

数据来源	数据说明
社交平台	如微博、微信等，可以分析用户平时在这些社交媒体上的行为动向，归纳出用户的喜好或关注点，这些数据能够为企业挖掘用户需求提供重要依据
设备运行产生	越来越多的机器配备了连续测量和报告运行情况的传感装置，这些数据也属于大数据的范围
设备采集产生	一些视频、音频等设备产生的数据，如视频监控数据
行业数据	企业内部的行业数据，如网络购物平台的商品及销售数据

阅读拓展

大数据中的日志数据采集

如今，各个公司、团体机构都在研究和构建自己的大数据处理平台，通过对海量数据的整合、分析与统计，充分利用数据中蕴含的价值，挖掘出对企业管理、业务改进及商机捕捉等更为有用的信息。而这些大规模的数据都需要从各个应用系统中获取，对于规模较大的公司，应用系统众多，数据生成、存储方式和规范都有所不同，使得不同应用系统中数据（尤其是日志数据）的采集变得困难。如何可靠、实时地将大规模的日志数据传送到大数据平台中，然后进行抽取、转置、加载处理，并对数据进行统计、分析、挖掘，是构建大数据平台必须面对的问题。

2.2.2　创建CSV数据文件

通过实测和调研采集的数据，需要利用专门的数据处理工具输入，从而创建成数据文件，以备后期的数据分析调用。

创建数据文件可以使用多种软件，如WPS中的表格工具、Excel，还有《记事本》、Notepad++，还可以用Python等编程软件来创建。因此，我们要根据数据的量和复杂程度，合理选择软件创建数据文件。

微信公众号：电子课本大全

CSV文件

CSV（comma-separated values）格式的文件是一种纯文本文件（以下简称CSV文件），是一组字符序列，字符之间用英文逗号或制表符分隔。在Windows系统环境中，《记事本》、Excel和Notepad++等文本编辑器都能打开CSV文件。网页中存在CSV文件，在采集时，需要对它们进行处理；而在处理数据时，也可以把数据保存为CSV文件。下面介绍两种创建CSV文件的方法。

阅读拓展

CSV文件的主要特点

CSV文件的特点主要有以下几个。

1. 文件结构简单，基本上和文本的差别不大，并具备很强的开放性。

2. 可以转换为电子表格软件生成的文件格式（如Excel文件），便于数据共享。

3. 由于其简单的存储方式，可以减少存储信息的容量，有利于网络传输以及客户端的再处理；同时，由于是一堆没有任何说明的数据，故具备基本的信息安全性。

使用《记事本》创建CSV文件

例如，创建一个包含某个城市近5个月交通运输中的货物运输情况信息（包括时间、铁路、水运、公路和机场）的CSV文件，可按以下方法操作。

（1）启动《记事本》，在文档编辑区输入数据。注意：关键字与关键字之间使用英文逗号隔开，第1行为引用字段，第2～6行数据为字段的对应值（图2.2.3）。

（2）执行“文件→另存为”命令。为文件命名时要加上“.csv”，编码选择“UTF-8”（图2.2.4）。

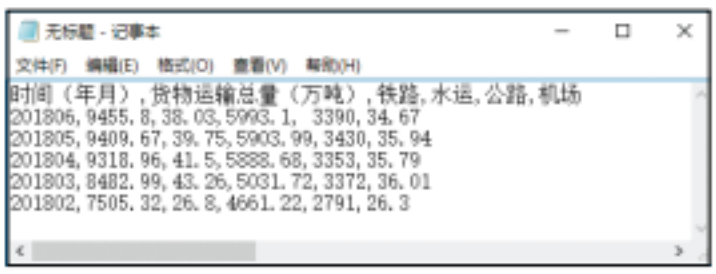


图2.2.3　在“记事本”中输入数据

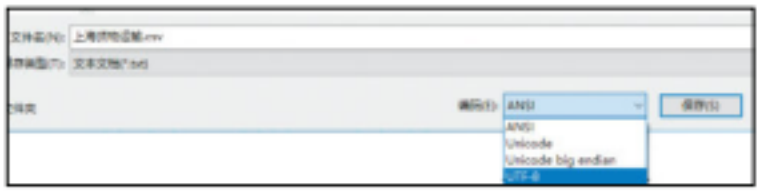


图2.2.4　选择编码方式

使用Python创建CSV文件

（1）打开Python编程窗口，执行“File→New File”命令，弹出

微信公众号：电子课本大全

Untitled窗口。

（2）执行“File→Save”命令保存文件，然后编写代码创建CSV文件并保存（图2.2.5）。

```
File  Edit  Format  Run  Options  Window  Help
import csv
filename = 'Python创建csv文件.csv'
f = open(filename,'w',newline='')
writer = csv.writer(f)
writer.writerow(['时间（年月）','货物运输总量（万吨）','铁路','水运','公路','机场'])
data = [('201806',9455.8,38.03,5993.1,3390,34.67),
        ('201805',9409.67,39.75,5903.99,3430,35.94),
        ('201804',9318.96,41.5,5888.68,3353,35.79),
        ('201803',8482.99,43.26,5031.72,3372,36.01),
        ('201802',7505.32,26.8,4661.22,2791,26.3)]
writer.writerows(data)
f.close()
print('创建csv文件完成')
```

图2.2.5　程序代码

创建好的CSV文件，可以随时把其中的数据调出并查看。

实践活动

把调查数据存为CSV文件

1. 做个小调查：调查班上同学的身高、体重、性别和视力，然后用一种电子表格软件输入数据并保存为CSV文件。

2. 在电子表格软件中调出这个CSV文件并查看。

2.2.3　从网络中采集数据

很多时候人们需要从网上采集数据，并从数据中挖掘或提取有价值的信息。在Python中，可以用Urllib库里的Request模块爬取统一资源定位符（Uniform Resource Locator，URL）内容。下面介绍采集某市统计局网站上的每月交通运输数据，并把数据保存为CSV文件。

阅读拓展

遵纪守法使用网络爬虫

网络爬虫是一个自动提取网页的程序，是搜索引擎的重要组成。通过网络爬虫，可以从网络上提取到我们需要的各种数据及资源。目前，互联网世界已经通过自身的协议建立起一定的规范（Robots协议），每一个使用网络爬虫的用户都应该遵守这项协议。在爬取网站数据的时候，需要遵守Robots协议，限制自己的操作，如约束网络爬虫程序的速度；在使用数据的时候，必须保护网站的知识产权。

微信公众号：电子课本大全

（1）分析网站结构。访问某市统计局的官方网站，依次找到需要爬取的交通运输数据页面。

（2）提取目标首页URL（这也是一个多页面的爬取）。URL的规则为：URL最后地址分别是index、2、3、4和5，分别表示第1页、第2页、第3页、第4页及第5页。每页有一定数量的交通运输情况条目（图2.2.6），每个条目分别是一个超链接，显示该月的交通运输情况数据，相对目标首页属于第2层链接。

（3）打开网页可以看到本月的交通运输情况数据。指标包括：货物运输总量（铁路、水运、公路与机场）；港口货物吞吐量（进港量与出港量）；国际标准集装箱吞吐量（进港量与出港量）；机场旅客吞吐量。这些数据主要以表格的形式呈现，大部分页面是具有相同行数和列数的表格（图2.2.7）。

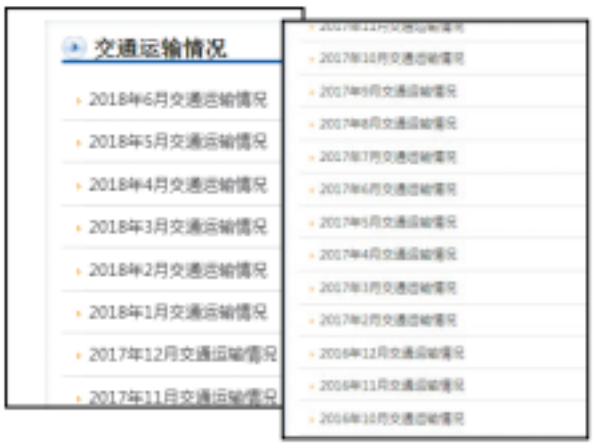


图2.2.6　网页条目



图2.2.7　网页中的表格

（4）安装Beautiful Soup扩展库。打开Python所在的文件夹，按住Shift键的同时，右击文件夹的空白区域，然后选择“在此处打开命令窗口（W）”，在命令行中输入：pip install beautifulsoup4。

阅读拓展

Beautiful Soup简介

Beautiful Soup是用Python编写的一个HTML/XML解析器。它提供了一些简单的、Python式的函数用来处理导航、搜索、修改分析树等操作。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，不需要太多代码就可以写出一个完整的应用程序。

Beautiful Soup会自动将输入的文档转换为Unicode编码，并将输出的文档转换为UTF-8编码。Beautiful Soup已成为好用的Python解释器，为用户灵活提供不同的解析策略和强劲速度。可以利用pip或者easy_install命令来安装。

微信公众号：电子课本大全

（5）使用get_urls()函数爬取网页地址，代码如下。

```
#获取网页地址列表
def get_urls(url):
    links=[]
    titles=[]
    try:
        kv = {'user-agent': 'Mozilla/5.0'}                    # 伪装浏览器访问
        r = requests.get(url, headers=kv)
        r.raise_for_status()
        r.encoding = r.apparent_encoding                      # 设置编码
        soup = BeautifulSoup(r.text,'lxml')
        for page in soup.findAll('ul',id = 'titlelist'):
            data = page.select('a')
            for each in data:
                href = each.get('href')
                title = each.get('title').strip()
                links.append(href)
                titles.append(title)
        return(links,titles)
    except:
        print(url, '爬取失败')
```

（6）使用get_data()函数爬取每个网页的表格数据，代码如下。

```
#爬取数据
def get_data(url):
    cdata=[]
    html=urlopen(url)
    soup = BeautifulSoup(html.read(),'lxml')
    trs = soup.find('tbody').findAll('tr')#获得表格行数据，用于计算表格行数
    for i in range(len(trs)-1):
       data = [c.get_text().strip() for c in soup.find_all('tr')[1+i].find_
all('td')[0:2]]              #获得每个表格中前两列数据
       cdata.append(data)    #获得的所有表的前两列数据
    return(cdata)
```



实践活动

get_urls()函数和get_data()函数

获取网页数据主要的两个函数是get_urls()和get_data()，请进一步了解它们的功能和使用方法。

（7）保存爬取的数据，代码如下。

```
def save_data(data,title):
    row1=['日期']
    try:
        with open("数据.csv",'w',newline='') as f:
            writer = csv.writer(f)
            row0=['上海全市每月交通运输情况']
            for i in range(len(data[0])): #生成爬取数据表的字段
                row1.append(data[0][i][0])
            writer.writerow(row0)          #写入一行数据标题
            writer.writerow(row1)          #写入一行数据字段
```


微信公众号：电子课本大全

```
            for i in range(len(data)):
                row=[title[i][:-6]]        #每行的初始值为日期
                for j in range(len(data[i])):
                    row.append(data[i][j][1])  #生成每行数据
                year=title[i][0:4]             #取日期中的年份
                month=title[i][5:-7]           #取日期中的月份
                #2018年11月之前没有客运数据，填充缺失值
                if len(month)==1:
                    month='0'+title[i][5:-7]    #月份使用两位表示
                if year+month <='201811':
                    row = row[:6]+[np.nan,np.nan,np.nan,np.nan,np.
nan]+row[6:]
                # 每行数据写入文件
                writer.writerow(row)
    except:
        pass
```

（8）载入所需模块并运行主函数，代码如下。

```
def main():
    #给定要爬取的页面网址
    url0 = 'http://tjj.sh.gov.cn/html/sjfb/ydsj/ydsj37/'
    #生成每月交通运输情况网址
    urls=[]
    names=[]
    for w in ['index.html','2.html','3.html','4.html','5.html']:
        url1 = url0+w
        datas = get_urls(url1)
        data_links = datas[0]    #获得网页中每月统计数据的网址
        data_titles = datas[1]
        urls.extend(data_links)
        names.extend(data_titles)
    #提取每个网页中的交通运输情况数据
    print('\n开始爬取交通数据...\n')
    print('数据网址...')
    tdata =[]
    for link in urls:
        url = 'http://tjj.sh.gov.cn'+ link
        print(url)
        tdata.append(get_data(url))
    print('\n数据爬取完毕！\n')
    save_data(tdata,names)

main()
```

程序运行结束后，就可以得到一个保存了指定月份交通运输数据的CSV文件，在电子表格软件中可以查看到类似图2.2.8的效果。

上海全市每月交通运输情况																	
日期	货物运输总量（万吨）	铁路	水运	公路	机场	旅客发送量（万人次）	铁路	港口	公路	机场	港口货物吞吐量（万吨）	进港量	出港量	国际标准集装箱吞吐量（万TEU）	进港量	出港量	机场旅客吞吐量(万人次)
2019年4月	9148.05	36.65	5812.74	3266	32.66	1891.22	1120.98	6.6	253	510.64	6146.68	3460.18	2686.5	361.15	175.45	185.7	1012.28
2019年3月	9021.09	44.44	5617.97	3324	34.68	1720.01	1007.4	5.84	201	505.77	6412.69	3664	2748.7	380.52	185.62	194.9	1013.26
2019年2月	7990.73	26.66	4824.3	2788	21.56	1610.11	838.1	9.93	296	466.08	4579.58	2640.14	1939.45	285.52	155.11	130.41	959.71
2019年1月	9321.33	39.62	5985.58	3263	33.15	1894.49	1121.64	10.28	238	524.57	6387.94	3673.69	2714.25	375.18	182.73	192.45	1005.51
2018年12月	9160.01	41.56	5736.41	3346	36.04	1643.75	905.73	10.37	241	486.65	6145.64	3488.9	2646.73	355.1	176.76	178.34	966.72
2018年11月	8628.59	36.11	5228.71	3326	35.78	nan	nan	nan	nan	nan	5824.42	3280.37	2544.05	352.93	175.09	177.83	949.1
2018年10月	9353.48	41.58	5988.18	3290	35.74	nan	nan	nan	nan	nan	6234.93	3580.07	2654.86	357.35	176.03	181.32	1022.53
2018年9月	9030.75	39.49	5980.54	3303	35.72	nan	nan	nan	nan	nan	6430.62	3636.42	2802.2	381.19	186.75	194.44	957.21
2018年8月	9617	39.07	6228.25	3317	34.68	nan	nan	nan	nan	nan	6149.29	3517.41	2631.89	348.39	169.52	178.88	1039.99
2018年7月	9186.5	37.47	5716.96	3377	35.07	nan	nan	nan	nan	nan	6222.9	3622.61	2600.29	355.32	174.38	180.94	1014.8

图2.2.8　爬取获得的数据

微信公众号：电子课本大全

实践活动

获取大学招生信息

访问一些大学的网站，查找近五年的招生信息，利用网络爬虫采集相关数据并保存为CSV文件。

2.2.4 导出CSV文件中的数据

采集的数据保存到CSV文件中后，后期也可以方便地把数据导出来使用。例如，使用Python导出前面用《记事本》生成的某市货物运输数据的CSV文件，可输入图2.2.9所示的代码。程序运行后可得到图2.2.10所示的结果。

```
import csv
filename = 'Python创建csv文件.csv'
f = open(filename,'r')
csv_reader = csv.reader(f)
for row in csv_reader:
        print(row)
f.close()
```

图2.2.9 导出CSV文件数据的代码

```
Python 3.6.2 (v3.6.2:5fd33b5, Jul  8 2017, 04:57:36) [MSC v.1900 64 bit (AMD64)]
on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\Administrator\Desktop\readPythonData.py ==========
['时间（年月）', '货物运输总量（万吨）', '铁路', '水运', '公路', '机场']
['201806', '9455.8', '38.03', '5993.1', '3390', '34.67']
['201805', '9409.67', '39.75', '5903.99', '3430', '35.94']
['201804', '9318.96', '41.5', '5888.68', '3353', '35.79']
['201803', '8482.99', '43.26', '5031.72', '3372', '36.01']
['201802', '7505.32', '26.8', '4661.22', '2791', '26.3']
>>>
```

图2.2.10 导出的结果

实践活动

在Python中导出Excel表格数据

既可以用Excel创建CSV文件，也可以把CSV文件转换为Excel文件。在Python中，还可以把Excel文件中的数据导出来。例如，要把名为“货物运输数据文档.xlsx”文件中的数据导出来，方法如下。

（1）安装Xlrd模块。在命令窗口输入pip install xlrd（图2.2.11），用来读取Excel文件里的内容。

（2）输入图2.2.12所示的代码，就可以导出指定的数据了。

```
D:\Python36>pip install xlrd
Collecting xlrd
  Downloading https://mirrors.ustc.edu.cn/pypi/web/packa
d8528bec4ea252a850bffec4be88ebc367e23a1a84b80b/xlrd-1.1.
108kB)
    47% |
    56% |
    65% |
    75% |
    84% |
    94% |
    100% |
0.8MB/s
Installing collected packages: xlrd
Successfully installed xlrd-1.1.0

D:\Python36>
```

图2.2.11 安装Xlrd模块

```
import xlrd
fileName = '货物运输数据文档.xlsx'
xls = xlrd.open_workbook(fileName)

sheet = xls.sheets()[0] # 选择表：括号内为第几个表
col = sheet.col_values(0) # 获取列：括号内为第几列的数据
print(col)
row = sheet.row_values(1) # 获取行：括号内为第几行的数据
print(row)
```

图2.2.12 导出的文件数据

微信公众号：电子课本大全

思考活动

采集数据要遵守哪些原则

使用网络爬虫采集数据时，应该遵守哪些原则？请把你能想到的都列出来，并与其他同学交流，在以后的数据采集中规范自己的行为。

项目实施

为“交通数据见发展”项目采集数据

一、项目活动

1. 小组讨论，共同确定每位成员所要采集数据的来源、采集的途径和工具，填写表2.2.3。

表2.2.3　数据采集说明

数　据	说　明
来源	如数据所在的网站地址
采集的途径和工具	使用的工具或技术（如网络爬虫）
创建数据的方法	
数据文件的名称	

2. 在Python中编写程序代码，把交通运输部门官方网站上的数据导入CSV文件。

3. 把数据文件同时保存为电子表格的格式，作为必要的备份。

二、项目检查

1. 组内展示各个成员所采集的数据成果，并注明数据的出处。

2. 每人总结自己采集数据过程中所用到的知识点，并把它们填写在图2.2.13中相应的横线上。

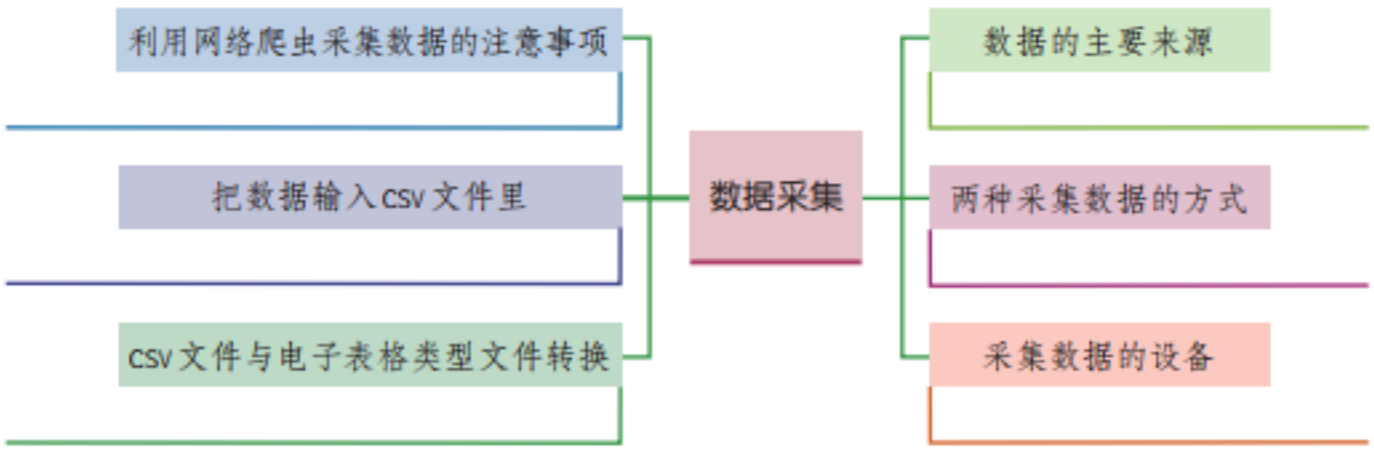


图2.2.13　知识总结

微信公众号：电子课本大全

练习提升

1. 结合日常学习生活中的例子，简述自己获取数据的直接来源有哪些。

2. 寻找身边可以采集数据的工具，并讨论这些工具是如何采集数据的。

3. 小明的爸爸开了一家玩具店，想通过对周围邻居进行问卷调查，从而了解玩具店是否受欢迎。你觉得通过这种方式收集的数据是否准确？你有更好的方法吗？

4. 打开Python编程界面，将下面A、B两位同学的成绩导出为CSV文件。

	Chinese	Math	Physics
A	93.0	91.0	92.0
B	95.0	92.0	85.0

5. 打开Python编程界面，尝试将上题中A、B两位同学的成绩导出为Excel文件，并用Excel求平均值，然后在Python中读取数据。

6. 选择1 ～ 2种你最感兴趣的数据获取方法，并提供本书中没有介绍的案例。

7. 目前人们更多地采用自动获取数据的方式，这是否意味着人工获取数据的方式将被淘汰？请谈谈你的观点。

微信公众号：电子课本大全

2.3 数据结构化与数据清洗

学习目标 ▶▶▶

- 能结合实例，说出噪声数据的现象与成因。
- 了解数据预处理的环节，掌握数据清洗的方法。
- 了解数据结构化的概念，说出不同结构化程度数据的区别。

体验探索

了解“商品大脑”

经常网购的球迷应该知道，只要在搜索商品时输入“××球星同款”，就能一键搜索到自己想要的商品。这个功能就像一个“商品大脑”，能从近几十亿件商品中找出对应的商品。例如，输入“我需要一条漂亮的真丝围巾”后，“商品大脑”会通过语法、词法分析来提取语义要点，如“一”“漂亮”“真丝”和“丝巾”关键词，从而帮助人们搜索到合适的商品。“商品大脑”还学习了大量的行业标准，如全棉、低糖、低嘌呤等标准。“商品大脑”可以从公共媒体、专业社区的信息中识别出近期热词，并跟踪热词的变化（图2.3.1）。



图2.3.1　大脑信息处理抽象图

思考：事实上，这一现象背后的技术之一是知识图谱技术。那么，什么是知识图谱？知识图谱的知识来源是否与不同结构化程度的数据有着密切关系呢？

通过查阅资料可以知道，知识图谱的第一个部分是知识获取，主要阐述如何从非结构化、半结构化和结构化数据中获取知识，这是很关键的一个环节。如今，高效地获取并处理大量的数据，把接收的数据（通常是没有结构或杂乱的）转换成具有结构的数据，已

微信公众号：电子课本大全

成为数据处理的重要环节。另外，现实世界中的数据大体上都是不完整、不一致的“脏数据”，无法直接进行数据挖掘，或挖掘结果差强人意。为了提高数据挖掘的质量，数据预处理技术应运而生。本节先介绍数据结构化的概念，然后介绍噪声数据和数据预处理的相关知识。

2.3.1 不同结构化程度的数据

人们时刻都在接收各种数据，但大多数人并不注意它的结构化程度。根据结构化程度的不同，在组织、存储和分析数据时需要区别对待。

例如，小李他们在调查同学对“餐厅是否应该提供地方特色小吃”这个问题的意见过程中，把每个被采访同学的性别和家乡名称都按统一的格式记录下来，把他们对特色小吃服务的意见通过录音记录下来。在这个情景中，用统一格式进行记录的性别和家乡名称就是结构化数据，而通过录音记录的采访意见就是非结构化数据。

结构化数据。能够用数据或统一的结构加以表示的数据称为结构化数据，如数字、符号。二维表结构数据（表2.3.1）就是结构化数据的典型。结构化数据由明确定义的数据类型组成，其模式可以使其易于被搜索。例如，企业财务系统、图书管理数据库、校园一卡通以及超市销售记录等数据都属于结构化数据。

表2.3.1 二维表结构数据

书名	作者	书号	书籍类型
利用Python进行数据分析	Wes Mckinney	9787111436737	信息技术
数据挖掘概念与技术	Jiawei Han等	9787111391401	信息技术
数据科学入门	Joel Grus	9787115417411	信息技术

非结构化数据。是指数据结构不规则或不完整、没有预定义的数据模型、不方便用数据库二维逻辑表来表现的数据。非结构化数据具有内部结构，但不通过预定义的数据模型或模式进行结构化。它可能是文本的或非文本的，也可能是人为的或机器生成的。在互联网飞速发展的今天，知识大量存在于非结构化数据中。非结构化数据已成为数据分析和挖掘中非常重要的资源。

微信公众号：电子课本大全

思考活动

非结构化数据举例

- 文本文件，如文字处理文档、电子表格、演示文稿、电子邮件和日志。
- 社交媒体，如来自微信、微博和脸书等的数据。
- 移动数据，如短信和定位等信息。
- 即时通信，如聊天、即时消息、电话录音和协作软件等数据。
- 媒体数据，如MP3、数码照片、音频文件和视频文件。
- 卫星图像，如天气、地形等图像数据。
- 科学勘探数据，如石油和天然气勘探、空间勘探、地震图像。
- 监控数据，如监控获得的数字照片和视频。

思考：在平时的应用中，你还能想到哪些是非结构化数据？请把它们列举出来。

半结构化数据。所谓半结构化数据，就是介于完全结构化数据与非结构化数据之间的数据。例如，邮件、XML、HTML文档、报表以及资源库等属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

结构化数据与非结构化数据，除了是否存储在关系数据库内的区别，还在数据分析便利性上存在区别。目前，分析结构化数据的工具比较丰富和完善，但用于挖掘非结构化数据的分析工具还处于发展阶段。在网络数据非常丰富的今天，非结构化数据比结构化数据多得多，并且增长速度非常快，非结构化数据的分析与挖掘已越来越受到重视。

实践活动

理解不同结构化程度的数据

根据自己的使用经验，从表2.3.2的几个方面谈谈对不同结构化程度数据的理解。

表2.3.2　对不同结构化程度数据的分析

不同类别	概　念	特　点	举例说明
结构化数据			
非结构化数据			
半结构化数据			

微信公众号：电子课本大全

2.3.2 噪声数据的现象与成因

有这样一个案例：银行给客户发放信用卡，结果遭到一些客户的投诉，原来是有些应该发卡的客户被系统标记为“不发卡”。究其原因，是因为这部分用户数据存在错误，如年收入少输入一个0、性别输错等。在这个案例中，输入错误的数据就是噪声数据，从而导致了发卡错误。

噪声数据指数据中存在着错误或异常（偏离期望值）的数据。在图2.3.2中，绝大部分数据都集中在中间的区域，而有少部分数据零散出现在周围。可以认为，这些零散出现的数据就很可能是噪声数据。噪声数据往往会影响数据分析和数据挖掘的结果。



图2.3.2 非数据集中区域存在异常值

噪声数据的成因主要包括以下三方面。

*数据采集工具导致的噪声。*采集工具工作时本身会产生误差，搜集到的数据就是含有噪声的数据。这些噪声数据是随机产生的，其数据大小也是随机的。

*数据输入时的人为或计算错误。*这类噪声数据是偶然产生的，产生的时间是随机的。噪声数据的大小与输入错误或计算误差值有关。

*数据传输过程中产生的错误。*由于通信线路会受到内部因素（本身的带宽、压缩及传输介质）及外部因素（磁场、温度、干扰等）的影响（图2.3.3），在传输的过程中或多或少都会产生一些噪声数据，使得接收到的数据与原始数据存在差别。例如，烟雾感应器受外部环境的影响，导致感应失灵，就会产生噪声数据。

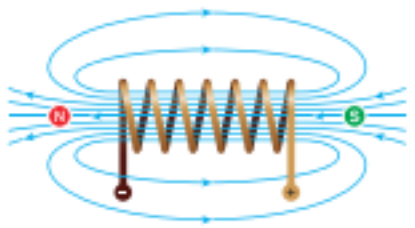


图2.3.3 干扰的出现

思考活动

大数据中的噪声数据

大数据来源可以分为来自物理世界和人类社会两类。物理世界的数据，是通过科学实验得到的数据，是直观的、可以量化的数据；人类社会数据具有多样性、时效性等特点，也是大数据的主要来源。但这样的数据有很多噪声。例如，当某一个热点事件发生时，互联网上会产生非常多的并发数据，从而产生很多噪声数据。因此，大数据并不是一个纯粹的技术或纯数学维度的学科，还需要与社会学、人类学、心理学以及行业知识交叉融合，才能做好大数据相关的数据分析和挖掘等工作。

思考：在大数据来源中，产生噪声数据的主要原因有哪些？请举例说明。

微信公众号：电子课本大全

2.3.3 数据清洗

基于数据量、数据来源的多重异质性和数据类型的不同，数据很容易受到噪声和不一致的影响。因此出现了一些数据预处理技术，如数据清洗、数据集成、数据压缩和数据转换等技术。数据清洗是把“脏”的数据“洗掉”，即把噪声数据剔除掉，是发现并纠正数据文件中可识别错误的一道程序，包括检查数据一致性、处理无效值和缺失值等。数据清洗的主要步骤包括纠正错误、删除重复项以及统一规格、修正逻辑、转换构造、数据压缩、补足残缺/空值、丢弃数据/变量。录入后的数据清洗一般由计算机完成。

在必修课程中，我们用过Pandas，它是Python的一个数据分析库，提供了大量处理数据的函数和方法。

下面介绍用Python中的Pandas进行数据清洗。

数据缺失检查

数据缺失在数据分析中很常见。Pandas使用浮点值NaN表示浮点和非浮点数组中的缺失数据，它只是一个便于被检测出来的数据而已。例如，要检查数据是否有缺失值，可输入以下代码。

```
from pandas import DataFrame
df=DataFrame({'c1':[0,1,2,None],'c2':[1,None,2,3]})
#DataFrame：二维的表格型数据结构

print(df)
print(df.isnull())    #数据缺失检查
```

```
   c1   c2
0  0.0  1.0
1  1.0  NaN
2  2.0  2.0
3  NaN  3.0
      c1     c2
0  False  False
1  False   True
2  False  False
3   True  False
>>>
```

图2.3.4 运行结果

运行结果如图2.3.4所示。可以看到，在Python中，None值会被识别为缺失值NaN。在当前缺失值不太多的时候，可以通过多种方法进行缺失值的填充。例如，可以直接指定特定的值来填充缺失值，代码如下。

```
from pandas import DataFrame
df=DataFrame({'c1':[0,1,2,None],'c2':[1,None,2,3]})
#DataFrame：二维的表格型数据结构

print(df.fillna('missing')) #数据缺失值指定填充为missing
print(df.fillna(df.mean())) #数据缺失值指定填充为该列数据的均值

df.ix[0,1]=None             #设第0行第1列数据为缺失值

print(df)
print(df.fillna(method='bfill',limit=1))#根据周围的值填补缺失值
```

```
    c1   c2
0  0.0  1.0
1  1.0  NaN
2  2.0  2.0
3  NaN  3.0
         c1       c2
0         0        1
1         1  missing
2         2        2
3  missing        3
    c1   c2
0  0.0  1.0
1  1.0  2.0
2  2.0  2.0
3  1.0  3.0
    c1   c2
0  0.0  NaN
1  1.0  NaN
2  2.0  2.0
3  NaN  3.0
    c1   c2
0  0.0  NaN
1  1.0  2.0
2  2.0  2.0
3  NaN  3.0
>>>
```

图2.3.5 运行结果

程序运行的结果如图2.3.5所示。程序指定了三种填充方法，填充值分别是missing、均值以及周围值。根据周围值填补缺失值时，指定了bfill（back fill）方法进行填充；使用limit参数是为了限制连续填充。这里选择1表示一列中有多个缺失值相邻时，只填充最近的一个缺失值。

微信公众号：电子课本大全

过滤数据

缺失值较少时，可以选择填充的方式来完善数据集。但缺失值较多且重要程度不高时，可选择去除这些没有价值的数据。例如，自定义去除缺失值的三种方式，可以输入以下代码。

```
from pandas import DataFrame,np
df=DataFrame({'one':np.arange(0,0.7,0.1),'two':np.arange(0,0.7,0.1)})
#DataFrame：二维的表格型数据结构

df.ix[0,0]=None                    #设置第0行第0列数据为缺失值
df.ix[1:3,0:2]=None                #设置第1、2、3行第0、1、2列数据为缺失值
df['three']=np.nan                 #增加1列为缺失值

print(df)

print(df.dropna())                      #输出过滤的数，默认how='any'
print(df.dropna(how='all'))             #输出过滤的数
print(df.dropna(how='all',axis=1))      #输出过滤的数
```

运行结果如图2.3.6所示。程序创建了df并插入了缺失值，利用dropna(how,axis)过滤数据，默认情况下，dropna()的参数how = 'any'，即有一个数据为NaN就去除整行或整列数据。axis用于指定删除行或列；0代表去除该行，1代表去除该列。

```
   one  two
0  0.0  0.0
1  0.1  0.1
2  0.2  0.2
3  0.3  0.3
4  0.4  0.4
5  0.5  0.5
6  0.6  0.6
   one  two  three
0  NaN  0.0    NaN
1  NaN  NaN    NaN
2  NaN  NaN    NaN
3  NaN  NaN    NaN
4  0.4  0.4    NaN
5  0.5  0.5    NaN
6  0.6  0.6    NaN
Empty DataFrame
Columns: [one, two, three]
Index: []
   one  two  three
0  NaN  0.0    NaN
4  0.4  0.4    NaN
5  0.5  0.5    NaN
6  0.6  0.6    NaN
   one  two
0  NaN  0.0
1  NaN  NaN
2  NaN  NaN
3  NaN  NaN
4  0.4  0.4
5  0.5  0.5
6  0.6  0.6
>>>
```

图2.3.6　运行结果

检测和过滤异常值

异常值的范围比较广，一般来说，数据格式不一致、数据范围异常等都属于异常值，主要根据一些生活和业务的常识来界定。例如通过年龄的界定来筛选想要的数据，可输入以下代码。

```
from pandas import DataFrame
df=DataFrame({'Name':['A','B','C'],'Age':[-1,14,125]})
#DataFrame：二维表格数据结构
print(df)
print(df.query("Age>=0 and Age<=110")) #输出去除异常值后的数据
```

运行结果如图2.3.7所示。

```
   Age Name
0   -1    A
1   14    B
2  125    C
   Age Name
1   14    B
>>>
```

图2.3.7　运行结果

移除重复数据与冗余信息

通常情况下重复或冗余的数据没有价值，而且占用一定的存储空间，有必要适当移除这些数据。例如，使用drop_duplicates方法可以方便地移除重复的数据，代码如下。

```
from pandas import DataFrame
df=DataFrame({'A':[1,1,2,2],'B':[3,3,4,4]})
print(df)
print(df.drop_duplicates()) #移除重复数据
```

运行结果如图2.3.8所示。

```
   A  B
0  1  3
1  1  3
2  2  4
3  2  4
   A  B
0  1  3
2  2  4
>>>
```

图2.3.8　运行结果

微信公众号：电子课本大全

实践活动

对“天气数据”进行数据清洗

一、数据来源

本书资源中有某个学习小组采集的天气数据。数据由放在室外的两台环境检测设备采集而得。其中一台每10 min采集一次温度及湿度，采集的数据按照时间递增排列，没有重复，但存在数据缺失，即某时间点的温度、湿度未采集；另一台每10 min采集一次细微颗粒物（Particulate Matter 2.5，PM2.5）值，采集的数据也是按时间递增排列，也存在缺失数据。两台设备采集时间并不同步。

二、实践任务

每10 min测量一次数据，使得数据冗余度比较高，现改为每个整点时间记录一次温度、湿度及PM2.5值，缺失值用NaN表示。如果整点时间没有实际测量的数据，把离整点时间最近的前后两条测量数据的平均值作为测量数据。

三、清洗步骤

（1）数据预处理，对两个设备上获取的数据（配套资源中有可参考的文件a.xls和b.xls）进行冗余清洗。

（2）使用Pandas中的工具读取Excel数据。

（3）使用函数merge()进行数据合并，按照时间排序，设置参数。

（4）运行代码，输出合并的结果。

四、程序代码

```
#-*- coding: UTF-8 -*-

#导入包
import xlrd, xlwt
import time,datetime
import pandas as pd

#设置路径
file_a='a.xls'
file_b='b.xls'
listdata_a = [['时间','温度','湿度']]      #用来保存处理后的数据
listdata_b = [['时间','PM2.5']]
clearedfile_a = 'a_copy'
clearedfile_b = 'b_copy'

def clear_data(file,listdata,clrfile):    #清洗数据并保存为Excel文件
    #打开数据文件
    data=xlrd.open_workbook(file)
    table=data.sheet_by_name(u'Sensor Data') #通过名称获得工作表
    nrows=table.nrows   #通过工作表的属性获得行数
    #清洗多余数据
    for rownum in range(1,nrows):  #用一个循环来遍历一次文件
        rowdata_bf = table.row_values(rownum-1) # 当前行的前一行数据
        rowdata_nw = table.row_values(rownum)   # 当前行数据
        hour_nw = (rowdata_nw[0])[11:13]        # 取当前行时间（小时）
```


微信公众号：电子课本大全

```
        new_time_nw = (rowdata_nw[0])[0:13]+":00:00"  # 构造新的整点时间数据,当
前行
        if ((rowdata_bf[0])[0:19] <= new_time_nw <= (rowdata_nw[0])[0:19]): #
判断整点并计算整点数据，保存至列表
            if (clrfile == 'a_copy'):     #处理温度及湿度数据，保留1位小数
                new_rowdate = [new_time_nw,round((rowdata_bf[1]+rowdata_
nw[1])/2,1),round((rowdata_bf[2]+rowdata_nw[2])/2,1)]
            if (clrfile == 'b_copy'):     #处理PM2.5数据，保留1位小数
                new_rowdate = [new_time_nw,round((rowdata_bf[1]+rowdata_
nw[1])/2,1)]
            listdata.append(new_rowdate)
    # 保存文件
    f = xlwt.Workbook() #创建工作簿
    sheet1 = f.add_sheet(u'sheet1',cell_overwrite_ok=True) #创建sheet
    for i in range(len(listdata)):
        for j in range(len(listdata[i])):
            sheet1.write(i,j,(listdata[i])[j])
    f.save(clrfile+'.xls')#保存文件

clear_data(file_a,listdata_a,clearedfile_a)
clear_data(file_b,listdata_b,clearedfile_b)

# 合并清洗后的a,b数据
df1 = pd.read_excel("a_copy.xls")
df2 = pd.read_excel("b_copy.xls")
df3 = pd.merge(df1,df2,how='outer',on='时间',sort='True')
print(df3)
df3.to_excel("MergeDates.xls",sheet_name='sheet1')
```

五、程序运行结果

运行程序的结果可参考图2.3.9。

```
>>>
= RESTART: C:\Users\Administrator\Desktop\testMergeExcle\VetherDataClear.py =
                    时间    温度    湿度  PM2.5
0     2016-02-25 12:00:00  12.6  47.5  33.5
1     2016-02-25 13:00:00  13.5  46.9  31.9
2     2016-02-25 14:00:00  14.2  43.2  37.4
3     2016-02-25 15:00:00  14.9  40.6  32.0
4     2016-02-25 16:00:00  14.6  40.6  33.0
5     2016-02-25 17:00:00  13.3  43.7  31.9
6     2016-02-25 18:00:00  12.0  46.1  35.4
7     2016-02-25 19:00:00  11.1  50.8  32.5
8     2016-02-25 20:00:00  10.1  54.2  42.0
9     2016-02-25 21:00:00   9.3  59.9  36.5
10    2016-02-25 22:00:00   8.8  64.7  30.6
11    2016-02-25 23:00:00   8.9  65.7  34.1
12    2016-02-26 00:00:00   8.7  67.9  37.6
13    2016-02-26 01:00:00   8.2  70.7  39.0
14    2016-02-26 02:00:00   7.1  72.7  33.2
15    2016-02-26 03:00:00   6.8  74.7  36.5
16    2016-02-26 04:00:00   7.3  73.2  46.0
17    2016-02-26 06:00:00   6.4  77.8   NaN
18    2016-02-26 07:00:00   5.8  81.6   NaN
19    2016-02-26 08:00:00   9.3  70.9  39.0
20    2016-02-26 09:00:00  12.2  62.7  33.8
21    2016-02-26 10:00:00  14.1  57.5  36.4
22    2016-02-26 11:00:00  16.3  51.4  38.9
23    2016-02-26 12:00:00  17.5  45.3  38.1
24    2016-02-26 13:00:00  19.0  41.1  33.5
25    2016-02-26 14:00:00  18.5  41.9  31.1
26    2016-02-26 15:00:00  18.3  41.7  33.4
27    2016-02-26 16:00:00  17.7  41.0  33.0
28    2016-02-26 17:00:00  16.4  48.4  33.3
29    2016-02-26 18:00:00  15.3  54.0  38.4
...                   ...   ...   ...   ...
5326  2016-12-30 10:00:00   7.0  68.3  30.9
5327  2016-12-30 11:00:00   8.4  64.1  31.2
```

图2.3.9　运行结果

微信公众号：电子课本大全

阅读拓展

数据预处理

数据预处理就是在对数据进行主要处理之前的处理，以利于计算机的运算。概括起来，数据预处理的过程包括数据审查、数据清洗、数据转换和数据验证四个环节。数据清洗前面已经介绍过，下面介绍另外三个环节。

·数据审查：检查数据的数量（记录数）是否满足分析的最低要求，字段值的内容是否与调查要求一致，是否全面等。

·数据转换：数据分析有时会造成数据不可比，如果统计指标的性质、计量单位不同，也容易引起评价结果出现较大误差，再加上分析过程中的其他一些要求，因此需要在分析前对数据进行转换，包括无量纲化处理、线性变换、汇总和聚集、适度概化、规范化以及属性构造等。

·数据验证：目的是初步评估和判断数据是否满足统计分析的需要，决定是否需要增加或减少数据量。

四个环节是一个逐步深入、由表及里的过程。最后是进一步检测数据内容是否满足分析需要，诊断数据的真实性及数据间的协调性等，确保优质的数据进入分析阶段。

项目实施

清洗所采集的交通运输数据中的噪声数据

一、项目活动

1. 小组讨论：所采集的交通运输数据中可能会出现什么样的噪声数据、产生的原因可能是什么。最后把讨论结果汇总到表2.3.3中。

表2.3.3　噪声数据分析

噪声数据情况	导　致　的　原　因
有个别值偏大了近10倍	

2. 编程对数据进行清洗，消除噪声数据，并写出清洗的大体步骤。

二、项目检查

1. 每个成员在组内汇报自己清洗数据的结果。

2. 组长把本章项目学习各个阶段中，每人的成果汇总到一个文档里，并提交给老师，由老师做出评价。

微信公众号：电子课本大全

练习提升

1. 在互联网数据中，哪些是非结构化数据？请举例说明。
2. 什么是噪声数据？哪些原因会导致噪声数据的出现？
3. 数据清洗的目的是什么？主要包含哪几个步骤？
4. 用Python语言编写程序，对数据进行清洗，主要用到哪几个关键的语句？
5. Pandas库的主要功能是什么？导入Pandas库的语句是什么？
6. 阅读以下材料，思考结构化数据和非结构化数据有什么联系？

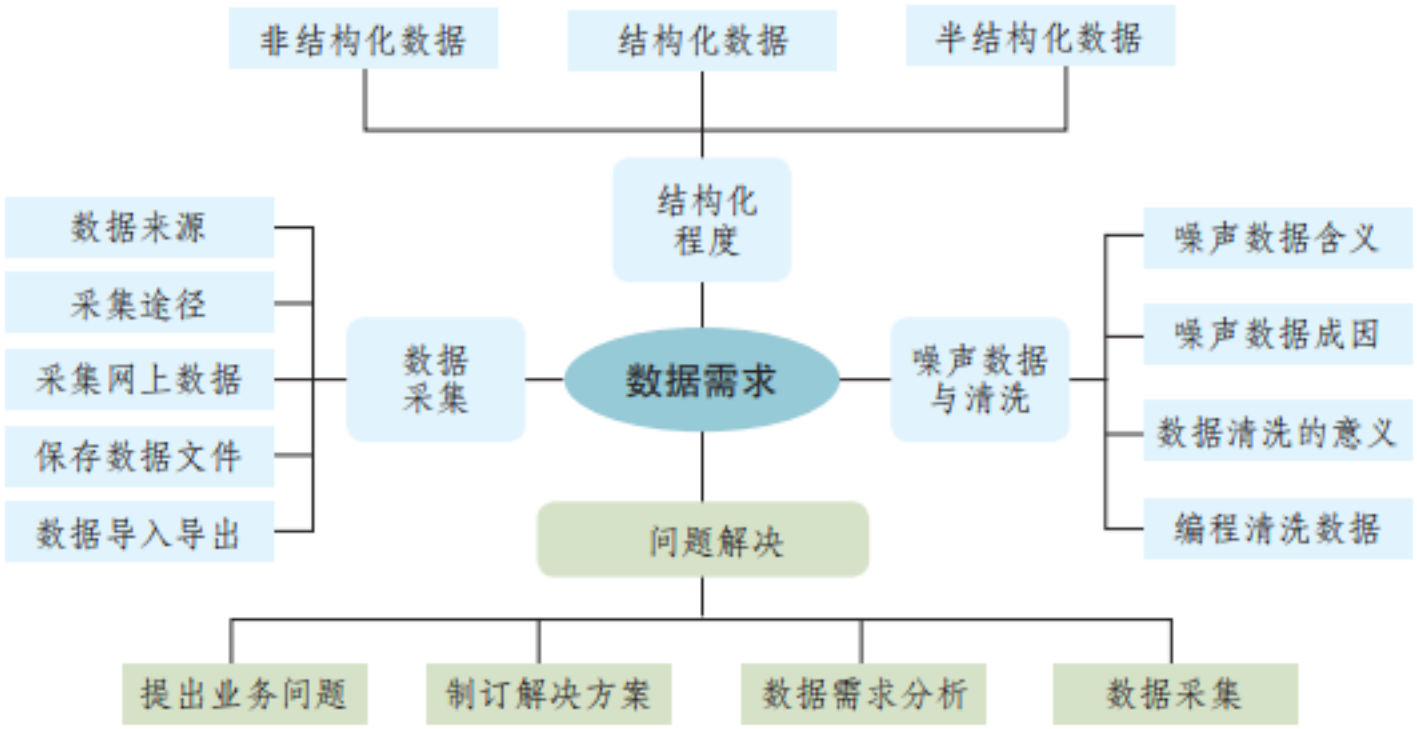
> 有观点认为，在移动互联网时代，数据的一大特征就是非结构化。数据是结构化还是非结构化，是一个相对的概念。例如，一般认为中文文本是非结构化数据，但是通过分词后，一个文档常常可以通过一个超高维的、关于词频的稀疏向量来表达。文档向量化后就不再是非结构化的了。对中文文本、网络结构、图像等一系列非结构化数据，通过基于人工智能的语义分析技术，可以获得很多有趣的信息。例如，有人对全唐诗进行深度的文本数据分析，得到了全唐诗中最常见的148个字排序表，找出了51个古语中常用的、表示颜色的单字，还根据诗歌中常见的主题类别总结出诗词的情绪类别。

7. 查阅资料了解：目前有哪些处理技术可以将非结构化数据转换为结构化数据。

微信公众号：电子课本大全

总结 评价

1. 下图展示了本章的核心概念与关键能力，请同学们对照图中的内容进行总结。



2. 根据自己的掌握情况填写下表。

学习内容	掌握程度		
数据采集的概念	□ 不了解	□ 了解	□ 理解
数据采集的途径	□ 不了解	□ 了解	□ 理解
创建数据文档	□ 不会	□ 会	□ 熟练
利用网络爬虫采集网络数据	□ 不会	□ 会	□ 熟练
导入和导出数据	□ 不会	□ 会	□ 熟练
三种不同结构化程度数据的特点	□ 不了解	□ 了解	□ 理解
噪声数据的含义	□ 不了解	□ 了解	□ 理解
噪声数据的成因	□ 不了解	□ 了解	□ 理解
编写程序清洗数据	□ 不会	□ 会	□ 熟练

3. 回答以下几个问题，对自己的学习情况进行总结与反思。

(1)是否熟悉常用的数据采集工具?

(2)在编写程序方面是否感到困难?自己是如何克服的?

微信公众号：电子课本大全

第 3 章
数据管理

数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的过程，其目的在于充分有效地发挥数据的作用。数据管理目前在整个社会中有广泛的应用，大到政务管理、国防科研、航天科技，小到校园一卡通、图书管理系统等。本章介绍目前应用广泛的数据管理技术——关系数据库（本书所提的数据库默认是关系数据库）系统。从数据库管理工具、数据库的创建与维护，到利用结构化查询语言（structured query language，SQL）实现数据的查询，再到数据库的备份和还原，从而认识数据库在数据管理中的重要作用。



微信公众号：电子课本大全

03 主题学习项目：数据管理助规划

项目目标

本项目中，要求在一定范围内调研和管理高中生在选择大学、专业方面的相关数据，为后期查看数据、获取信息打下基础，从而更好地帮助他们规划未来。在这个项目学习过程中，要达到以下目标。

1. 采集学生信息、大学信息、专业信息等数据，进一步掌握数据采集的方法。

2. 能根据需求设计管理数据的数据库，并建立概念数据模型。

3. 能创建数据库和数据表，掌握备份数据的常用方法。

项目准备

为了完成项目，需要做以下准备。

- 组建学习小组。开展学习过程中，小组成员要互相讨论，独立思考、共同协作。
- 掌握数据采集的途径和方法。数据需要采用问卷调查和网络爬虫两种方式来获取。
- 安装和熟悉一种数据管理工具。了解工具的操作界面和基本功能。
- 简要回顾Python的常用语法。搭建Python开发环境，熟悉其编程界面和常用语句的格式。

项目过程

为了保证顺利完成本项目的学习活动，在不同学习阶段，小组长要注意检查组员项目学习的进度，并做好协调互助工作。

1 确定数据库功能

了解设计数据库的基本流程，确定数据库的功能，用E-R图呈现数据表之间的关联，并适当优化。

P72

2 建立关系模型

建立数据的关系模型，创建数据库和数据表，输入调研数据。在这个过程中掌握数据类型和相关操作方法。

P83

3 改进方案

用SQL实现对数据的查询和提取，为后期的数据分析、信息获取打下基础。

P95

4 完善方案

通过查阅资料、分析和研讨，了解数据备份的重要性；学会用数据管理工具备份数据。

P103

项目总结

通过本章的项目学习，加深对数据管理的认识，学会使用数据管理工具创建数据库，并利用结构化查询语言实现对数据的查询和提取及备份，为后阶段的数据分析做准备。

微信公众号：电子课本大全

3.1
数据库与数据管理

学习目标 ▶▶▶

- 认识到数据库系统是管理数据的重要方式。
- 掌握建立概念模型和关系模型的基本方法。
- 掌握一种数据库管理系统的基本操作方法。

体验探索

图书管理系统探秘

现在很多学校的图书馆都实现了信息化、系统化管理。同学们能直观感受到信息技术带来的便捷。通过图书管理系统能便捷地查询图书信息、借阅图书（图3.1.1）。但系统要实现这些功能，就需要对图书、读者等数据进行科学有效的管理。这其中用到的数据管理方式就是数据库系统。

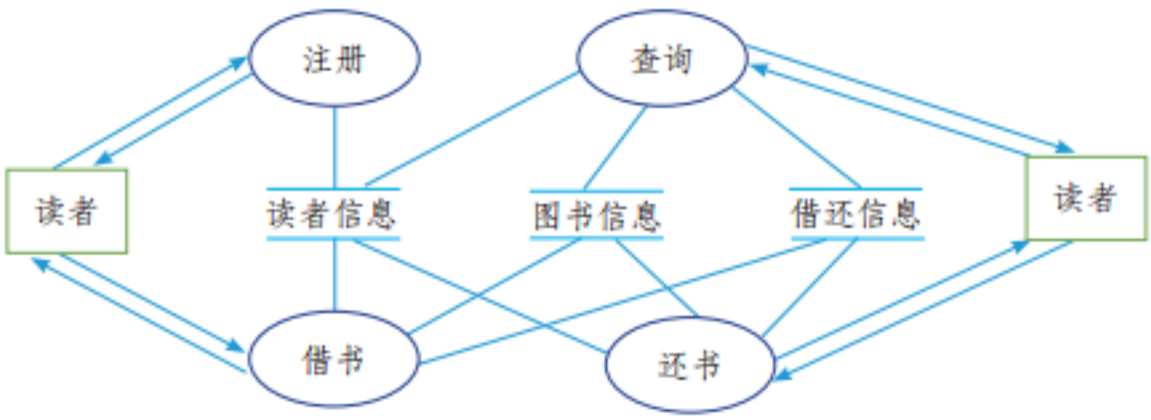


图3.1.1　一个图书管理系统提供给读者的基本功能

学习小组成员去图书馆向系统维护员请教以下问题。

1. 图书管理系统中包含哪些基本数据？这些数据是以什么样的形式存储的？

2. 开发类似的图书管理系统可以使用哪些软件？

微信公众号：电子课本大全

3.1.1 数据库与数据库管理系统

数据、数据库、数据库管理系统、数据库系统是几个经常遇到的概念，但对这些概念的解释有多种，同学们可以查阅资料深入了解。

简单来说，数据库是按照数据结构来组织、存储和管理数据的“仓库”。严格来说，数据库是长期存储在计算机内、有组织的、可共享的数据集合。在日常工作中，常常需要把某些相关的数据放进这样的集合中，并根据管理的需要进行相应的处理。数据库管理系统是一组软件，负责数据库的访问、存取、维护、管理和控制。用户对数据库的各种操作请求，均由数据库管理系统来完成，它提供了数据库操作的环境。



实践活动

了解常用的数据库管理系统

1. 上网查阅相关资料，了解数据库与数据库管理系统的发展历史以及常用的数据库管理系统，并填表3.1.1。

表3.1.1 常用的数据库管理系统

系统名称	简单介绍
MySQL	
Oracle	
Microsoft SQL Server	

2. 下载并安装MySQL，然后登录其控制台，初步认识命令输入的格式。



阅读拓展

大数据的处理

大数据的处理，需要由专门设计的硬件和软件工具来进行。

· Hadoop是一个能够对大量数据进行分布式处理的软件框架。Hadoop的框架最核心的设计就是HDFS（Hadoop Distributed File System）和MapReduce。HDFS为海量的数据提供存储功能，而MapReduce为海量的数据提供计算功能。

· Hive是一个建立在Hadoop上的开源数据仓库基础设施，通过Hive可以对数据进行抽取、转换和加载，以及结构化处理，并对Hadoop上大数据文件进行查询和处理等。Hive提供了一种简单的类似SQL的查询语言HiveQL，这为熟悉SQL的用户查询数据提供了方便。

微信公众号：电子课本大全

3.1.2 确定数据库的基本功能

获取数据之后，可以利用数据库管理系统设计和创建数据库，从而对数据进行科学有效的管理。在目前初学阶段，我们建立数据库可按图3.1.2所示的过程进行。首先是对已有数据进行需求分析，明确用户对管理数据有何需求，从而确定数据库的基本功能；然后设计数据概念模型，把用户需求转化为概念模型；接着把概念模型转换为关系模型；最后利用数据库管理系统创建数据库、运行维护数据库，不断调试和修改，最终得到相对完善的数据库。

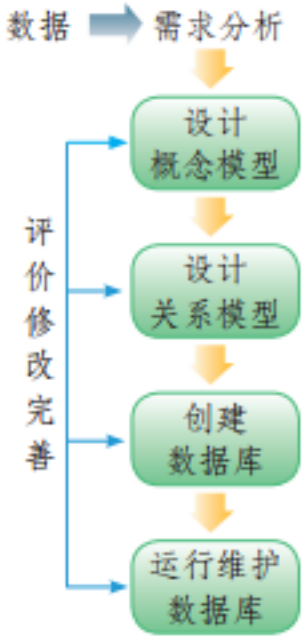


图3.1.2 大体过程

数据库的功能往往是在数据库计划建立时就已有初步的预设，在后期调研中获得用户需求后确定下来。例如，图书资料部门为了有效管理各种图书数据，需要建立一个图书管理数据库，然后就会调查了解读者对这个数据库有哪些要求，从而构建了该数据库的基本功能。在本章的项目学习中，希望通过了解高中生对大学以及专业的选择意向，帮助大家树立一个目标，激发学习动力。在为本章项目设计“学生专业规划”数据库时，希望其具备图3.1.3所示的基本功能。

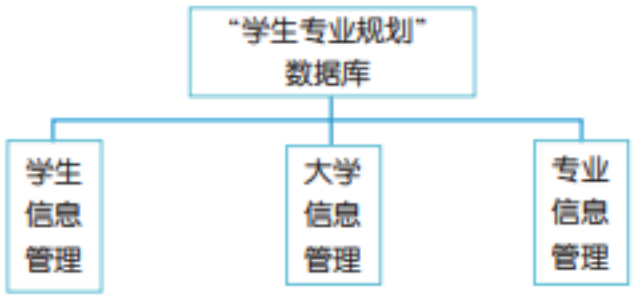


图3.1.3 数据库基本功能模块

本节的需求分析建立在已有数据基础之上；物理结构设计是为数据逻辑结构模型选取一个最适合应用环境的物理结构（包括存储结构和存取方法），本书不涉及该内容。



实践活动

数据库功能分析

“学生专业规划”数据库是为更有效地了解、管理学生对大学和专业的选择意向而建立的数据库，那么你希望数据库中所包含的三个管理模块分别承担什么功能呢？请按照你自己的设计填写表3.1.2。

表3.1.2 “学生专业规划”数据库功能分析

基本功能模块	具体功能介绍
学生信息管理	
大学信息管理	
专业信息管理	

微信公众号：电子课本大全

根据数据库的功能需求，可以进一步细化数据库的基本功能。例如，在“学生专业规划”数据库中，可以进一步设计出预期的功能，大体如图3.1.4所示。

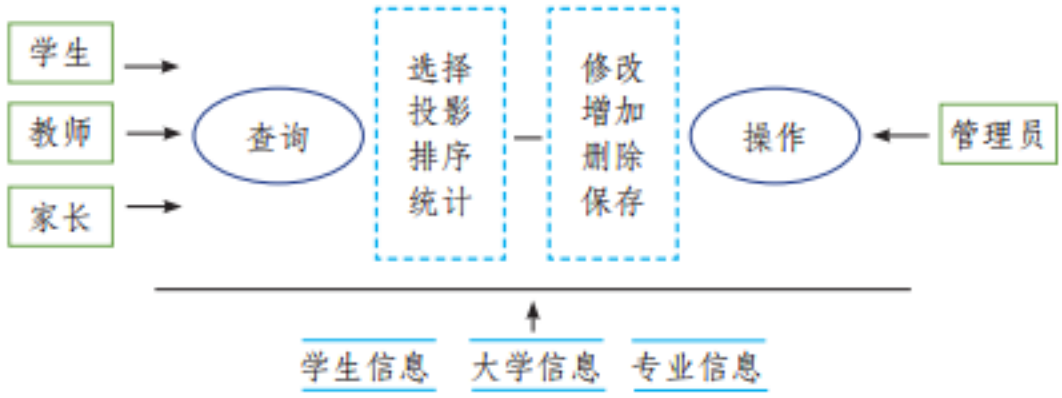


图3.1.4　数据库细化功能图

思考活动

用于大数据的数据库管理系统

与当今大数据时代的数据量比起来，同学们目前所要管理的数据规模并不大，利用数据库技术就能把这些数据存储并组织起来。如果要管理的是海量数据，就需要通过大数据技术进行高效的存储管理。在大数据应用中，在海量空间数据场景下，需要通过空间索引技术将节点内的数据尽可能地保持其完整性和独立性，尽量避免节点间的数据交换，从而实现高效的分布式计算。下面简单介绍几种可用于大数据处理的系统。

1. PostgreSQL是一个开源的关系数据库管理系统，可以支持主流的操作系统。它支持外键、连接、视图、触发器和存储过程。

2. MongoDB是一个基于分布式存储的NoSQL文档数据库管理系统，介于关系数据库与非关系数据库之间。

3. HDFS是一种分布式文件系统，由于它是Hadoop生态体系的基石，且与Spark技术无缝结合，因此在大数据管理时，经常被作为一种通用型存储方案来使用。

思考：查阅资料，了解更多的大数据管理系统，并想一想，目前所学的数据库知识会为以后学习大数据相关知识带来哪些帮助。

3.1.3　建立概念数据模型

有了用数据库管理数据的思路，下面就要建立起信息世界的概念数据模型。概念数据模型简称概念模型，是面向数据库用户的现实世界的模型，主要用来描述信息世界的概念化结构，它使数据库的设计人员在设计的初始阶段，可以摆脱计算机系统及数据库管理系统的具

微信公众号：电子课本大全

体技术问题，集中精力分析数据以及数据之间的联系等，与具体的数据库管理系统无关。概念数据模型用于信息世界的建模，是现实世界到信息世界的第一次抽象，是数据库设计人员进行数据库设计的有力工具，也是设计人员和用户之间交流的语言。

E-R模型是20世纪提出的一种语义模型。由于它能将现实世界中概念的含义和相互关联映射到数据库概念模型中，因此许多数据库设计工具都利用了E-R模型的概念。

E-R图

描述信息世界的概念数据模型最常用的方法是绘制E-R图。E-R图也叫实体联系图，它通过图示将实体以及实体间的联系描述出来，为客观事物建立概念数据模型。图3.1.5就是一个E-R图。

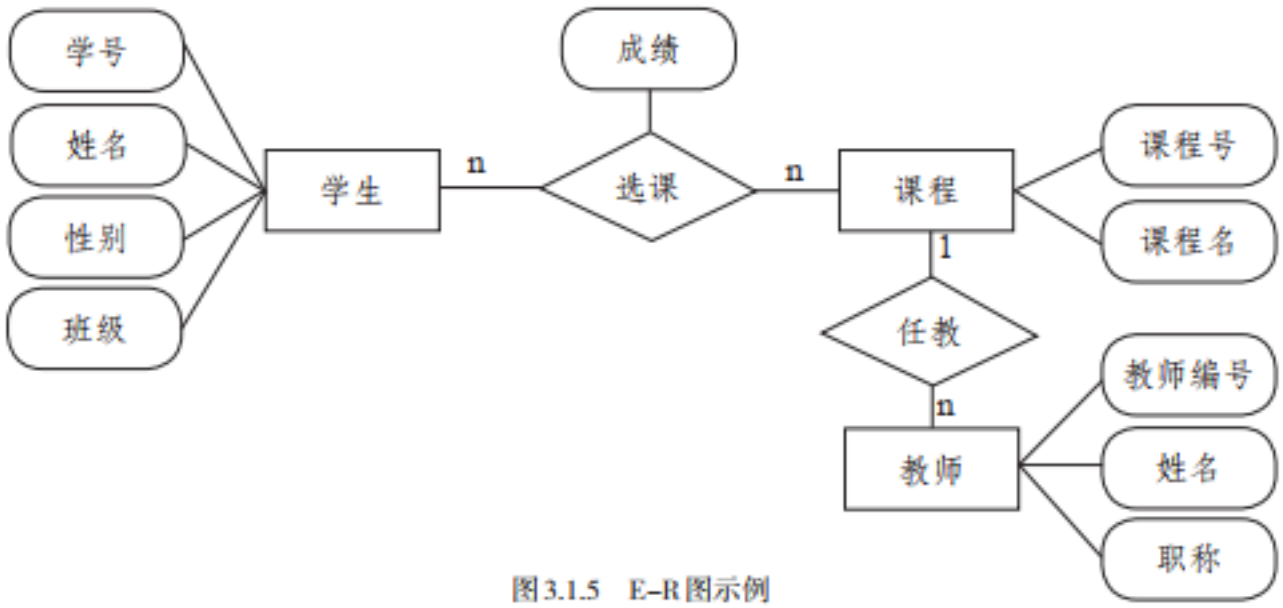


图3.1.5　E-R图示例

E-R图中要用四种图形来描述，具体见表3.1.3。

表3.1.3　E-R图中各种图形的含义

图形	含　义
矩形框	表示实体，在框中记入实体名
菱形框	表示联系，在框中记入联系名
椭圆框	表示实体或联系的属性，将属性名记入框中。
直线	连接实体与属性、实体与联系、联系与属性，并在直线上标注联系的类型

实体

实体是现实世界中客观存在并可相互区分的人或事物。实体可以是实际的对象，也可以是抽象的概念（如事物之间的联系）。例如，学生、教师、课程、班委、工厂、职工、产品等都是实体。

微信公众号：电子课本大全

属性

属性是实体所具有的特性。一个实体可以由若干属性来刻画。例如：教师作为一个实体，教师的属性有教师编号、姓名、性别、籍贯、工龄等；图书作为一个实体，它的属性有编号、书名、类别、出版社等。

实体间的联系

实体之间的联系可分为一对一、一对多、多对多三种，具体说明见表3.1.4。

表3.1.4　实体之间的联系

联系类型	举例说明	图　示
一对一联系 记成1 ： 1	班长和班级，一个班长任职于一个班级，一个班级有一位班长	
一对多联系 记成1 ： n	班级和学生，一个班级有多个学生，一个学生属于一个班级	
多对多联系 记成m ： n	学生和课程，一个学生可以选择多门课程，一门课程可以被多个学生选择	

实体与属性之间、实体与联系之间、联系与属性之间用直线相连，并在直线上标注联系的类型。对于一对一联系，要在两个实体连线方向各写1；对于一对多联系，要在一的一方写1，多的一方写n；对于多对多关系，则要在两个实体连线方向分别写n、m。两个实体之间的这三种联系总结如图3.1.6所示。

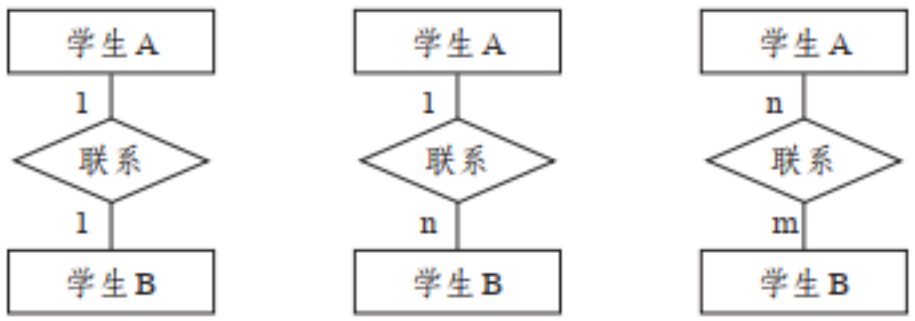


图3.1.6　实体之间的三种联系

微信公众号：电子课本大全

实践活动

画E-R图

根据你对学校图书馆数据管理系统中数据库的了解，列出图书、学生两个实体可能包含的属性，并画出E-R图。

E-R图的优化

初步建立的概念模型并不一定是最优的，可能还存在冗余的数据，因此要对E-R图进行优化。例如，在项目“学生专业规划”数据库中，最开始建立了如图3.1.7所示的学生E-R图。

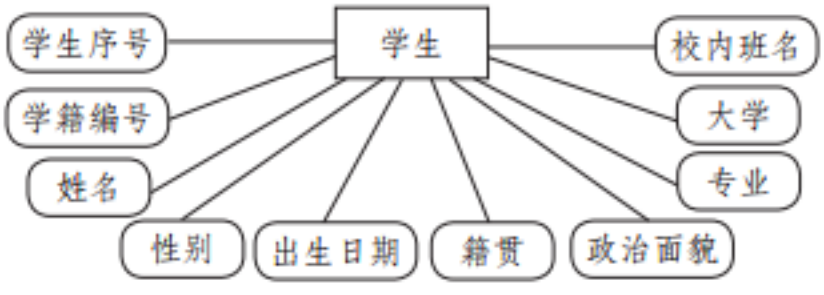


图3.1.7　最初的学生E-R图

对E-R图进行优化时要注意，属性不再具有需要进一步描述的性质，属性也不能与其他实体有联系。图3.1.7中的专业还可以有更加细分的属性（如专业序号、专业名称、专业代码），因此应该把专业作为一个单独的实体列出来；大学也有更加细分的属性，如序号、院校代码、学校名称、主管部门、所在地、办学层次等，所以大学也要作为一个单独的实体。

确定实体之后，需要考虑任意两个实体之间是否存在联系。在确定联系时要尽量取消冗余的联系。例如，本章项目中，一个学生可以确定一个专业，一个专业可以有很多学生选择，因此确定实体“学生”和“专业”之间存在一对多联系。一个专业可以属于不同的大学，一个大学也可以拥有不同的专业，所以实体“专业”和实体“大学”之间存在多对多的联系。优化后的E-R图如图3.1.8所示。

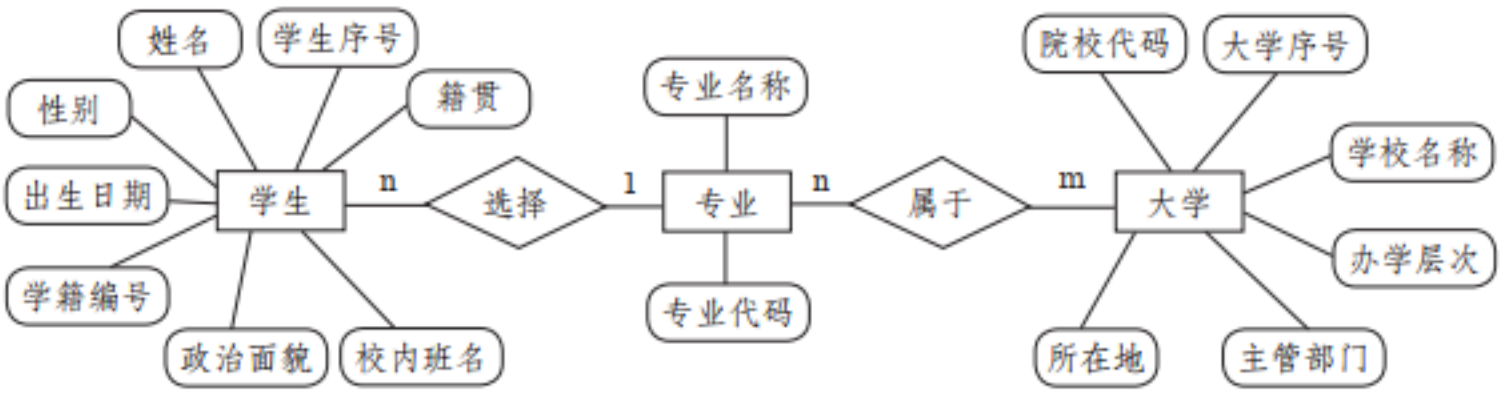


图3.1.8　优化后的E-R图

微信公众号：电子课本大全

实践活动

优化E-R图

打开名为“学校食堂数据库介绍”的文档，阅读和观察文档中关于食材、菜谱的E-R图，请把它进一步优化。

项目实施

“学生专业规划”数据库的数据调研与建模

一、项目目的

高中是人生中学习和未来专业规划紧密结合的阶段。同学们怀有对大学、对未来职业的设想。通过建立高中生专业规划数据库，可以实现对学生、大学及专业等数据的管理和查询，从而增强对大学招生、专业设置、同学意向的了解。

二、项目活动

1. 数据调研。采用适当的调研方法，在一定范围内获取数据，并填写表3.1.5。

表3.1.5 “学生专业规划”数据库系统功能分析

调研内容	范围（或数量）	方法（工具）	具体内容	小组分工形式
学生数据	高一或高二	利用问卷进行面谈或网络访谈	姓名 性别	
大学数据		网络		
专业数据				

2. 建立数据模型。首先设计数据库的基本功能，然后建立学生、大学、专业三个数据模型，并绘制在下面空白处。

学生E-R图　　大学E-R图　　专业E-R图

三、项目检查

1. 简要汇报数据调研阶段的工作成果。

2. 展示数据库设计的一些思路，以及所设计的数据库功能模块和数据模型。

微信公众号：电子课本大全

练习提升

1. 用“数据库系统”“数据管理”等关键词上网查阅资料，进一步了解用数据库系统管理数据的一些特点。

2. 根据自己的理解，通过表格的形式描述数据库、数据库管理系统和数据库系统三者之间的关系。

3. 请介绍你所熟悉的一种数据库管理系统的主要功能。

4. 说出建立数据库的大体过程以及各个阶段的主要工作。

5. 说出E-R图的作用以及各种图形框的含义。

6. 请结合实例，在表3.1.6中描述实体之间的三种联系。

表3.1.6　实体之间的三种联系

联系类型	举例说明	图　示
一对一联系		
一对多联系		
多对多联系		

微信公众号：电子课本大全

3.2
设计逻辑结构与建立数据库

学习目标 ▶▶▶

- 掌握设计简单数据库逻辑结构的基本方法。
- 理解MySQL数据库管理系统提供的数据类型。
- 掌握创建数据库和数据表（包括输入数据）的方法。
- 认识到数据库是管理数据的一种途径，增强科学有效管理数据的意识。

体验探索

学生信息管理系统中的数据库

与学生信息相关的管理系统有很多，如选课系统、学籍管理系统、学生成绩管理系统等（图3.2.1），它们虽然侧重点有所不同，但基本功能都是管理学生的基本信息和课程学习情况。在这些系统里，学生的基本信息、课程、成绩、课余活动等数据都被存放在数据库中。



图3.2.1　学生成绩管理系统

1. 尝试登录与自己相关的信息管理系统，体验它的主要功能。

2. 利用课余时间，走进学校教务处，请老师展示学籍管理系统的主要功能并介绍系统包含哪些数据库。

3. 初步思考：系统中的数据库是如何建立的？数据是如何录入的？

微信公众号：电子课本大全

在体验探索各种学生信息管理系统的过程中，可以大体感受到系统中数据管理的便利性。另外，在前面的学习中，我们已经建立了信息世界的概念模型，并用E-R图表示出来，接下来要把它们转换为关系模型，并用数据库管理系统建立相应的数据库。

3.2.1 概念模型转换为关系模型

关系模型是指用二维表的形式表示实体和实体之间联系的数据模型。关系模型目前应用很广泛，而且当前的数据库系统多为基于关系模型的关系数据库系统。

关系模式是对关系的逻辑结构和特征的描述，一般表示为：关系名（属性1，属性2，…，属性n）。例如，根据足球世界杯的一些数据（表3.2.1），世界杯实体的关系模式可以表示为“世界杯（届次，年份，地点，冠军）”。进行数据库的逻辑结构设计，主要是将概念模型设计中的E-R图转换成关系模型，即将实体、实体的属性和实体之间的联系转化为关系模型。其中实体和联系都可以表示成关系，E-R图中的属性可以转换成关系的属性。

表3.2.1 足球世界杯一些数据

届次	年份	地点	冠军
18	2006	德国	意大利
19	2010	南非	西班牙
20	2014	巴西	德国
21	2018	俄罗斯	法国

主键的值能唯一标识表中的每一行。例如，在学生信息表（学号，姓名，性别，班级）中，学号能唯一标记每一条记录，所以是一个主键。

实体的转换

一个实体转换为一个关系模式，实体的属性就是关系的属性，实体的主键就是关系的主键。

例如，在“学生专业规划”数据库中，把E-R图中实体转换为关系模式可以表示如下（带下画线的属性表示实体的主键）：

学生（学生序号、学籍编号、校内班名、姓名、性别、籍贯、出生日期、政治面貌）
专业（专业代码、专业名称）
大学（大学序号、院校代码、学校名称、主管部门、所在地、办学层次）

联系的转换

对于实体之间的联系，有以下几种情况。

一个m：n联系的转换。这种联系要转换为一个独立的关系模式，与该联系相连的各实体的主键以及联系本身的属性转换为关系的属性，该关系的主键为各实体主键的组合。

微信公众号：电子课本大全

1 ： n联系的转换。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端实体所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主键以及联系本身的属性转换为关系的属性，n端实体的主键为该关系的主键。一般情况下，1 ： n联系不转换为一个独立的关系模式。

1 ： 1联系的转换。一个1 ： 1联系可以转换为一个独立的关系模式，也可以与任意一端实体所对应的关系模式合并。一般情况下，1 ： 1联系不转换为一个独立的关系模式。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主键以及联系本身的属性转换为关系的属性，每个实体的主键均可作为该关系的主键。如果是与联系的任意一端实体所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实体的主键和联系本身的属性。

例如，数据库“学生专业规划”E-R图中的三个实体（学生、专业、大学）分别转换成以下三个关系模式，一个m ： n联系要转换为一个独立的关系模式（表3.2.2）。

表3.2.2　转换后的关系模式

实体名	关系模式
学生	学生（学生序号、学籍编号、校内班名、姓名、性别、籍贯、出生日期、政治面貌，专业代码）
专业	专业（专业代码、专业名称）
属于	属于（专业代码、大学序号）
大学	大学（大学序号、院校代码、学校名称、主管部门、所在地、办学层次）

注：表中带下画线的属性为实体的主键。



实践活动

把E-R图转换为关系模型

通过关系转换和优化后一般可以得到一些基本表。根据名为“图书管理E-R图”文档中的提示，把相关E-R图转换为关系模型，并设计对应的数据表结构。

3.2.2　创建和查看数据库

用MySQL创建数据库的方法主要有两种：一种是在MySQL控制台中输入命令语句；另一种是采用图形化界面的数据管理工具。采用命令语句有助于理解所要执行任务的目的，利于提高编程能力。

微信公众号：电子课本大全

采用图形化界面，操作直观简单。

图形化操作界面的数据库管理工具有很多种，Navicat for MySQL是其中常用的一种。

在Navicat for MySQL中创建数据库的方法如下。

右击“连接树”，在快捷菜单中单击“新建数据库”（图3.2.2）。在对话框中设置数据库，如输入数据库名称、选择字符集和排序规则（图3.2.3），最后单击“确定”按钮。

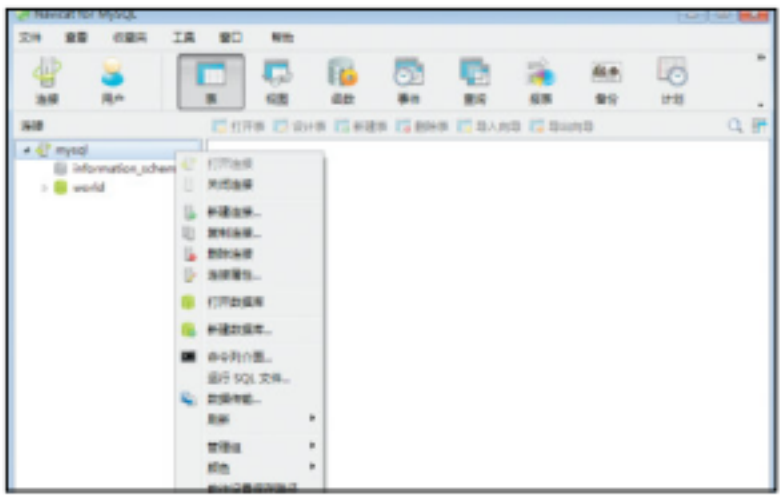


图3.2.2　Navicat for MySQL界面



图3.2.3　Navicat for MySQL对话框

字符集是用来定义字符在数据库中的编码的集合。常见的字符集有GB 2312（简体中文的编码）、GBK（简体中文及繁体中文编码）、BIG5（繁体中文）、UTF-8等。不同字符集编码格式不同，在编写应用程序或者网页应用操作数据库时遇到的乱码，一般是由于调用的字符集不同而导致的，像MySQL和Oracle都会有字符集问题。

首先，登录MySQL控制台（窗口界面类似图3.2.4所示），然后输入要执行操作对应的命令语句。

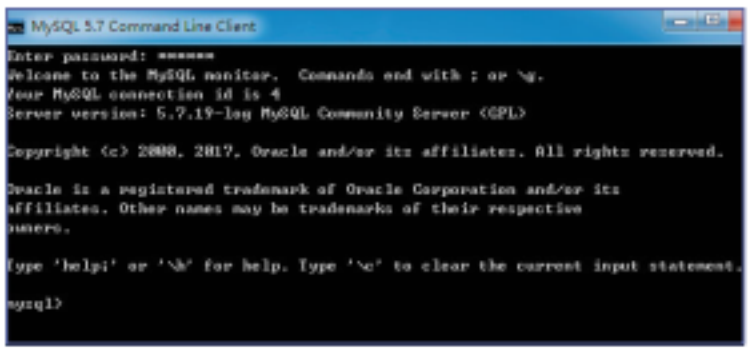


图3.2.4　MySQL命令语句输入界面

输入以下语句，可以查看MySQL支持的字符集。

```
SHOW CHARACTER SET;
```

查看当前关于MySQL各种字符集的语句如下。

```
SHOW VARIABLES LIKE 'character%';
```

MySQL是一个关系数据库管理系统，它所使用的SQL是用于访问数据库的常用标准化语言。

微信公众号：电子课本大全

排序规则是指字符比较时是否区分英文字母大小写，以及是按照字符编码进行比较还是直接用二进制数据比较。实际操作时，一般按默认方式处理。

前面用菜单命令创建的数据库，也可以用SQL语句进行创建。例如，要创建名为database_stu的数据库，可以输入以下语句。

```
CREATE DATABASE database_stu;
```

阅读拓展

MySQL语句的输入

在MySQL控制台中输入语句时，每个语句以分号结束。如果输入语句后敲回车键，发现忘记输入分号，这时并不需要重输一遍语句，只要输入分号后敲回车键即可。

MySQL数据库一旦创建成功，创建的编码也就确定了。如果想修改数据库的属性，可以在Navicat for MySQL中“连接树”里选择相应的数据库并右击，在弹出的快捷菜单中选择“数据库属性”，然后在数据库属性中修改数据库（图3.2.5）。也可以用命令语句实现该操作。例如，要选择名为database_stu的数据库，所用的SQL语句如下。

```
USE database_stu;
```

删除数据库是将数据库系统中已经存在的数据库删除。一旦删除数据库，该数据库中的所有数据都将被清除，原来分配的存储空间也会被回收。

删除名为database_stu的数据库，所用的SQL语句如下。

```
DROP DATABASE database_stu;
```

实践活动

创建图书管理数据库

创建一个图书管理数据库，然后在表3.2.3中总结自己的操作体会。

表3.2.3　创建图书管理数据库的总结

总结项目	简要描述
创建的方法	
创建的步骤	
操作的难点	

微信公众号：电子课本大全

3.2.3 MySQL的数据类型

要充实数据库的内容，就需要往数据库里输入数据，也就是要创建数据表并输入数据。创建数据表时，首先要定义数据表的结构，这项工作主要包括设置数据表中各个字段的名称，确定各个字段的数据类型，确定数据表的主键。下面主要介绍数值型、字符串型、日期和时间型三种数据。

数值型

数值型数据可大致划分为两种：整数、浮点数或小数。MySQL允许指定数值字段中的值为正、负或者用零填补。表3.2.4列出了各种数值类型以及它们的允许范围和占用的内存空间。

表3.2.4 数值类型表

类 型	大小/字节	范围（有符号）	范围（无符号）	用 途
TINYINT	1	(−128，127)	(0，255)	小整数值
SMALLINT	2	(−32 768，32 767)	(0，65 535)	大整数值
MEDIUMINT	3	(−8 388 608，8 388 607)	(0，16 777 215)	大整数值
INT 或INTEGER	4	(−2 147 483 648，2 147 483 647)	(0，4 294 967 295)	大整数值
BIGINT	8	(−9 233 372 036 854 775 808， 9 223 372 036 854 775 807)	(0，18 446 744 073 709 551 615)	极大 整数值
FLOAT	4	(−3.402 823 466 E+38，1.175 494 351 E−38)	0，(1.175 494 351 E−38， 3.402 823 466 E+38)	单精度 浮点数值
DOUBLE	8	(1.797 693 134 862 315 7 E+308， 2.225 073 858 507 201 4 E−308)	0，(2.225 073 858 507 201 4 E−308， 1.797 693 134 862 315 7 E+308)	双精度 浮点数值
DECIMAL (M，D)	如果M>D，为M+2；否则为D+2	依赖M和D的值	依赖M和D的值	小数值

字符串类型

MySQL提供了八种基本的字符串型数据，可以存储的范围从简单的一个字符到文本块或二进制字符串数据（表3.2.5）。

微信公众号：电子课本大全

表3.2.5　字符串类型表

类　型	大小/字节	用　途
CHAR	0 － 255	定长字符串
VARCHAR	0 － 65 535	变长字符串
TINYBLOB	0 － 255	不超过255个字符的二进制字符串
TINYTEXT	0 － 255	短文本字符串
BLOB	0 － 65 535	二进制形式的长文本数据
TEXT	0 － 65 535	长文本数据
MEDIUMBLOB	0 － 16 777 215	二进制形式的中等长度文本数据
MEDIUMTEXT	0 － 16 777 215	中等长度文本数据
LONGBLOB	0 － 4 294 967 295	二进制形式的极大文本数据
LONGTEXT	0 － 4 294 967 295	极大文本数据

日期和时间型

在处理日期和时间型的数据时，MySQL中有五种数据类型可供选择（表3.2.6）。子类型在每个分类型中都可以使用，并且MySQL带有内置功能，可以把多样化的输入格式自动转变为标准格式。

表3.2.6　日期和时间类型表

类　型	大小/字节	格　式	用　途
DATE	3	YYYY-MM-DD	日期值
TIME	3	HH:MM:SS	时间值或持续时间
YEAR	1	YYYY	年份值
DATETIME	8	YYYY-MM-DD HH:MM:SS	混合日期和时间值
TIMESTAMP	8	YYYYMMDD HHMMSS	混合日期和时间值，时间戳

例如，在“学生专业规划”数据库中，数据库由多个相互关联的数据表组成。根据之前得到的学生关系模式：学生（学号，校内班名，姓名，性别，籍贯，出生日期，政治面貌，专业代码），将其转换为数据表中具体的字段。字段名具有唯一性和描述性，可以和原来的属性名一致，也可以另外取名字。

根据“学生”数据的特征，数据库中学生信息表中的字段名、数据类型、主键设定如表3.2.7所示。

微信公众号：电子课本大全

表3.2.7 “学生信息”数据表结构

字段名	数据类型	说　明
学生信息表ID	INT	主键
学籍编号	VARCHAR	
校内班名	VARCHAR	
姓名	VARCHAR	
性别	TINYINT	
出生日期	DATE	
政治面貌	VARCHAR	
籍贯	VARCHAR	
专业代码	VARCHAR	

字段名除了ID表示“序号”，其他数据表中的字段名都采用原来的中文名称，这样设计可读性高并且易于理解。

3.2.4　创建和查看数据表

设计数据表的结构后，可以在数据库中创建数据表。数据库管理系统中往往存在多个数据库，在操作之前要确定是哪一个数据库。

例如，要选择名为data_book的数据库，对应的SQL语句如下。

```
USE data_book;
```

要创建名为t_student的数据表，对应的SQL语句如下。

```
CREATE TABLE t_student;
```

在Navicat for MySQL中创建数据表的步骤是：在“连接树”中选择数据库，在“主工具栏”中选择“表”，单击“对象列表工具栏”中的“新建表”。

通过单击“添加栏位”来增加一个字段，然后在“类型”列表中选择相应的数据类型和设置长度。如果不允许字段为空，则不选择“允许空值”。另外，还可以设定某个字段为主键。主键用来唯一标识数据表中的每条记录，一个数据库中往往包含多张数据表，需要通过主键建立表之间的关系，使各表协同工作（图3.2.5）。

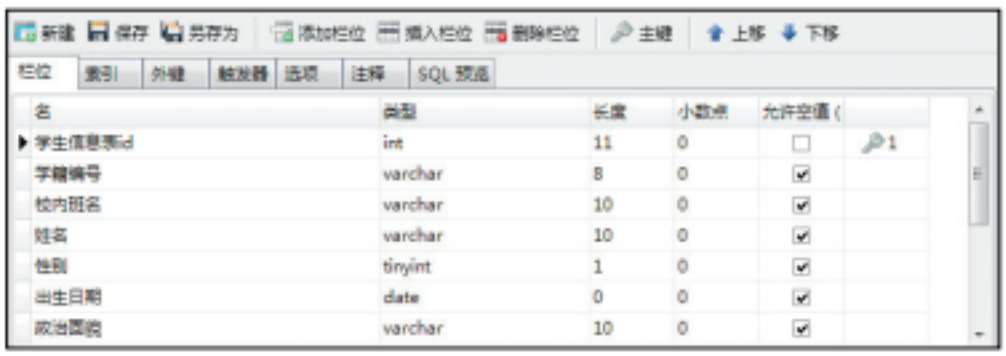


图3.2.5　创建的数据表

微信公众号：电子课本大全

思考活动

关于数据表名称的思考

在同一个数据库中，能否创建数据表名称相同的两张数据表？请说出理由。

3.2.5 修改和删除数据表

在实际操作过程中，可能要对表的结构进行一些修改。例如，要修改某张数据表中字段的名称、数据类型、长度等，可在Navicat for MySQL界面的“连接树”中选择“表”，再单击“对象列表工具栏”中的“设计表”，就可以对数据表进行修改了。

删除数据表是指删除数据库中已经存在的数据表。在删除数据表的同时，数据表中存储的数据都将被删除。单击“对象列表工具栏”中的“删除表”即可删除数据表。也可以用命令语句来实现该操作。例如，要删除名为t_student的数据表，对应的SQL语句如下。

```
DROP TABLE t_student;
```

3.2.6 将数据输入数据表

数据表创建好之后，就可以录入或导入数据了。

在数据采集阶段，往往会把数据保存为电子表格文件或CSV文件。遇到这种情况，可以通过导入的方式把它们存进MySQL数据库中。

例如，在Navicat for MySQL中，将Excel格式的“大学信息”表数据导入“学生专业规划”数据库的“大学信息”数据表中，可单击“导入向导”，然后在“导入类型”中选择对应的文件格式，再按提示逐步操作。

也可以用SQL语句来实现数据的导入。例如，要在专业信息表中导入数据，该数据文件的格式是CSV，可以在MySQL命令行窗口中输入以下语句来完成。

```
mysql>Load data infile 'd:\zyxx.csv' into table 专业信息表 fields
terminated by ',' lines terminated by '\n';
Load data infile：从指定文件加载数据；
Into table ：导入到指定的数据表；
Fields terminated by ',': 指字段（列）的值以逗号',’为分隔符；
Lines terminated by '\n'; 指分行为换行；
```


微信公众号：电子课本大全

通过前面的操作可以知道，数据表对于整个数据库来说只是一个容器，而数据则是容器中的内容。数据库和数据表创建完成之后，就可以录入数据了。数据的输入方法与电子表格软件的操作方法大体类似，但要注意的是：在数据表中输入的数据必须与数据表结构中的数据类型一致。



实践活动

在图书管理数据库中创建数据表

1. 打开前面实践活动中创建的图书管理数据库文件，然后在数据库中创建“学生信息”数据表。

2. 根据前面绘制的图书信息数据表，在数据库中创建“图书信息”数据表，输入一些图书数据。

3. 在该数据库中增加一个“出版社信息”数据表，然后用输入命令语句的方法，在数据表中输入一些出版社的信息。



项目实施

创建“学生专业规划”数据库并导入数据

一、项目活动

1. 根据E-R图创建逻辑结构并适当优化。

2. 创建数据库和数据表，然后根据调研获得的数据，设置数据的类型；分别在对应的数据表中导入数据，并写出导入的方法。

二、项目检查

把数据库和数据表在班里进行交流和展示；根据自己的学习情况，利用表3.2.8进行自我总结。

表3.2.8　项目学习总结

总结内容	说　明
对图形界面工具的使用	
对命令语句的使用	
操作上的难点	
对数据库和数据表的理解	
数据导入对数据采集提出的要求	

微信公众号：电子课本大全

练习提升

1. 根据自己的理解，简要阐述：什么是概念模型？概念模型和关系模型有哪些联系？数据库和数据表之间有什么样的联系？

2. 你常用到MySQL中的哪些数据类型？请完善表3.2.9进行说明。

表3.2.9　MySQL中的数据类型

数据类型	举　例　说　明
数值型	

3. 数据表中设置主键有什么作用？

4. 如果误删某个数据表中的记录后，能否在Navicat for MySQL中进行撤销操作？

5. 请举例说明以下操作对应的SQL命令语句。

（1）创建数据库 _____

（2）删除数据库 _____

（3）打开数据库 _____

（4）创建数据表 _____

微信公众号：电子课本大全

3.3 结构化查询与提取

学习目标 ▶▶▶

- 了解结构化查询语言（SQL）及常用的查询方法。
- 学会使用SQL实现简单的数据查询。
- 了解使用程序设计语言（Python）调用SQL语句实现数据的提取。

体验探索

在知网平台与网购平台中查询数据

1. 访问中国知网的国学宝典资源库（图3.3.1），查找一本书并进行以下操作：设置一种查询条件，写出自己所用到的关键词，然后查找需要的书籍。

2. 访问一个网络购物平台（图3.3.2），查询一款自己喜欢的电子产品。注意观察平台提供的查询条件，并列出自己所使用的查询关键词。



图3.3.1　知网查询系统



图3.3.2　商品查询系统

讨论：这些大型数据库系统提供的查询功能是否便利？在自己设计的数据库中，能否完成类似的查询操作？

微信公众号：电子课本大全

3.3.1 结构化查询语言

一些具备相对完善功能的数据库往往数据量庞大。例如，一个年级的学生成绩数据库，学生数量少则几百人，多则上千人；一个超市的商品数据库，一般包含成千上万的品种记录；而一些公共大型的数据库更是包含海量的数据。如何快速便捷地管理和检索数据库中的数据，是数据库系统建设的重要任务。目前大部分关系数据库都支持结构化查询语言。

检索时，首先由用户根据业务要求发出SQL指令，数据库管理系统接收到指令后对数据库执行相应的操作，最后数据库管理系统将处理的结果返回给用户。

阅读拓展

SQL与国际标准

1974年，IBM公司的两位计算机科学家雷蒙德·F.博伊斯（Raymond F. Boyce）和多恩·钱伯林（Don Chamberlin）为了方便操作该实验室的大型数据库SYSTEM R，设计了SEQUEL语言，后来在SEQUEL的基础上发展成为SQL。1987年国际标准化组织（ISO）颁布了国际标准，SQL成为关系数据库的国际标准语言。

可以在Navicat for MySQL窗口中执行查询的相关命令来查询数据，也可以采用SQL语句创建查询。

例如，之前创建的“学生专业规划”数据库，里面包含了学生信息、大学信息、专业信息等数据表。下面就以这些数据表为例，介绍SQL语句的使用方法。

实践活动

在数据表中检索数据

在图3.3.3所示的两个关联的数据表中，可以尝试做哪些有意义的数据检索？

学生信息表id	学籍编号	校内班名	姓名	性别	籍贯	出生日期	政治面貌	专业代码
1	20180001	高一(02)班	褚晨怡	0	天津市	2002-01-14	无	010101
2	20180002	高一(03)班	方晨洁	0	北京市	2001-12-10	共青团员	040312
3	20180003	高一(03)班	顾建松	1	上海市	2002-08-24	共青团员	030409
4	20180004	高一(03)班	胡亚清	1	上海市	2002-05-27	无	030508
5	20180005	高一(02)班	黄湘湘	0	北京市	2002-07-08	无	020111
6	20180006	高一(05)班	刘天华	0	北京市	2001-10-06	无	030408
7	20180007	高一(02)班	陆鸿芳	0	浙江省	2002-08-16	无	040327

专业代码	专业名称
010101	哲学
010102	逻辑学
010103	宗教学
010104	伦理学
020101	经济学
020102	国际经济与贸易
020103	财政学
020104	金融学

图3.3.3 数据表

微信公众号：电子课本大全

3.3.2 数据库的查询方法

数据库中的基本数据查询方法有选择、投影、排序、统计等多种（图3.3.4）。

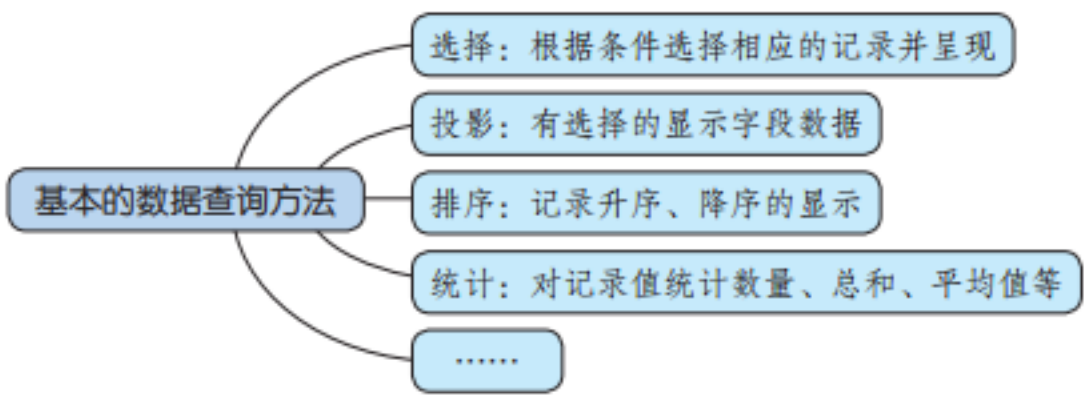


图3.3.4 数据查询的主要方法

选择法

选择法是数据查询最基本的方法，它利用select、from、where等关键字来实现数据的查询，其基本结构如下。

```
select <字段名1或表达式1>,<字段名2或表达式2>,… <字段名n或表达式n>
from <表1或子查询1>,<表2或子查询2>,…<表n或子查询n>
where <条件表达式>
      <其他关键字>;
```

select关键字是选择法查询的基本命令，后面可接不同字段名称；from关键字后面接的是数据表名或子查询。

例如，要显示学生信息表中所有的数据，可采用以下SQL语句。

```
select  *  from  学生信息表
```

在SQL中，通配符“*”代表所有字段，即显示学生信息表所有字段。

实践活动

查询数据表的数据以及了解通配符的含义

1. 使用选择法显示大学信息表中的所有数据。
2. 查阅资料，了解SQL中常用的通配符有哪些，各自对应的含义是什么。

where关键字后面的条件表达式可以筛选出与字段值相匹配的数据，用法为：

```
… where 字段名 = 字段值
```


微信公众号：电子课本大全

例如，在学生信息表中采用选择法查询出“政治面貌”为“共青团员”的数据，可采用以下SQL语句。

```
select * from 学生信息表 where 政治面貌='共青团员'
```

条件中的字段值带有字符时，可使用单引号（''）或双引号（""），查询结果如图3.3.5所示。

2	20180002	高一(03)班	方晨浩	0	2001-12-10	共青团员	北京市
3	20180003	高一(03)班	顾建松	1	2002-08-24	共青团员	上海市
9	20180009	高一(01)班	陆掌明	1	2002-01-10	共青团员	上海市
11	20180011	高一(01)班	潘艺泽	0	2002-08-06	共青团员	北京市
12	20180012	高一(05)班	钱昕怡	0	2002-01-06	共青团员	北京市
19	20180019	高一(01)班	沈伟	1	2001-02-18	共青团员	浙江省

图3.3.5　查询结果

实践活动

SQL支持的运算符

SQL支持算术运算、关系运算、逻辑运算，从而实现更加复杂的查询功能（表3.3.1～表3.3.3）。

表3.3.1　SQL中的算术运算

运算符	功能
+	加法
−	减法
*	乘法
/、div	实数除法、整数除法
%、mod	求余数

表3.3.2　SQL中的关系运算

运算符	功能
=	等于
<>、!=	不等于
<	小于
<=	小于或等于
>	大于
>=	大于或等于

表3.3.3　SQL中的逻辑运算

运算符	功能
and 、&&	与
or 、||	或
not、！	非

where关键字可以结合各种运算实现相应的查询。如果要“查询学生成绩表中语文成绩85分以上，数学成绩95分以上的学生姓名”，请写出这个语句。

微信公众号：电子课本大全

思考活动

如何根据需要显示指定数据

通过上述的学习，小明发现每次显示的数据都是所有字段的数据。但当数据量比较大时，想从显示的数据中快速找到所需要的数据有一定的难度。是否能够将小明所需要的字段显示，而不需要的字段数据不显示呢？

投影法

投影法是指在进行查询的过程中选择部分的字段数据显示，或格式化字段显示，从而精简查询结果，提高查询效率。

要实现部分字段的显示，只要在select关键字的后面添加相应的字段名称，不同字段之间用逗号“,”隔开。

例如，要显示学生信息表中“校内班名”和“姓名”字段的记录，可采用以下SQL语句。

```
select 校内班名,姓名 from 学生信息表
```

字段显示内容格式化是指通过使用as关键字，使得字段名称显示为指定的名称。as关键字用法如下。

```
字段名称 as 新名称
```

例如，在上述查询中，字段名称为“校内班名”，想让该字段显示为“班级”，使其更加精简，可以按以下格式写SQL语句。

```
select 校内班级 as 班级,姓名 from 学生信息表
```

图3.3.6显示的是在Navicat for MySQL中查询的结果。



图3.3.6　校内班名、姓名查询结果

实践活动

用投影法查询图书信息

打开前面创建的图书管理数据库，采用投影法查询图书的信息，并写出所使用的语句。

微信公众号：电子课本大全

阅读拓展

多表联查

一个数据库中往往有多个相互关联的数据表，有时一个查询中，需要显示多个相互关联表的数据。例如“大学信息”与“专业信息”是多对多的关系，可以通过“大学专业表”将两个表关联起来（图3.3.7）。

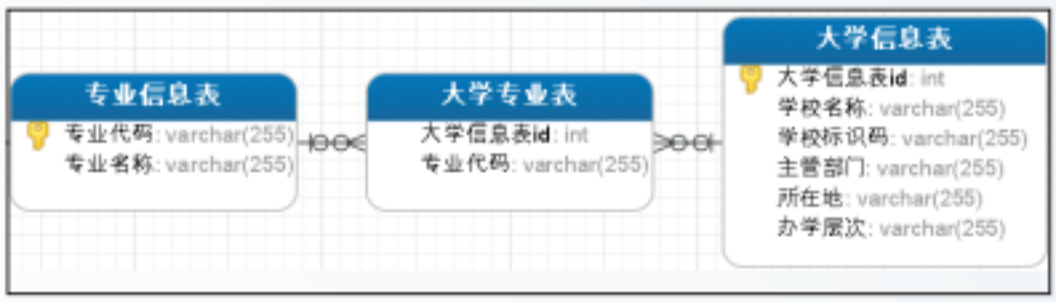


图3.3.7　表的关联

关联的规则是：“大学信息表”中的“大学信息表id”字段等于“大学专业表”中的“大学信息表id”字段；“专业信息表”中的“专业代码”字段等于“大学专业表”中的“专业代码”字段。在SQL语句中，表与所属字段用“.”分开，所有关联语句可以写成：

```
大学信息表.大学信息表id = 大学专业表.大学信息表id and
专业信息表.专业代码 = 大学专业表.专业代码
```

我们通过SQL语句来检索大学与专业的部分信息，可输入：

```
select * from 大学信息表,专业信息表,大学专业表  where 大学信息表.大学信息表id=大学
专业表.大学信息表id  AND
专业信息表.专业代码 =大学专业表.专业代码
```

多表联查时，不同数据表的名称放在关键字from后面，用逗号分开；同时表与表之间通过关联字段进行匹配，放在关键字where后面。

排序法

在文字处理软件和电子表格软件中，对数据进行排序是很常见的操作。使用SQL也可以实现类似的功能。

思考活动

如何用SQL语句来排序

小明的姐姐在分析和研讨高考志愿填报系统时，发现有一个叫“位次号”的项目。“位次号”是指对某一个专业来说，在录取前首先对考生按总分从高到低排名。在总分相同的情况下，再按照语文成绩从高到低排名……最后形成的序号就是“位次号”。那么，如何采用SQL语句来实现这个排序呢？

微信公众号：电子课本大全

排序法是指使用order by关键字，对数据表中指定字段值根据升序或降序进行排序。order by关键字一般写在SQL语句的最后，语法如下。

```
… order by 字段1 ASC/DESC,字段2 ASC/DESC,…
```

其中关键字ASC表示升序，DESC表示降序。

例如，使用SQL实现对“学生信息表”按字段名为“校内班名”进行升序排序。这里的SQL语句可按以下格式编写。

```
select * from 学生信息表 order by 校内班名 ASC
```

在使用order by关键字排序中，默认字段排序顺序为升序，所以上述语句中的ASC实际上可以省略。

图3.3.8显示的是在Navicat for MySQL中查询的结果。

```
1  select * from 学生信息表 order by 校内班名 ASC
2
```

息 结果1 概况 状态

学生信息表id	学籍编号	校内班名	姓名	性别	籍贯	出生日期	政治面貌	专业代码
31	20180031	高一(01)班	周乐洋	0	山东省	2002-04-29	无	010104
24	20180024	高一(01)班	杨海霞	0	江苏省	2002-02-09	无	040324
9	20180009	高一(01)班	陆掌明	1	天津市	2002-01-10	共青团员	020119
11	20180011	高一(01)班	潘艺萍	0	北京市	2002-08-06	共青团员	040308
21	20180021	高一(01)班	屈建清	1	上海市	2002-07-05	无	040110
19	20180019	高一(01)班	沈帅	1	北京市	2001-02-18	共青团员	040203
18	20180018	高一(01)班	沈诗芸	0	浙江省	2001-10-06	共青团员	040309
1	20180001	高一(02)班	褚晨怡	0	天津市	2002-01-14	无	010101
30	20180030	高一(02)班	赵民华	1	上海市	2002-03-14	无	010102
5	20180005	高一(02)班	黄淮湘	0	北京市	2002-07-08	无	020111

图3.3.8　查询结果

实践活动

用SQL语句排序

打开名为“学生成绩”的数据库文件，使用SQL语句实现从“学生信息表”和“学生成绩表”中显示学生的校内班名、姓名、语文、数学、外语、总分字段数据，并且根据“总分”从高到低排序。

统计法

统计法是指利用SQL对数据求和、计数、求平均值以及求最大值和最小值等，从而实现对数据的统计。SQL中的统计功能一般采用函数来实现。表3.3.4列举了SQL中常见的统计函数。

表3.3.4　SQL中常见的统计函数

函数名称	统计功能
sum(字段)	对字段值求和
count(字段)	统计记录的数量
avg(字段)	对字段值求平均值
max(字段)	对字段值求最大值
min(字段)	对字段值求最小值

微信公众号：电子课本大全

例如，使用SQL的统计功能，统计大学信息表中记录的数量，SQL语句可以写成以下格式。

```
select count(学校名称)  from 大学信息表
```

图3.3.9显示的是在Navicat for MySQL中查询的结果。



图3.3.9　查询结果

3.3.3　查询数据的提取

用SQL在数据库中查询的结果，假如想在其他数据分析软件里对它们进行分析，就需要把这个查询结果提取出来。

例如，要在Excel中提取MySQL的数据，可按以下方法操作。

（1）进入Excel界面，在“数据”页面单击“新建查询”，选择“从MySQL数据库（M）”选项。

（2）连接MySQL数据库。填写MySQL数据库服务器地址和数据库名称（图3.3.10仅供参考）及数据库用户名和密码（图3.3.11），完成和MySQL数据库的连接。



图3.3.10　连接数据库



图3.3.11　输入用户名和密码

阅读拓展

如何填写MySQL服务器地址

在填写MySQL服务器地址时，如果是本机作为MySQL服务器，则服务器名称可填写“localhost”或“127.0.0.1”，如果是由网络中其他的服务器作为MySQL服务器，则填写该服务器的IP地址或服务器名称。例如，在学生机房中，如果MySQL服务器的IP地址是“192.168.119.130”，则在服务器地址中填写该地址。

数据库的用户名和密码一般由数据库管理员提供，在操作过程中可咨询老师。

微信公众号：电子课本大全

完成登录后，窗口中显示了当前数据库中的所有数据表，选择合适的数据表，单击“加载”按钮，数据就被提取出来了。

如果只想提取SQL查询的结果，则可以在创建连接的界面中选择“高级选项”，然后输入SQL语句（图3.3.12），数据选择界面将显示SQL查询结果，最后加载相应的数据。



图3.3.12　输入SQL语句

实践活动

把用电量统计结果存为Excel表格文件或CSV文件

打开SQL语句编写代码界面，把名为“用电量”的数据库打开，统计2018年几个重点城市的每个月居民用电量、工业用电量的平均值，并且将结果提取到Excel或CSV文件里，为以后的数据分析做准备。

3.3.4　编程实现SQL查询

SQL可以嵌入其他程序开发语言中，用于数据的提取和分析。下面就用Python来实现SQL语句的调用和数据的提取。

实践活动

Python与MySQL连接

首先，进入CMD命令提示行状态，输入以下语句，安装PyMySQL（该库用于连接Python与MySQL）。

```
pip3 install PyMySQL
```

然后，打开Python的编程窗口，新建Python文件，输入下面的语句就可以引入PyMySQL库。

```
import pymysql
```

打开Python的编程窗口，新建Python文件，引入PyMySQL库，并执行PyMySQL.Connect命令创建一个MySQL连接。连接命令参数如表3.3.5所示。

例如，前面已经建立了“学生专业规划”数据库。下面就通过编程命令，实现调用这个数据库和提取数据的功能。

微信公众号：电子课本大全

表3.3.5　连接参数

参数名	类型	说明
host	字符串	MySQL服务器地址
port	数字	MySQL服务器端口号
user	字符串	用户名
passwd	字符串	密码
db	字符串	数据库名称
charset	字符串	字符编码

数据库连接创建命令如下。

```
db=pymysql.connect(host ='localhost', port = 3306,
user = 'root', passwd = '123456', db = '学生专业规划' ,
charset='utf8')
```

由于数据库中的数据含有中文，为了防止乱码，建议在数据库、数据表及数据库连接中采用统一的字符集，例如UTF-8。

完整的程序代码如下。

```
import pymysql
db=pymysql.connect(host='localhost',port=3306,user='root', passwd= '123456',
db= '学生专业规划', charset='utf8')
cursor = db.cursor()                                    #创建游标
sql='select 姓名,校内班名 from 学生信息表'                  #编写和执行SQL语句
cursor.execute(sql)                                     #执行SQL语句
results=cursor.fetchall()                               #得到数据集
for row in results:                                     #迭代显示
    xm=row[0]
    bj=row[1]
    print('姓名：',xm,' 班级：',bj)
cursor.close();                                         #释放游标
db.close();                                             #关闭数据库连接
```

```
姓名： 褚晨怡  班级： 高一(02)班
姓名： 方晨洁  班级： 高一(03)班
姓名： 顾建松  班级： 高一(03)班
姓名： 胡亚清  班级： 高一(03)班
姓名： 黄渝湘  班级： 高一(02)班
>>>
```

图3.3.13　运行结果

执行代码后，运行结果就会显示出来（图3.3.13）。

实践活动

提取图书信息并显示

利用Python语言编程，调用前面创建的图书管理数据库，并提取里面的数据，把图书名称、书号、出版社、作者等信息显示出来。

微信公众号：电子课本大全

项目实施

用SQL实现对“学生专业规划”数据库的查询

一、项目活动

1. 利用网络爬虫，获取更多的大学和专业数据，并把它们导入本项目数据库中对应的数据表里。

2. 运用SQL语句，实现以下操作。

（1）运用多表联查，获取指定大学所设立的专业。

（2）指定一个专业，获取可报考的大学。

（3）指定一个专业，获取所有对该专业有选择意向的同学名单。

3. 小组研讨：还能给该数据库增加哪些功能？能否给它添加一个友好的查询界面，让不会编程的同学也能查询里面的数据？尝试使用Python，设计一个可以便利查询该项目数据库的小系统。

二、项目检查

1. 组内成员两两合作，对数据库进行查询操作，一个同学提出查询要求，另一个同学上机操作。

2. 每个同学针对下面几个问题进行自我检测。

（1）对数据库在数据管理作用方面有哪些新的体会？

（2）现阶段，对所学内容是否感到困难？主要困难是什么？将如何克服？

练习提升

1. 打开文件名为“学生专业规划”的数据库文件，执行以下操作。

（1）结合关系运算和逻辑运算，获取“校内班名”为“高一(01)班”并且“籍贯”为“浙江省”的所有数据。

（2）在学生信息表中只显示“姓名”和“性别”字段的记录。

（3）在学生信息表和学生成绩表中显示姓名、语文、数学和外语字段，并且按“语文”为主要关键字降序排序，“数学”为次要关键字降序排序。

（4）统计指定大学专业的数量。

2. 查阅SQL的相关材料（网络资料或图书），掌握更多SQL语句的使用方法。

微信公众号：电子课本大全

3.4 备份和还原数据库

学习目标 ▶▶▶

- 了解数据备份的重要性，知道数据库备份的含义。
- 学会利用数据库备份功能实现对数据的备份，确保数据的安全和完整。
- 学会根据备份数据实现对数据库的还原，确保在数据库发生意外时能够及时恢复。

体验探索

数据丢失的遭遇

【案例1】李满有一天独自外出，他和父母约定一个小时后给家里打个电话。但过了约定时间，父母等不到他的电话，也打不通他的电话。虽然最后联系上了，但也让父母颇为担心。事后通过新闻才知道，一个通信技术公司由于误操作，导致几十万用户数据丢失，用户的手机通信业务也受到了影响。

【案例2】小明喜欢摄影，外出旅游时拍了几千幅照片，但是回家后忘了及时导出来保存。过几天发现，相机存储卡出现了故障，照片无法读取，他痛心无比。

【案例3】王博在单位负责人事数据库的更新维护。有一天由于他的误操作，把员工薪资数据删除了，导致当月无法按时发放工资。他也因此受到相应的处罚。

讨论：你有数据丢失的遭遇吗？或者你听说过哪些数据丢失的事件？数据丢失会带来什么样的后果？

在数据管理中，数据安全防范意识的培养尤为重要。例如，我们的手机、计算机或移动硬盘中都存储着很多重要的数据，如通信录、照片、视频、文档等。你是否想过：手机突然坏了或丢了怎么办？存储在里面的数据还能完好无损地找回来吗？该怎么做才会尽可能地找回更多更完整的数据？对于隐私数据，又该如何设置数据保护方式呢？增强数据保护意识，掌握备份方法是根本。

微信公众号：电子课本大全

3.4.1 数据丢失常见的原因

数据丢失常见的原因有以下几种。

数据保护意识不强

如果平时缺乏足够的数据保护意识，就有可能因为意外导致数据丢失。尤其是一些重要资料，如果没有及时备份，一旦出现问题，都可能造成无法挽回的损失。“体验探索”案例2中的小明就遇到这种情况。

设备故障

硬盘和系统在使用过程中偶尔会出现一些意外，软件也随时有崩溃的风险。这些意外都有可能导致数据丢失。

人为误操作

在日常工作中，人为的误操作会导致数据安全受到很大的影响，例如，一些误格式化、误分区、误删除等操作引发的文件丢失。在“体验探索”案例3的情况中，就属于人为误操作造成的数据丢失。

总的来说，在数据丢失事件中，关键还是人的意识。只要我们具备及时备份、多次备份的意识，并按照正常的操作流程来完成日常工作，就能把数据丢失的损失程度降到最低。

思考活动

如何避免小概率错误导致的数据丢失

在“云”上备份数据已成为很多个人和组织备份数据的一种途径（图3.4.1）。但一家提供云服务的知名公司，却把一个大客户放在其云服务器上的数据全部丢失了，且无法恢复。据该公司披露的信息显示，该故障起源于因磁盘静默错误导致的单副本数据错误，再加上数据迁移过程中的两次不规范操作，导致云盘三个副本安全机制失效，最终导致客户数据的完整性受损。



图3.4.1 “云”上的数据库

据分析，这次事故是因为运维人员为了尽快完成搬迁任务、降低仓库使用率，违规关闭数据校验，违规对源仓库进行数据回收而引起的。

思考：事件的发生，归根到底是人为的因素还是技术的因素？在实际工作中，对于非常重要的数据，如何避免因为小概率错误而导致的数据丢失？

微信公众号：电子课本大全

3.4.2 常见的备份方法

任何原因导致的数据丢失或损坏都将带来不可弥补和无法估量的损失。在所有保护数据安全的战略中，数据备份是最基础的工作。数据备份，就是把数据从原来存储的地方复制到其他地方的操作，其目的就是在设备发生故障或发生其他威胁数据安全的灾害时保护数据，将数据遭受破坏的程度降到最低。

例如，小明和几个同学合作，为学校食堂建立了一个数据库，用来管理食材、食谱、学生等数据，从而帮助食堂提高管理水平。当数据库里存储了大量的数据时，无法分清楚哪些数据重要、哪些不重要，为了防止数据丢失，他们应该如何选择备份方法？

下面就来介绍几种常用的备份方法：全备份、增量备份、差异备份、实时备份和定时备份（表3.4.1）。在实际应用中，要根据具体情况合理选择恰当的方法。

表3.4.1 数据备份的方法

方法	简单定义	主要优点	主要缺点
全备份	对某一时间点的系统所有数据进行备份	备份的数据全面完整，只需利用一份副本就可以恢复全部数据	备份的数据量大，消耗的存储空间多，备份过程慢
增量备份	对新增加和修改的文件进行备份（与前一次备份比较）	与全备份结合，可以在使用较少存储空间的同时能够对数据进行全部备份	备份一环扣一环，若中间一个备份数据丢失，将导致还原数据库时失败或数据不完整
差异备份	对上一次全备份之后有变化的数据进行备份	只需对第一次全备份和最后一次差异备份进行数据还原	每一次的差异备份是在上一次的差异备份数据上进行累加备份，备份文件的容量一般也是逐渐增加
实时备份	利用主数据库服务器和从数据库服务器，通过同步日志事件备份数据	若主服务器出故障时，从服务器可以替代主服务器，减少还原数据的时间，提高了效率	实时备份需要2台或多台服务器，相比较其他的备份方式，成本较高
定时备份	固定的时间间隔进行数据备份	一定程度上减轻了管理员的工作量	规划时间点之外的其他突发事件不能自动备份，还需要人工结合其他备份方式

数据的备份是一个长期的过程，而恢复数据（数据还原）只在事故发生后进行。恢复可以被看作备份的逆过程，恢复程度的好坏很大程度依赖于备份的完整性。因此，我们应该重视数据的备份，确保数据的安全。

微信公众号：电子课本大全

实践活动

巩固数据备份的知识与操作

1. 与同学一起讨论，分析各种数据备份方法的特点。
2. 了解各种备份方法对应的常用备份工具有哪些。
3. 为小明的项目团队制订一个数据备份的方案。

3.4.3 备份与还原数据库

数据库的备份是数据备份的一个重要方面。在数据库备份中，数据库管理系统（如MySQL、SQL Server等）不同，备份的命令和步骤会有一定的差别。接下来介绍用MySQL实现数据库备份的方法。

阅读拓展

MySQL数据库备份前的准备

BinLog是MySQL实时记录数据变化的日志文件。在MySQL中，要实现备份和还原，需要开启BinLog，开启方法如下。

1. 找到MySQL的配置文件my.ini（默认在“C:\ProgramData\MySQL\MySQL Server 5.7”）。

2. 打开my.ini文件，在[mysqld]小节中找到“# Binary Logging”，将下面的“#log-bin”设置为“log-bin = mysql-bin”。

3. 重新启动MySQL。

备份策略的选择

不同的数据库备份类型各有优缺点。针对较小的数据库，可以采取每天全备份的策略；针对较大的数据库，可以采用定期“全备份＋日志增量备份”或差异备份的策略；对于实时性要求比较高的数据库应用（如银行领域），可以采用实时备份。

例如，对本章“学生专业规划”数据库进行备份，可以采用“全备份＋日志增量备份”的方式。

首先制订未来一周的备份计划（表3.4.2）。

表3.4.2 数据库备份计划

周一18：00	周二18：00	周三18：00	……
全备份，重置日志文件	增量备份1	增量备份2	……

微信公众号：电子课本大全

全备份

在备份的过程中，为了防止对数据库的写入或修改，首先要锁定数据库，命令语句如下。

```
mysql -uroot -p 密码                    #进入MySQL控制台
flush tables with read lock;            #输入锁定命令
```

在MySQL中对数据库进行全备份，可以采用mysqldump命令来实现。该命令语句的具体格式如下。

```
mysqldump [参数] 数据库1[表格] [数据库2…]  > 保存位置
```

例如，要将“学生专业规划”数据库备份到计算机中的D盘，名称为stu.sql，则可以写以下命令语句。

```
mysqldump --default-character-set=gb2312  -u root -p 学生专业规划> d:\stu.sql
```

由于这里的数据库名称是中文，所以使用“--default-character-set=gb2312”参数，确保命令能被正确执行。

通过该操作，D盘会自动生成全备份文件stu.sql。

当完成一次全备份后，剩下的几天里只需要进行增量备份。增量备份的数据记录在日志文件bin-log中，所以当完成全备份操作后，应该立即将当前的日志文件存盘（如文件名称为mysql-bin.000001），同时产生新的日志文件（如文件名称为mysql-bin.000002）。那么，接下来对数据库进行的任何操作，都将记录在日志文件mysql-bin.000002中，而进行的第一份增量备份文件，正是这个新产生的mysql-bin.000002文件。

对日志文件进行存盘并产生新日志文件的操作命令语句如下。

```
flush logs;
```

完成日志切换后，也要记得将数据库解锁，命令语句如下。

```
unlock tables;
```



阅读拓展

日志文件的位置

一般日志文件与数据库在同一个位置，若安装数据库时设置为默认安装，则数据库和日志文件一般都在“C:\ProgramData\MySQL”文件夹中。

增量备份

根据计划表，第二天进行的增量备份，其实就是第一天对日志文

微信公众号：电子课本大全

件的操作，即把全备份后一天的日志文件（如mysql-bin.000002）存盘，并且产生新的日志文件（如mysql-bin.000003），用于记录下一天的数据库操作过程。

第二天的增量备份，首先是锁定数据库，命令语句如下。

```
flush tables with read lock;
```

将日志文件存盘并产生新文件，命令语句如下。

```
flush logs;
```

解锁数据库，命令语句如下。

```
unlock tables;
```

最后将日志文件mysql-bin.000002保存起来。后面几天的增量备份与该操作类似，这里不再重复叙述。

实践活动

备份图书管理数据库中的数据

打开图书管理数据库的文件，对它进行一次全备份和两次增量备份，熟悉备份的流程和命令语句。

还原数据库

备份数据库是为了确保数据的安全，以备在数据库发生故障时可以及时恢复。

思考活动

如何对数据进行还原

赵军同学按照表3.4.2的计划，对数据库进行了全备份和增量备份。由于服务器异常，导致数据库在周三晚上21点出现了故障，现在要还原数据库，他该怎么做呢？

根据数据备份的顺序，在数据库发生故障前，存在以下几个备份文件（表3.4.3）。

表3.4.3　备份后产生的备份文件

周一18：00	周二18：00	周三18：00	周三21:00
全备份文件stu.sql	增量备份文件1 mysql-bin.000002	增量备份文件2 mysql-bin.000003	出现故障，原数据库无法读取，18:00-21:00的日志文件：mysql-bin.000004

全部备份文件放在新数据库服务器的D盘中，接下来按照备份的顺序进行逐个还原。

微信公众号：电子课本大全

（1）全备份的还原

首先要新建数据库。例如，要新建“学生专业规划”数据库，用于后面还原过程中将数据还原到该数据库。命令语句如下。

```
mysql -uroot -p 密码                          #进入mysql控制台
mysql> create database 学生专业规划;    #创建新数据库
```

这个时候创建的新数据库是空的数据库。开始还原全备份的数据库文件stu.sql，命令语句如下。

```
mysql -uroot -p 密码 学生专业规划<d:\stu.sql  #开始还原
```

至此，全备份文件已经还原成功。但是该还原数据是到周一18:00时的数据库状态数据，所以还要通过以下的增量备份数据还原恢复到故障时的数据状态。

（2）增量备份的还原

增量备份的还原用mysqlbinlog命令，命令语句如下。

```
mysqlbinlog 日志文件位置 | mysql管理员登录权限
```

例如，要还原第一个增量备份文件mysql-bin.000002，还原命令语句如下。

```
mysqlbinlog  d:\mysql-bin.000002 | mysql -uroot -p 密码
```

通过本次增量备份的还原，使得数据库还原到周二18:00的状态。同样，对mysql-bin.000003和mysql-bin.000004增量备份文件的还原，使得数据库恢复到故障时的状态，从而完全还原数据库。

实践活动

还原已备份的图书管理数据库

我们已经对图书管理数据库进行了全备份和增量备份。现在需要将已备份的数据文件还原到同一局域网的另一台计算机上，请完成相应的操作。

阅读拓展

mysqldump命令

mysqldump命令可以将数据库中的数据备份成一个文本文件。表的结构和表中的数据将存储在生成的文本文件中。mysqldump命令的工作原理是：先查出需要备份的表的结构，再在文本文件中生成一条create语句；将表中的所有记录转换成一条insert语句；然后通过这些语句就能够建表并插入数据。

提示：可以查阅mysqldump的用法实例，了解用该命令进行数据备份的方法。

微信公众号：电子课本大全

项目实施

对“学生专业规划”数据库进行备份

一、项目活动

1. 制订数据库备份的周策略，并填写计划表。

2. 尝试数据还原的操作。

3. 结合本章的学习研讨以下问题：数据管理与数据库的关系；大数据时代数据管理面临的挑战；数据备份在数据管理中的重要意义。

二、项目检查

汇总本章项目学习各阶段的成果，在老师的指导下，按调研论文的格式进行编辑加工和排版，并在班里进行交流和展示。

论文要包含以下几方面内容。

- 项目开展的背景情况。
- 小组成员分工情况和项目工作计划。
- 获取的数据（可用数据表格展示）。
- 建立数据库的成果（展示E-R图，写出数据库名称和各个数据表名称）。
- 数据备份的策略。

练习提升

1. 数据丢失会带来什么风险？请结合实例进行说明。

2. 数据丢失的主要原因有哪些？谈谈人们主观意识对数据保护的重要性。

3. 数据备份包括全备份、_____、_____、实时备份、定时备份。

4. 备份数据库时需要对数据库进行锁定，防止写入，锁定数据库的命令语句为：

_____

5. 需要对一个名为student的数据库进行一次全备份，备份文件为student.sql，并且生成存到D盘根目录，则全备份的命令语句为：

_____

6. 在进行增量备份时，需要对日志文件存盘并生成新的日志文件，则该命令语句为：

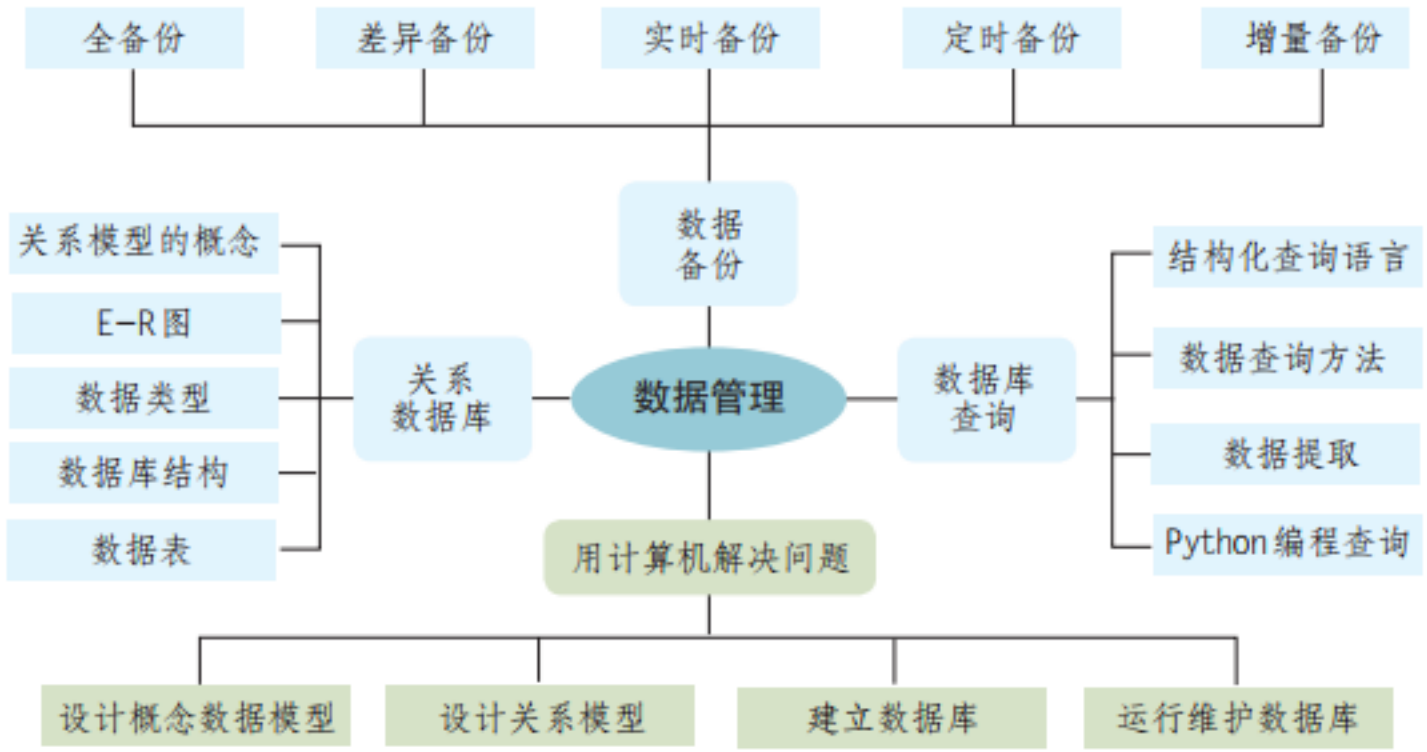
_____

7. 小东的D盘中包含一个名为student.sql全备份数据和一个名为mysql-bin000003的增量备份，请写出相关命令语句，将该全备份和增量备份进行恢复。

微信公众号：电子课本大全

总结评价

1. 下图展示了本章的核心概念与关键能力，请同学们对照图中的内容进行总结。



2. 根据自己的掌握情况填写下表。

学习内容	掌握程度		
数据库与数据库管理系统	□ 不了解	□ 了解	□ 理解
建立数据的概念模型	□ 不会	□ 会	□ 熟练
把概念模型转换为关系模型	□ 不会	□ 会	□ 熟练
创建数据库和数据表	□ 不会	□ 会	□ 熟练
数据录入或导入	□ 不了解	□ 了解	□ 理解
数据查询	□ 不会	□ 会	□ 熟练
数据提取	□ 不会	□ 会	□ 熟练
常用的SQL语句	□ 不了解	□ 了解	□ 理解
备份方法和备份数据库	□ 不会	□ 会	□ 熟练

3. 回答以下几个问题，对自己的学习情况进行总结与反思。

(1)在本章的项目学习中，你在小组合作方面有什么体会?

(2)项目实施中，涉及思考、调研、实践、评价等环节，你表现较好的是哪个环节?

(3)通过学习数据库的相关知识，你是否对数据管理有了更深刻的认识?

(4)你对目前的学习感到最困难的地方是什么?

微信公众号：电子课本大全

第 4 章
数据分析

本章的目的是了解数据分析的内涵。通过学习，我们能够选择恰当的数据分析工具，使用科学有效的数据分析方法对数据进行分类整理；能够转换数据所承载的信息，提取与发现其中有价值的信息；学会选择适合的展示方式，将数据分析结果进行可视化展示；能够撰写完整的数据分析报告，从而体验数据分析的完整过程。



微信公众号：电子课本大全

4 主题学习项目：数据分析知天气

项目目标

本项目首先要采集和处理某地区一段时期内的天气数据（含空气质量指数），然后对这些数据进行分析，从大量看似杂乱无章的数据中发现内在规律和变化趋势，从而挖掘数据背后的现象和一些有价值的事实依据。

1. 根据要分析的问题，采用恰当的分析方法对数据进行分析，增强数据可视化的能力。

2. 撰写天气数据的分析报告，阐述获得的事实并发表自己的观点，进一步提高数据分析能力和自我表达能力。

项目准备

为了完成项目，需要做以下准备。

- 组建学习小组，建议2 ~ 3人一组为宜。
- 采集数据。利用网络爬虫、搜索引擎等工具搜索某地区近几年的天气数据。
- 回顾和熟悉Python编程。熟悉软件编程的界面，自主学习常用的编程语句。
- 学会使用数据库系统管理搜索到的天气数据。

项目过程

为了保证顺利完成本项目的学习活动，在不同学习阶段，小组长要注意检查组员项目学习的进度，并做好协调互助工作。

1 选择工具和方法

把前期获取的天气预报数据和空气质量指数导入数据库中进行有效管理，确定要分析的问题，选择合适的数据分析工具和分析方法。

P120

2 分析数据

采用恰当的方法对数据进行分析，并把数据可视化，获得相应的图表。

P130

3 汇总与表达

小组成员一起讨论，确定数据分析报告的格式，把过程中的成果进行汇总，形成完整的报告并在一定的场合展示。

P130

项目总结

完成本章项目学习后，各小组要提交关于某地区天气数据的分析报告，并与他人交流分享学习体会。通过本项目的学习，增强团队合作意识，增强对数据进行分析的能力和获取有价值观点的能力。

微信公众号：电子课本大全

4.1 数据分析的工具与方法

学习目标 ▶▶▶

- 进一步了解数据分析的基本步骤，增强数据分析的过程性和有序性。
- 了解常用数据分析方法的内涵及特点，并能用于数据分析的实践中。
- 能够使用Python语言编写数据分析程序。
- 理解数据分析的意义，增强用数据阐述自己观点的意识与能力。

体验探索

数据分析带来商业价值

一家大型零售商的销售产品总数超过3万种，产品的价格因地区和市场条件而异。由于产品种类繁多，成本变化比较频繁，每年商品调价次数高达12万次，给定价促销策略的制定带来非常大的困难。因此，公司组建了数据分析团队，旨在通过分析消费者的购买记录和相关信息（图4.1.1），提高定价的准确度和响应速度。通过这一系列的活动，提高零售商的销售额和利润。

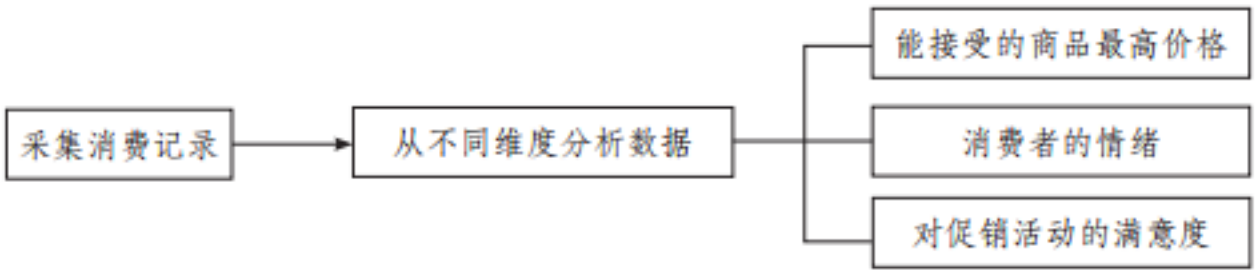


图4.1.1 分析的内容

思考：如果你是本案例中的一名数据分析师，可以从消费记录中获得哪些方面的信息？同时通过哪些数据能预测消费者的消费倾向？

微信公众号：电子课本大全

数据的价值已在商业、科研等领域得到体现。但数据的价值不是直接表现出来的，而是需要人们通过恰当的方法、工具和技术去分析和挖掘。为了获取数据的价值，就要掌握数据分析的工具和方法。同时，也要了解数据挖掘的概念和作用，为从数据中获得信息、知识和智慧打下基础。

4.1.1 数据分析的工具

工欲善其事，必先利其器。想要获得更多的数据价值，就需要借助恰当的数字化工具。简单的数据分析工具有电子表格类软件，功能丰富的数据分析工具有SPSS、SAS。另外，还可以通过编程实现数据分析，如Python语言中的Numpy、Pandas、Matplotlib等一些专业数据分析库。Numpy是Python的一种开源数值计算扩展库；Pandas提供了大量能快捷处理数据的函数和方法；Matplotlib是一个Python的2D绘图库，可以生成直方图、功率谱、条形图和散点图等，是Python中最常用的可视化工具之一。除此之外，还有Xlrd、Xlwt、Pymysql等读取文件、连接数据库的扩展工具。这些数据分析工具要根据分析任务的需要来选用。

实践活动

认识Python扩展库

安装Python中与数据分析相关的扩展库，进一步熟悉它的操作界面和常用语句的语法规则。

4.1.2 常用的数据分析方法

实际工作中，数据分析的任务往往是要求从已有数据中找出规律和背后的现象。这时就需要根据要求选择恰当的分析方法。常用的数据分析方法有对比分析法、分组分析法、交叉分析法、平均分析法和相关分析法。

对比分析法

对比分析法也称为比较分析法，通常是把两个相互关联的客观事物或指标数据加以比较，研究它们的特点（如规模大小、速度快慢、水平高低）以及关系，分析它们背后的现象。

微信公众号：电子课本大全

思考活动

店铺客流量的对比分析

某品牌公司需要找店商代理产品，现在公司需要在两家店铺中挑选一家。如果交给你这个任务：分析哪一家店铺人气旺。你是不是可以采用对比分析法，对某一段时间范围内，这两家店铺的客流量进行对比分析呢？

例如，想对比某两家店铺顾客人数的变化情况，首先选用这两家店铺同一天的单位时间顾客统计数据，并用Pandas库的DataFrame模块创建一个名为“顾客数”的数据帧，再从Matplotlib库中引入Pyplot模块创建两家店铺客流量对比折线图，从而能直观地看出二者的区别。

利用Python可以编写下面这段程序。

```
import pandas as pd
from matplotlib import pyplot as plt
#调用中文字体'SimHei'（黑体）
plt.rcParams['font.sans-serif']=['SimHei']
#创建图表x轴刻度组，命名为“时间轴”
时间轴='10时','11时','12时','13时','14时','15时','16时','17时'
x = range(len(时间轴))
#引用数据创建“顾客数”数据帧
顾客数=pd.DataFrame({'店铺一':[20,28,29,30,24,18,21,32],'店铺二':[21,28,30,33,
25,24,22,24]},index=时间轴)
#通过“顾客数”数据帧创建折线图
顾客数.plot()
#将“时间轴”导入折线图横轴
plt.xticks(x,时间轴)
#给图表横轴、纵轴以及表头标上标题
plt.xlabel('时间');plt.ylabel('顾客数量/人');plt.title('店铺客流量')
#显示图表
plt.show()
```

这个程序前两行语句的作用是导入Pandas和Matplotlib中的Pyplot。利用Pyplot里的函数，可以创建图形、在图形里创建绘图区、在绘图区画线、用标签装饰图形等。运行程序后，可以直观地对数据进行对比分析（图4.1.2）。

在对比分析中，选择合适的对比标准十分关键，标准合适才能做出客观的评价。

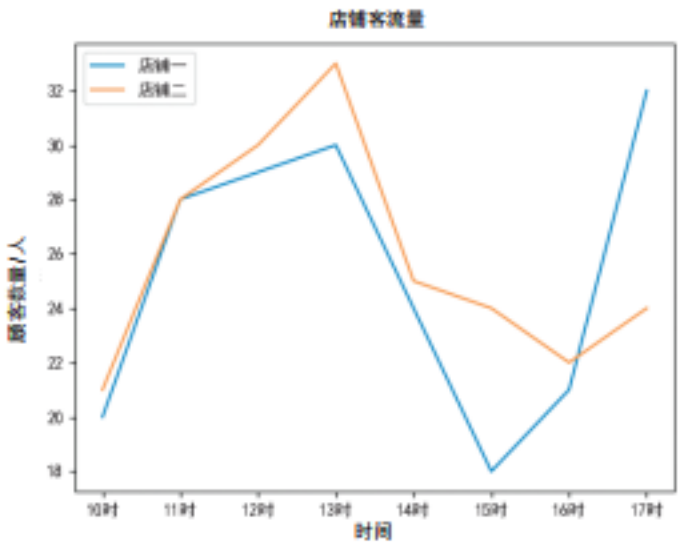


图4.1.2　对比分析图

微信公众号：电子课本大全

阅读拓展

对比分析法中的同比与环比

在一些关于统计数据的新闻报道中，经常能听到同比增长、环比下降等说法。例如，国家统计局的数据显示，“2018年7月，全国居民消费价格同比上涨2.1%，环比上涨0.3%”（图4.1.3）。其实这是对比分析法中的不同类型。按照发展速度采用基期的不同，对比分析可分为同比、环比和定基比分析，三者均用百分数和倍数表示。

2018年7月工业生产者价格主要数据

	环比涨跌幅（%）	同比涨跌幅（%）	1-7月平均同比涨跌幅（%）
一、工业生产者出厂价格	0.1		
生产资料	0.1		
采掘	0.6		
原材料	-0.1		
加工	0.1		
生活资料	0.2		
食品	0.2		
衣着	0.4		
一般日用品	-0.1		
耐用消费品	0.2		

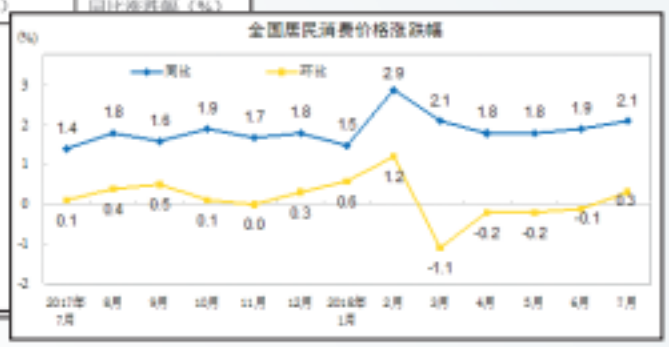


图4.1.3　来自国家统计局的数据

实践活动

对比分析空气质量变化的特点

打开名为“空气质量指数”的文件，根据其中的数据，采用对比分析法，对其中的数据进行同期比较，观察其变化特点。

分组分析法

分组分析法是指通过统计分组的计算和分析，来认识所要分析对象的不同特征、不同性质及相互关系的方法。分组分析法是在分组的基础上，从定性或定量的角度，对现象的内部结构或现象之间的依存关系进行分析研究，以便寻找事物发展的规律，正确地分析问题和解决问题。

数据分析前，首先要确定组数、确定组距以及将数据归纳进组。分组时要遵循穷尽原则和互斥原则。穷尽原则就是使总体中的每一个单位都有组可归；互斥原则就是在特定的分组标志下，总体中的任何一个单位只能归属于某一个组，而不能同时或可能归属于几个组。

微信公众号：电子课本大全

例如，某三甲医院抽选20位医生，通过分析他们某一天的接诊人数（表4.1.1），以分析接诊能力接近的医生之间是否存在差异，从而便于调整工作量。

表4.1.1　医生一天内接收的就诊人数

医生序号	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
人数（人）	66	49	54	52	61	63	62	55	48	42	47	55	48	39	33	60	48	53	46	52

首先，可用Excel软件将上述数据转置后做成表格，然后用Pandas库中的read_excel()函数（需要Xlrd支持）读取数据，利用Matplotlib库中的直方图将上述人数分段进行分组统计并显示。

Xlrd扩展库是Python语言中读取Excel的扩展工具。

编写的程序如下。

```
import pandas as pd
from matplotlib import pyplot as plt
#调用中文字体'SimHei'（黑体）
plt.rcParams['font.sans-serif']=['SimHei']
#从'就诊人数.xlsx'中导入数据并转换成数据帧
就诊人数=pd.DataFrame(pd.read_excel('D:\\就诊人数.xlsx'))
#将就诊人数由低到高的数据跨度等分成四个数据范围，并统计每个数据范围内的医生数量，制作
成直方图
就诊人数.hist(bins=4)
#给图表中的横轴、纵轴标上标题
plt.xlabel('就诊人数/人');plt.ylabel('医生人数/人')
plt.show()
```

上面的程序将就诊人数情况跨度由低到高均分为四组，并且自动统计各个区间内的样本数量，然后以直方图的形式显示出来（图4.1.4），可以直观地看到每组中的医生人数，还可以深入分析同组医生职称、学历等数据的规律等。

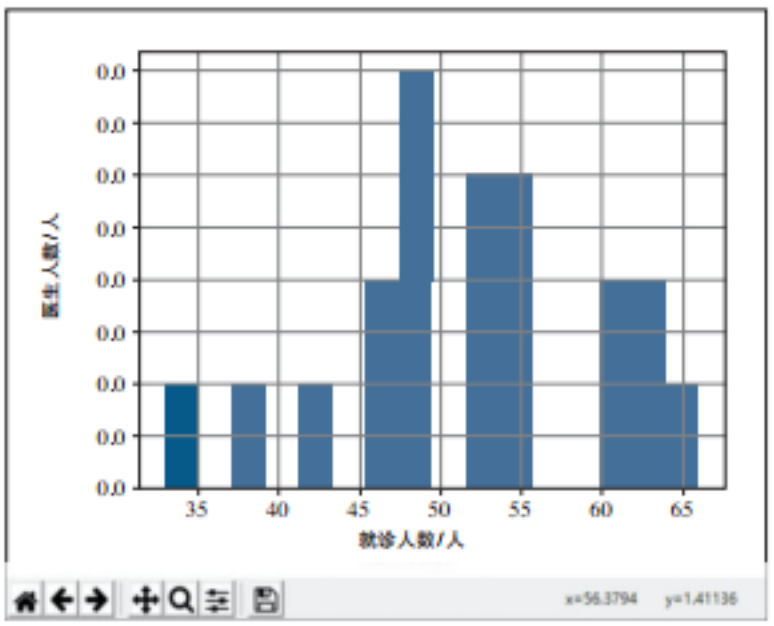


图4.1.4　直方图

利用分组分析法，可以在同一标准范围内对数据进行分析，客观性更强。例如，在对某个地区高中教育质量进行监测时，往往就会把监测数据按经济发展水平不同区域、学校、学生性别、年级、学科等进行分组，然后再进行统计分析，得出每种分组的数据变化规律。

微信公众号：电子课本大全

实践活动

分组分析法的使用

使用分组分析法时，分组是关键。请查阅相关资料，简单了解这种方法的分组原则和适用情况。然后对“学生专业规划”数据库中的“学生”数据按学籍、性别分组，然后进行统计，了解这两个组别的学生选择专业的规律。

交叉分析法

交叉分析法又称立体分析法，是从交叉、立体的角度出发，由浅入深、由低级到高级的一种分析方法。这种方法虽然复杂，但它弥补了“各自为政”分析方法所带来的偏差。交叉分析的主要作用就是从多个维度细分数据，从中发现最为相关的维度来探索数据变化的原因。

例如，某店铺出售一批国产和进口商品（表4.1.2），库存管理员按商品类别、产地、品种、数量和价格对每种商品进行了数据处理。

表4.1.2　商品信息

类别	水果				蔬菜			肉类		
产地	美国	中国	中国	新西兰	中国	新西兰	美国	新西兰	美国	中国
品种	苹果	梨	草莓	菲油果	番茄	黄瓜	胡萝卜	羊肉	牛肉	鸡肉
数量（份）	5	5	9	4	3	2	7	10	8	12
价格（元/份）	5	5	10	8	3	2	12	13	20	6

接下来用Python将上述数据处理成数据帧。

```
import pandas as pd
from matplotlib import pyplot as plt
#调用中文字体'SimHei'（黑体）
plt.rcParams['font.sans-serif']=['SimHei']
#根据上文数据创建一个名为“商品”的数据帧
商品 = pd.DataFrame({'类别':['水果','水果','水果','水果','蔬菜','蔬菜','蔬菜',
'肉类','肉类','肉类'],
                     '产地':['美国','中国','中国','新西兰','中国','新西兰','美国',
'新西兰','美国','中国'],
                     '品种':['苹果','梨','草莓','菲油果','番茄','黄瓜','胡萝卜','羊
肉','牛肉','鸡肉'],
                    '数量':[5,5,9,3,2,10,8,7,4,12],
                    '价格':[5,5,10,3,3,13,20,12,8,6]})
#将“商品”的“类别”作为列索引，将商品的“产地”作为行索引创建一个交叉数据表
产地类别 = pd.crosstab(产品['类别'],商品['产地'])
print(产地类别)
#将交叉数据表“产地类别”中的“肉类”数据制作成饼图，数据显示至百分比小数点后两位
plt.pie(产地类别.ix['肉类'],labels=产地类别.columns,autopct='%1.2f%%')
plt.title('肉类产地来源比')
plt.show()
```


微信公众号：电子课本大全

程序运行后输出的结果如图4.1.5和图4.1.6所示。

产地 类别	中国	新西兰	美国
水果	2	1	1
肉类	1	1	1
蔬菜	1	1	1

图4.1.5　输出的数据行

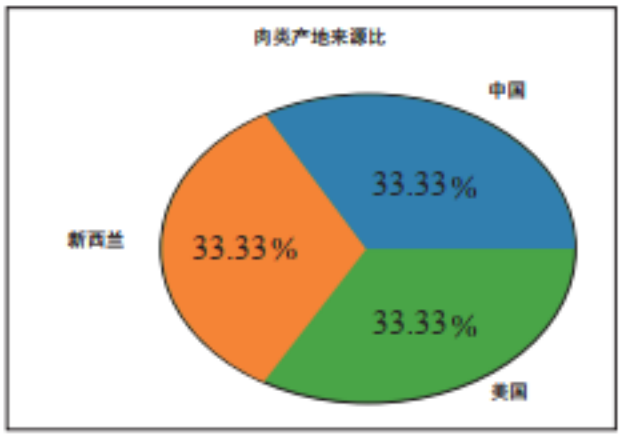


图4.1.6　饼图

阅读拓展

交叉分析法

利用交叉分析法，能对数据进行多角度的观察与操作。报表字段的定义可以通过拖动的方式实现，操作简便。用户还可以对字段进行一系列设置，例如，可对汇总维度字段进行预警、显示格式、统计方式等的设置；可对维度字段进行分组统计方式、超链接、排序等设置（图4.1.7），满足用户的多种设计需求。

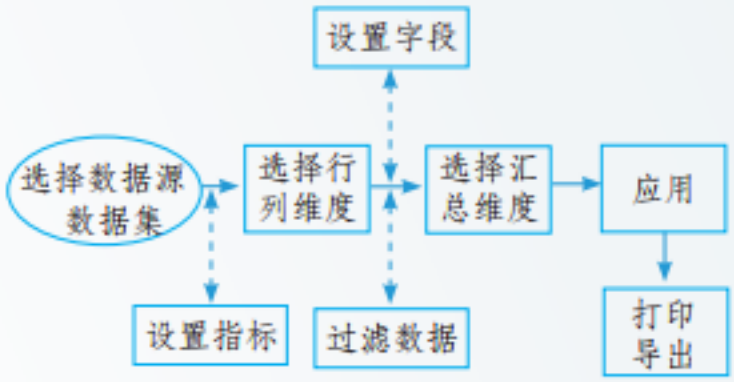


图4.1.7　交叉分析法

实践活动

制作产品相关数据的饼图

如果想要分析以上案例中所有肉类产品来自哪些产地，并且各有多少类别，需要利用crosstab()函数创建一个交叉数据表来仔细分析，并且可以用pie()函数来做成饼图直观分析。请尝试制作一个饼图（参考图4.1.6）。

平均分析法

平均分析法是分析数据在一定条件下某一数量特征的一般水平，可以对某一现象在不同时期的变化进行比较，以说明现象的发展趋势及规律。其中，平均数是一个抽象化的数值，用来说明总体各单位标志值的集中趋势。根据平均数的不同，平均分析法分为数值平均数和位置平均数。例如，在分析某个新闻平台的文章阅读量时，借助工具导出的数据可以快速找到阅读量大于平均值的文章，接下来可以继续挖掘这些文章的标题、排版、配图等规律，便于提升后

微信公众号：电子课本大全

续内容质量。

例如，某校组织学生把沙湖区大致划分为几个区域，然后统计每个区域内野鸭与湖鸥的栖息数量（表4.1.3），他们希望能够将上述数据统计出一个平均值，并做成柱状图来直观分析野鸭与湖鸥在沙湖的分布密度。

表4.1.3　野鸭与湖鸥的栖息数量

沙湖区域	野鸭/只	湖鸥/只	沙湖区域	野鸭/只	湖鸥/只
湖区1	59	27	湖区14	74	53
湖区2	40	71	湖区15	67	50
湖区3	60	9	湖区16	63	59
湖区4	50	73	湖区17	41	0
湖区5	54	60	湖区18	46	67
湖区6	53	69	湖区19	54	36
湖区7	47	81	湖区20	63	8
湖区8	55	53	湖区21	37	25
湖区9	64	54	湖区22	58	52
湖区10	77	75	湖区23	91	43
湖区11	63	58	湖区24	34	57
湖区12	41	56	湖区25	29	35
湖区13	77	87			

编写的程序如下。

```
import pandas as pd
#调用Numpy库，并命名为np
import numpy as np
from matplotlib import pyplot as plt
#调用中文字体'SimHei'（黑体）
plt.rcParams['font.sans-serif']=['SimHei']
#从“野鸭湖鸥数量.xlsx”中导入数据并转换成数据帧“数量”
数量=pd.DataFrame(pd.read_excel('D:\\野鸭湖鸥数量.xlsx'))
#mean()是numpy中求平均值的函数，axis=0表示求每一列的平均值。这里是对“数量”中的两组数据取平均值并将结果创建成新的数组“平均数量”
平均数量=np.mean(数量, axis=0)
x=range(0,2)
#以x数据为横轴，“平均数量”数据为纵轴，创建一个柱状图
plt.bar(x,平均数量)
plt.xticks(x,平均数量.index)
plt.ylabel('数量/只');plt.title('沙湖野鸭湖鸥分布密度')
#为图中柱状标识数字
for a,b in zip(x,平均数量):
    plt.text(a, b+0.05, '%s' %int(b), ha='center', va= 'bottom',fontsize=11)
plt.show()
```


微信公众号：电子课本大全

程序首先利用DataFrame()函数导入统计数据，然后用Numpy的mean()函数计算野鸭和湖鸥的平均数量，再用Pyplot模块中的bar()函数将结果制作成柱状图（图4.1.8）。从图中可以看出野鸭和湖鸥在各个湖区分布的平均数量。

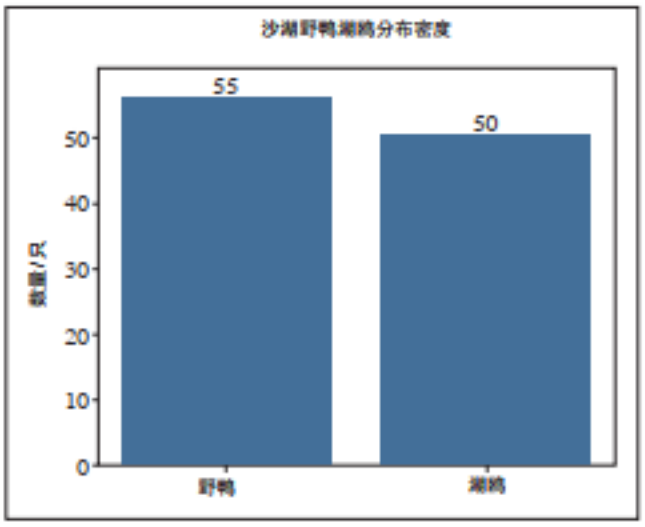


图4.1.8　平均分析法得到的柱状图

阅读拓展

DataFrame简介

DataFrame是一种表格型的数据结构，包含一组有序的列，每列可以是不同的值类型。DataFrame同时有行索引index和列索引columns。下面给出一个比较字典和DataFrame的程序。

```
>>> import pandas as pd
>>> data = {
'grade':['Grade1','Grade1','Grade1','Grade2','Grade2'],
'class':['Class1','Class2','Class3','Class1','Class2'],
'member':[43,45,44,46,47]}
>>> s = pd.DataFrame(data)
>>> s
    class   grade  member
0  Class1  Grade1      43
1  Class2  Grade1      45
2  Class3  Grade1      44
3  Class1  Grade2      46
4  Class2  Grade2      47
```

由程序可知，在字典data中依次存放了grade、class、member三组数据，但是DataFrame的列索引并不按照字典顺序生成，而是按字符首字母顺序自动排列。如果希望DataFrame的列按照指定顺序排列，就需要用columns来指定列索引的序列。如果需要删除数据，则使用drop()函数。

相关分析法

相关分析法是对总体中具有联系的标志进行分析的方法，其主体是对总体中具有因果关系标志的分析，它是描述客观事物相互间关系的密切程度并用适当的统计指标表示出来的过程。主要包括以下几方面：确定现象之间有无关系；确定现象之间关系的密切程度；测定两个变量之间的一般关系值；测定因变量估计值和实际值之间的差异。例如，在一段时期内出生率随经济水平的增长而上升，说明这两个指标是正相关关系；而在另一时期，随着经济水平的增长，

微信公众号：电子课本大全

相关的程度可分为完全相关、不完全相关和不相关三种。

出生率下降，说明这两个指标就是负相关关系。

例如，表4.1.4是某年春运期间部分城市出发地所占全国出行比例以及相关国内生产总值（gross domestic product，GDP）、人口、人均GDP等数据。接下来利用Python对上述数据进行相关分析。

表4.1.4　出行比例与GDP数据

城市	出发比例/%	GDP/亿元	人口/万人	人均GDP/元
广州	18.91	21 500	1 404	153 133.9
上海	15.68	30 133	2 418	124 619.5
深圳	14.28	22 286	1 090	204 458.7
北京	12.13	28 000	2 171	128 972.8
杭州	3.94	12 556	919	136 626.8
武汉	1.91	13 400	1 077	124 419.7
福州	1.89	7 128	757	94 161.16
温州	1.78	5 485	919	59 684.44

编写的程序如下。

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
#调用中文字体'SimHei'（黑体）
plt.rcParams['font.sans-serif']=['SimHei']
#从“春运出行.xlsx”中导入数据并转换成数据帧“春运数据”
春运数据=pd.DataFrame(pd.read_excel('D:\\春运出行.xlsx'))
#将“春运数据”中的数据组进行两两相关度计算，结果写入数组“相关度”
相关度=春运数据.corr()
#统计“相关度”中与出发比例全部相关的数据种类（包括自身）
maxn=len(相关度.ix[0])
#创建一个可以忽略出发比例自身的纵轴刻度组
y=range(1,maxn)
#创建与出发比例的相关度由大到小排序的横向柱状图
for i in y:
    #横向柱状图中创建对应数据柱
    plt.barh(i,相关度.ix[0][maxn-i])
    #为数据柱标识数字
    plt.text(相关度.iloc[0][i]-0.05, maxn-i, '%.2f%%' %(相关度.iloc[0][i]*100),
    ha='center', va= 'bottom',fontsize=11)
plt.xticks(np.arange(0,0.9,0.1),labels=['0%','10%','20%','30%','40%','50%','60%',
'70%','80%','90%'])
#给图表纵轴标上标题
plt.yticks(y,相关度.index[::-1])
plt.title('出发比例相关度排行')
plt.show()
```

以上程序利用corr()函数对四列数据进行相关分析计算，所得结果越接近1，表明相关性越强。从图4.1.9显示的结果可以看出，GDP的值和出发比例的相关度最高，达到85.39%，其次是与人口的相关

微信公众号：电子课本大全

度，最后是与人均GDP的相关度。

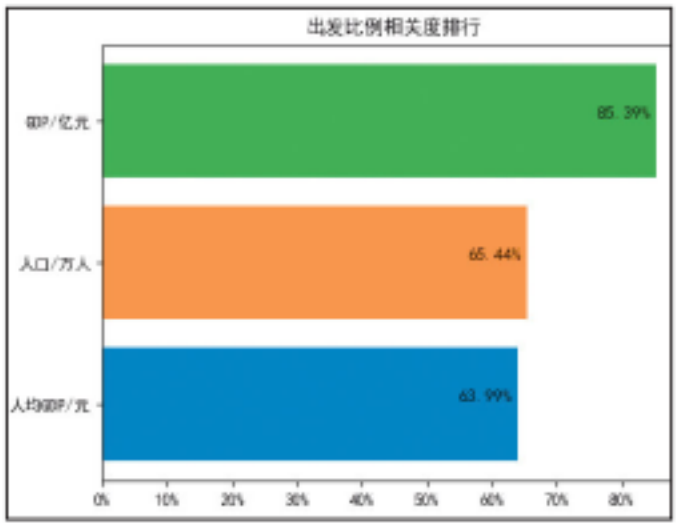


图4.1.9　用相关分析法得到的柱状图

大数据分析中，很多看似不相关的数据却能被找到相关点。例如，一家零售商通过数据分析发现，天气变冷，肉桂葡式蛋挞的销量上升了5倍，羊奶干酪打折能促进红酒的销售。这些看似互不相干的事物其实存在相关性。

思考活动

如何编程实现排行榜分析法

排行榜分析法是一种简单、大众化的分析法。排行榜存在于各种分析和各种媒体报道中，如商品排行榜、销售排行榜、福布斯排行榜、中国高校排行榜、空气质量排行榜等。

思考：如果要制作一个各省市GDP数据排行榜（参考图4.1.10），如何用Python编程实现？

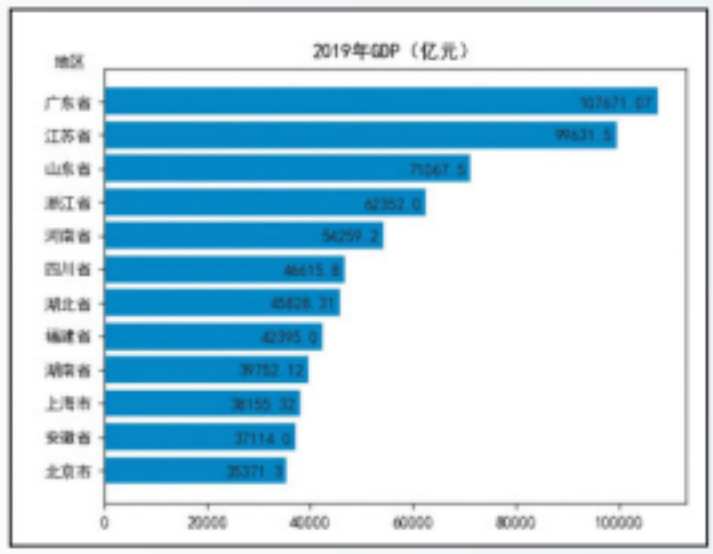


图4.1.10　用排行榜分析法得到的柱状图

实践活动

总结数据分析方法

1. 在表4.1.5中总结几种数据分析方法。

表4.1.5　常用的数据分析方法汇总

城市	基本概念	举例说明
对比分析法		
分组分析法		
交叉分析法		
平均分析法		
相关分析法		

2. 访问中华人民共和国生态环境部网站，获取2018年地表水监测数据和空气质量预报数据，选择适当的数据分析方法分析二者之间是否存在关联。

微信公众号：电子课本大全

4.1.3 数据挖掘

关系数据库系统是从20世纪70年代发展起来的。人们通过查询语言、查询处理优化和事务管理，就可以便捷地访问数据。近年来出现了“数据仓库”这种数据存储技术，其包括数据清理、数据集成和联机分析处理。联机分析处理是一种分析技术，具有汇总、合并和聚集，以及从不同的角度观察信息的能力。

随着大数据系统成为存储、访问和运营工作的重心，很多企业着眼于构建全局数据结构，从而可以全面访问来自多个源头的数据，并为真正的多用户系统提供计算服务。同时，越来越多的企业用数据流进行计算，而不只是利用经过处理并存入数据库的数据。这些数据流收集了关键业务事件并可反映业务结构，而统一的数据结构将成为构建这些大规模数据流的基础。

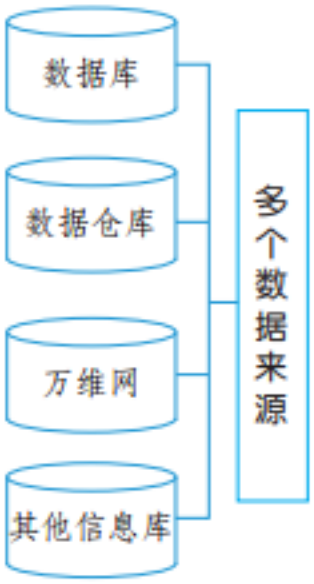


图4.1.11　数据来源多样性

数据库、万维网、各种异种数据库等都成了数据处理的对象（图4.1.11）。要从这些不同形式的数据中，通过检索技术、数据挖掘和分析技术提取知识，已成为当前研究的热点。

实践活动

自主学习了解学科新技术

查找并阅读与“数据管理与分析的新技术”相关的资料，自主了解数据管理与分析领域出现的新技术以及学科面临的挑战。

数据挖掘就是从大量、不完全、有噪声、模糊、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又具有潜在价值的信息和知识的过程。关于数据挖掘，更形象的比喻是资料探勘、数据采矿，它是知识发现的一个步骤。

数据挖掘要从大量的数据中通过算法搜索隐藏于其中的信息，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

思考活动

数据挖掘案例

【案例1】如何判断细胞是否属于肿瘤细胞。肿瘤细胞和普通细胞存在差别，过去往往需要非常有经验的医生通过病理切片才能判断。现在，通过数据挖掘，利用机器

微信公众号：电子课本大全

学习的方式，可以使系统自动识别出肿瘤细胞。这种通过“主观（医生）+客观（模型）”的方式进行多重识别，并将结果交叉验证的方式，会使结论更加可靠。

【案例2】电商知道你喜欢什么。访问一些著名的网购商城，总会有“猜你喜欢”“根据您的浏览历史记录精心为您推荐”“购买此商品的顾客同时也购买的商品”“浏览了该商品的顾客最终购买的商品”等信息提示，这些都是数据挖掘的结果。在协同过滤算法（同时考虑其他顾客的选择和行为）的基础上，搭建产品相似性矩阵和用户相似性矩阵，基于此找出最相似的顾客或最关联的产品，从而给你做推荐。

思考：数据挖掘与数据分析有哪些区别？要从事数据挖掘这个职业，应该学习哪些学科知识？

数据挖掘和数据分析都是从数据中提取一些有价值的信息，二者有很多联系，但侧重点和实现方法有所不同。进行数据挖掘，需要掌握一些机器学习所用的方法和模型知识，通过模型的训练，可以得到处理数据的最优模型。

数据挖掘涉及统计学、机器学习、模式识别、数据库和数据仓库、信息检索、可视化、算法以及高性能计算机和许多应用领域的大量知识与相关技术。

查看一些公司对数据分析、数据挖掘职位的招聘要求，可以看出这两个领域专业课程的区别。

实践活动

基金公司对客户数据的挖掘

图4.1.12是一个基金公司通过对客户数据进行挖掘，找到客户的特征，最后形成相应营销、理财策略的简单过程。

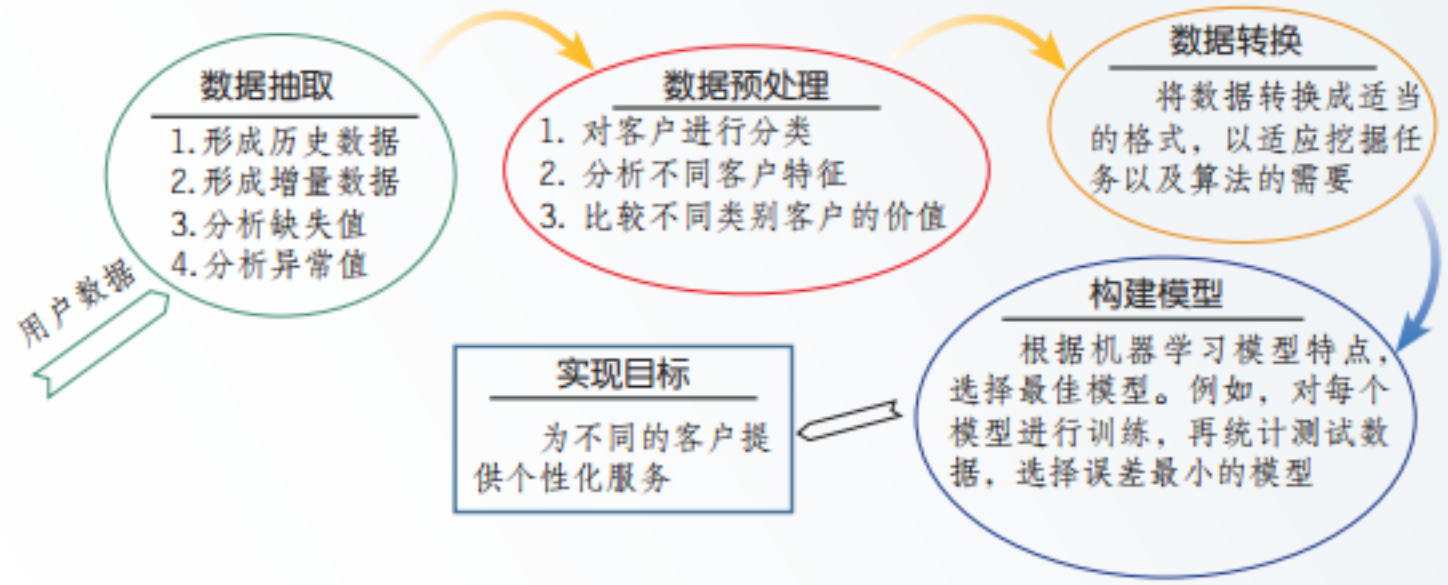


图4.1.12　基金公司客户数据挖掘

请根据图4.1.12，简要阐述数据挖掘的过程，并说出数据挖掘的意义。

微信公众号：电子课本大全

项目实施

对“地区天气数据”展开数据分析

一、项目活动

打开已经获取的地区天气数据的文件，进行以下分析工作。

1. 采用对比分析法，分析两年来同一季度的空气质量变化情况。

2. 采用分组分析法，把空气质量数据按程度分组，分析不同组里空气质量与城市的相关性。

3. 对几个月份的空气质量进行排序。

二、项目检查

1. 你在数据分析方面有哪些收获？与用电子表格软件分析数据相比，你认为通过编程分析数据有哪些好处？

2. 通过对空气质量的数据进行分析，你获得了哪些有用的信息？请通过演示文稿向其他同学展示。

练习提升

1. 简要阐述你对不同数据分析方法的理解。

2. 举例说明，你在生活中需要对某个问题中的数据进行分析时，采用的是哪种方法，最终分析结果是什么？

3. 分别给分组分析法和平均分析法列举一个案例，并给出分析方案。

4. 上网搜索数据分析的典型案例，并阐述其数据分析的方法及过程。

5. 了解数据分析与数据挖掘的区别，了解社会上对相应岗位（职业）的专业要求。

6. 阅读数据挖掘的书籍，了解更多关于数据挖掘的知识。

微信公众号：电子课本大全

4.2 数据可视化与数据报告

学习目标 ▶▶▶

- 了解数据可视化的基本过程，认识数据可视化的意义。
- 掌握数据可视化的常用方法，提高可视化操作的能力。
- 能够使用Python语言编写数据可视化的程序。

体验探索

感受数据（大数据）可视化

【案例1】访问国家统计局网站，查看一些统计数据报告，感受图文并茂的数据展示形式（图4.2.1）。

【案例2】利用数据可视化技术可以生动地呈现我国人口流动的景象。使用相应的关键词，如人口迁徙、劳动力迁徙、数据可视化等，上网查找相关资料，感受其动态效果。

【案例3】我国很多方面的发展都取得了较大的成就。请选择你想了解的数据（如GDP数据、居民消费数据、城市夜间灯光数据等），上网查找并欣赏这些数据的动态可视化效果，增强自己对国家发展变化的自豪感。



图4.2.1　国家统计局网站上的统计图

思考：数据可视化可以给人们带来什么样的感受？对展示数据有什么帮助？

数据可视化是指将数据转换成适当的可视化图表，从而将隐藏在数据中的信息直观地呈现出来。它通过计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性和动画的显示，对数据加以可视化解释。数据可视化可以帮助人们更直观地理解数据中所隐藏的

微信公众号：电子课本大全

变化趋势，挖掘数据更深的内涵。

4.2.1 数据可视化中的图形

数据可视化是为了观测和跟踪数据，更直观地发现数据之间的潜在关联。如果要实时观察数据变化的趋势，可以生成一份动态的、可读性强的图形。

图形是最直接的数据可视化方法。常用的数据图形有饼图、柱状图、折线图、散点图等，此外还有气泡图、面积图、省份地图、词云、瀑布图、漏斗图等类型，不同的图形能满足不同的展示和分析需求。图形可以是静态的或者动态的，有二维的，也有三维的。

实践活动

汇总数据统计图的样式和特点

1. 回顾用过的数据统计图，上网查阅更多的统计图样式，把它们的名称、特点和适用情况用表格的形式进行说明。

2. 查阅动态图的特点，了解数据技术的发展对丰富统计图样式带来的促进作用。

4.2.2 数据可视化的步骤

一般来说，对数据进行可视化的大体步骤如图4.2.2所示。

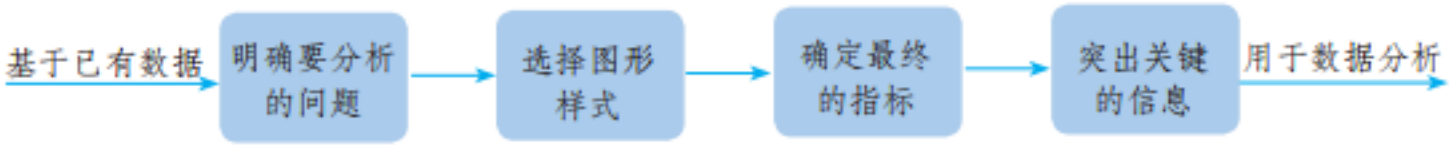


图4.2.2 数据可视化的大体步骤

明确分析的问题。要思考“这个可视化结果如何帮助读者理解这些数据”这个问题，明确这个问题有助于在后续步骤合理选择图形。

选择基本的图形。如果你非常熟悉各种图形样式的特点，就可以直接选择自己想要的图形样式。如果无法准确预知数据的特点，在数据可视化时，可以先尝试选择基本的图形样式，如饼图、柱状图、折线图等。这样就能更方便地了解自变量和因变量的关系。

确定最终的指标。通过建立基本的图形样式，对所要展现的数据情况有了一个基本的判断。此时要剔除不相干的数据变量，选择最适合展现数据关系的图表类型。

突出关键的信息。当数据可视化图形做好后，可以选择适当的

微信公众号：电子课本大全

加工方法，对图形进行修饰，目的是突出关键信息。例如，对图中关键的数据点位增加标注或者加粗坐标轴的标识等。

思考活动

数据的哪些方面可以可视化

可以从以下几个方面对数据进行可视化。

· 将指标值图形化。简单来说，就是将数据的大小以图形的方式表现。例如，用柱状图的长度或高度表现数据大小。

· 将数据关系图形化。当存在多个维度的数据时，挖掘它们之间的关系，并将其用图形表达出来，可提升图表的可视化深度。

· 将时间和空间可视化。例如，当图表要突出显示地域信息时，可用地图将空间可视化，将地图作为主背景呈现所有信息点。

· 让图“动”起来。数据图形化完成后，可结合实际情况，将其变为动态和可操控的图形，增强其交互体验。

思考：如果老师交给你一个班级成绩表，里面包含全班同学各门学科的期中成绩和期末成绩，你将从哪些方面对成绩数据进行可视化？

4.2.3 编程实现数据可视化

Matplotlib是Python的二维绘图库，利用它可以制作多种统计图形，另外还可以把它作为绘图控件，嵌入GUI应用程序中。

例如，小明所在的学习小组已经对本市某个时间段的天气数据进行了处理，现在想分析这些数据的变化情况，他们选择了折线图作为分析图表。

采用plot()函数来绘制折线图（图4.2.3），部分程序如下。

```
import matplotlib.pyplot as plt
plt.title("某城市2019年2-11月温度、湿度、
PM2.5平均值")                          #图表的标题
plt.xlabel("月份")                     #横轴标签
plt.ylabel("平均值")                   #纵轴标签
plt.plot(yf,wd,linestyle='-
',color='red',marker='<')              #wd表示温度值
plt.plot(yf,sd,linestyle='-
',color='green',marker='<')            #sd表示湿度
plt.plot(yf,pm,linestyle='-
',color='blue',mar
ker='<')                               #PM2.5
plt.legend(['温度','湿度','PM2.5']) #图例
plt.show()                             #显示图表
```

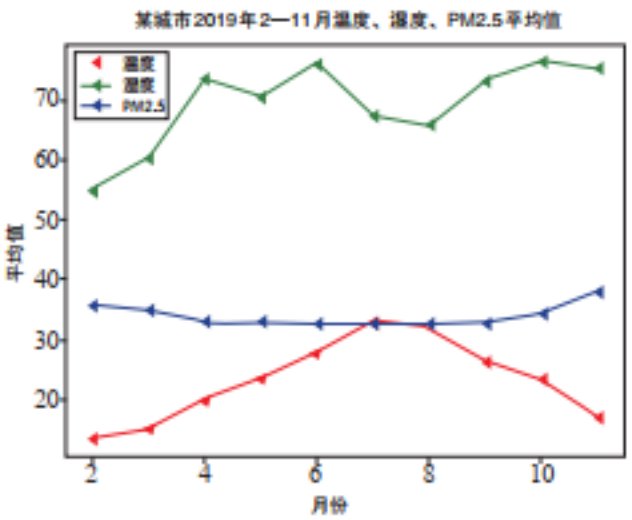


图4.2.3 天气变化的折线图

微信公众号：电子课本大全

Pyplot 模块是 Matplotlib 库中的一个绘图模块，它的每个函数或方法会根据其参数值修改绘制的图形。

首先，这段程序从 Matplotlib 中引入 Pyplot 作为 plt。接着设置图表的标题、横轴和纵轴的对象，横轴调用了月份的数据，纵轴调用了温度、湿度及 PM2.5 的值，分别用红色、绿色和蓝色的线来表示。三条折线的节点处都用朝左的三角形来表示。legend() 函数用来设置图表的图例。最后用 show() 函数的方法输出折线图。

实践活动

绘制柱状图分析数据

打开名为“供暖季 PM2.5 数据”的文件，根据数据绘制柱状图并分析数据隐含了哪些信息。

提示：柱状图的生成可以采用 bar() 函数来实现。例如，要把图 4.2.3 所示的折线图改为柱状图，可参考以下程序。

```
for row in results:                    #循环显示
    yf.append(row[0])                  #结果为：[2,3,4,5，…]
    yf2.append(row[0]+0.2)             #结果为：[2.2,3.2,4.2,5.2，…]
    yf3.append(row[0]+0.4)             #结果为：[2.4,3.4,4.4,5.4，…]
plt.title("某城市2016年2月-11月温度、湿度、PM2.5平均值")   #图表的标题
plt.xlabel("月份")                     #横轴的标签
plt.ylabel("平均值")                   #纵轴的标签
plt.bar(left=yf,height=wd,color='red',width=0.2)#left表示从原点到第一个月份的
距离,height表示数值的高度，color 表示颜色,width表示立柱的宽度
plt.bar(left=yf2,height=sd,color='green',width=0.2)
plt.bar(left=yf3,height=pm,color='blue',width=0.2)
plt.legend(['温度','湿度','PM2.5'])
plt.show()
```

思考活动

如何使图表呈现数据分布情况

小明在分析班上同学的身高、体重、肺活量等数据时，希望能使用一种图表直观形象地呈现这些数据的分布情况，该如何操作？

对于小明的要求，用散点图比较恰当。绘制散点图的程序如下。

```
# -*- coding: utf-8 -*-

import numpy as np
import pandas as pd
import csv
import matplotlib.pyplot as plt

plt.rcParams['font.sans-serif'] = ['SimHei'] #设置中文字体 宋体
df1 = pd.read_csv("boy.csv")                 #读取男生数据文件
height1 = df1['Height']                      #取出男生身高数据
weight1 = df1['Weight']                      #取出男生体重数据
```


微信公众号：电子课本大全

```
vc1 = df1['vital capacity']                                    #取出男生肺活量数据
N1 = len(height1)

df2 = pd.read_csv("girl.csv")                                  #读取女生数据文件
height2 = df2['Height']                                        #取出女生身高数据
weight2 = df2['Weight']                                        #取出女生体重数据
vc2 = df2['vital capacity']                                    #取出女生肺活量数据
N2 = len(height2)

area1=[int(x/10) for x in vc1]                                 #男生肺活量数据点的面积
area2=[int(x/10) for x in vc2]                                 #女生肺活量数据点的面积

plt.scatter(height1, weight1,s=area1,c='r',alpha=0.5,marker='s')#设置x、y取值范围
plt.scatter(height2, weight2,s=area1,c='b',alpha=0.5,marker='o')#设置x、y取值范围

#设置title和横轴，纵轴的标签
plt.title("男生女生身高体重分布图")
plt.xlabel("身高/cm")
plt.ylabel("体重/kg")
label = ["男生", "女生"]
plt.legend(label, loc = 0, ncol = 2)
plt.show()                    #展示图片
```

程序中，用scatter()函数绘制散点图，前两个参数是数组，分别指定每个点的坐标；参数s指定每个点的大小，值和点的面积成正比；参数c指定每个点的颜色，可以是数值或数组。这里使用一维数组为每个点指定了一个数值。通过红色和蓝色来表示女生和男生的数据分布情况（图4.2.4）。

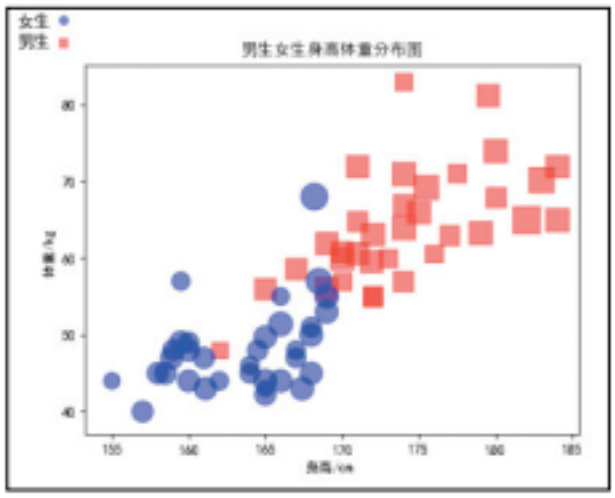


图4.2.4　反映数据分布情况的散点图

又如，某校高二年级学生各门课程的成绩被保存在score.xls文件中（表4.2.1）。现在小明要辅助班主任，利用Python的数据分析方法，展示学生不同课程的成绩分布情况。

表4.2.1　课程成绩

班级	学号	姓名	语文	数学	英语	物理	化学	生物	政治	历史	地理	技术	总分
209	20920	沈杰	75	33	71				54	37		31	301
209	20935	杨坚	77	23	37				55	42		42	276
209	20910	吴昊	87	67					65	0		48	267
209	20919	魏烨怡	44	28	66				51	34		42	265
209	20922	李雨婷	11	23	26.5				26	19		33	138.5
209	20937	宋海阳	52		35.5				36	29			152.5
210	21007	黄鑫	103	107	93.5	76					67	48	494.5
210	21012	张俊杰	85	88	79	88					65	66	471
210	21001	李思雨	98	71	104	75					74	55	477
210	21006	罗雷	85	77	71	74					66	67	440
210	21024	邓佳锋	89	84	73.5	71					58	63	438.5
210	21019	吴宇超	92	70	77	76					70	57	442
210	21014	刘灵涛	85	98	78.5	76					52	50	439.5
210	21008	金润强	93	93	79.5	62					48	56	431.5
210	21002	魏荣	96	79	108	54					51	47	435
210	21018	俞鑫超	88	75	80.5	76					54	52	425.5
210	21021	吕腾	85	72	99	63					58	50	427
210	21011	樊成富	85	91	69.5	63					58	55	421.5
210	21016	陈思阳	82	84	96.5	69					52	46	429.5
210	21022	吴振烨	80	90	79	69					55	49	422

微信公众号：电子课本大全

小明编写了以下程序。

```
import numpy as np
import pandas as pd
import pylab as pl
import xlrd
pl.rcParams['font.sans-serif']=['SimHei']
pl.rcParams['axes.unicode_minus']=False
excelpath = r"D:\score.xls"
df = pd.read_excel(excelpath,'Sheet1')
pl.figure(figsize=(10,6))
ax1=pl.subplot(231)
ax2=pl.subplot(232)
ax3=pl.subplot(233)
ax4=pl.subplot(234)
ax5=pl.subplot(235)
ax6=pl.subplot(236)
pl.sca(ax1)
pl.plot(df['语文'].dropna(),'o',color='red',label='语文')#散点图
pl.legend()
pl.xlabel('学生序号')
pl.ylabel('成绩/分')

pl.sca(ax2)
pl.hist(df['英语'].dropna(),color='blue',label='英语')#直方图
pl.legend()
pl.xlabel('成绩区间/分')
pl.ylabel('人数/人')

pl.sca(ax3)
pl.hist(df['数学'].dropna(),color='green',label='数学')
pl.legend()
pl.xlabel('成绩区间/分')
pl.ylabel('人数/人')

pl.sca(ax4)
pl.hist(df['物理'].dropna(),color='#CCCC66',label='物理')
pl.legend()
pl.xlabel('成绩区间/分')
pl.ylabel('人数/人')

pl.sca(ax5)
pl.hist(df['技术'].dropna(),color='#8B0000',label='技术')
pl.legend()
pl.xlabel('成绩区间/分')
pl.ylabel('人数/人')

pl.sca(ax6)
pl.hist(df['总分'].dropna(),color='yellow',label='总分')
pl.legend()
pl.xlabel('成绩区间/分')
pl.ylabel('人数/人')
pl.show()
```

运行程序后可以得到相关成绩统计图，其中图4.2.5所示的是物理、技术学科成绩以及总分的分布情况。

微信公众号：电子课本大全

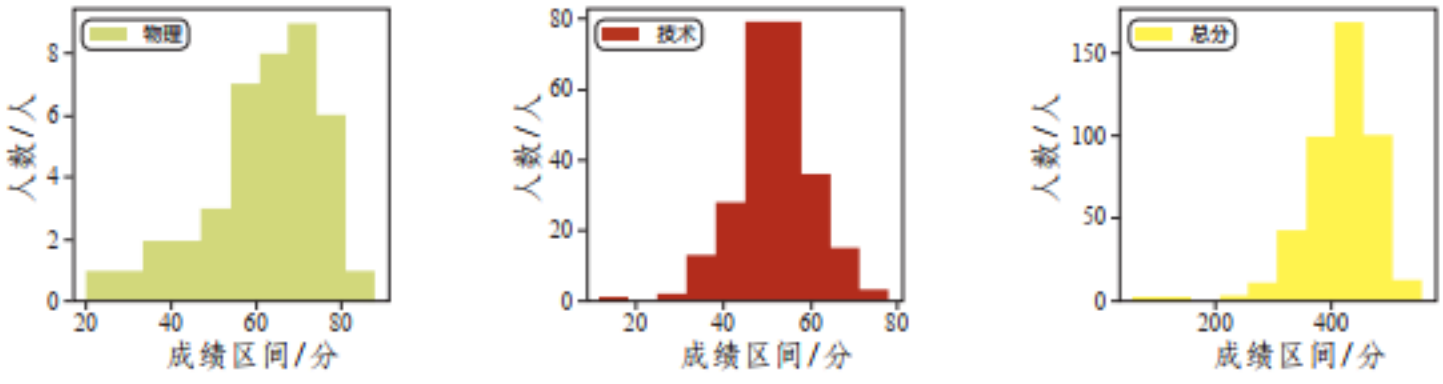


图4.2.5　物理、技术成绩及总分分布情况

思考活动

分析图4.2.5中的数据

图4.2.5中各个统计图的数据反映了什么？请按课程给出你的思考结果。

4.2.4　撰写数据分析报告

信息时代，提高数据处理和应用意识，掌握数据、应用数据以及表达自身观点是信息社会公民的重要能力。在数据分析阶段，我们已经获得了很多信息，并形成了自己对问题的看法。这时，可以进一步剖析和整合这些内容，形成有理有据的数据分析报告，并通过适当的场合展示给他人。

思考活动

如何更好地用数据说话

立恒所在的创业团队计划买一辆汽车。每个人都提出了自己的要求，最后形成了两派意见。立恒决定用汽车销售的各项数据来说服对方。于是他访问汽车数据分析平台，搜集和对比了两个品牌汽车的车型、价格、性能等方面的数据（图4.2.6），并撰写了数据分析报告在会议上讲解（图4.2.7），最终赢得了团队伙伴的支持。



图4.2.6　获取数据



图4.2.7　表达观点

思考：如果是你，会如何搜集和分析数据？如何更好地说服对方？

微信公众号：电子课本大全

根据选题的范围，数据分析报告分为专题性分析报告和综合性分析报告。例如，用户流失分析、提升用户消费分析、高中生营养配餐问题分析、小学生近视原因分析等对应的报告，都属于专题性分析报告；世界人口发展报告、全国经济发展报告、企业财务分析报告等，属于综合性分析报告。

实践活动

浏览现有的数据分析报告

1. 访问中华人民共和国生态环境部的网站，浏览环境质量方面的各种报告，体会这类综合性分析报告（图4.2.8）的特点。

2. 访问国家统计局的网站，浏览一些可视化的数据图表，获得撰写分析报告的启发。





图4.2.8　数据分析报告局部样例

撰写数据分析报告需要注意以下一些原则。

规范。数据分析报告要“以数据说话”，所使用的数据单位、名词术语一定要规范，所使用数据的来源要加以标注。



严谨。数据分析报告的编制过程一定要谨慎，体现在基础数据上，就要求这些数据真实、准确、完整，分析过程要科学、合理、全面，尊重客观事实。

突出重点。数据分析报告一定要体现分析的重点。在各项数据分析中，重点选取真实性、合法性指标，主次分明，并且对同一类问题的描述也要按照问题的重要性来排序。

有创新。首先，可以利用先进的技术手段，提高数据分析的科学性和多样性；其次，要有创新思维，应结合数据提出一些有前瞻性、可操作性的决策依据，使所做的数据分析产生一定的效益。

数据分析报告往往有一个基本的框架要求。通过学习，我们已经知道了分析报告主要包括标题页、前言、正文、结论和建议、附录等几部分。总结起来，数据分析报告应该回答以下几个问题：研

微信公众号：电子课本大全

究的问题是什么？通过什么途径获取哪些数据？利用什么工具分析数据以及分析的结果是什么？推理出哪些观点以及这些观点是否能解决最初提出的问题？

例如，在本章的项目学习中，小明所在的小组需要撰写一份数据分析报告，阐述他们的研究过程和研究结果。并要求在具体撰写报告时，可以在基本结构的基础上，根据自己的实际情况进行调整，但核心内容应该包括图4.2.9所示的几部分。



图4.2.9　数据分析报告的主要部分

实践活动

编写数据分析报告的提纲

请你为小明所在小组编写一份数据分析报告的提纲，并利用思维导图的形式来绘制，然后在学习小组内展示。

阅读拓展

数据可视化的动态效果

长期以来，数据可视化作品几乎都是静态的，如图表、地图等。随着数据处理技术、图像处理技术等领域的迅猛发展，动态交互式的数据可视化作品越来越丰富。这些作品通过灵活运用造型和色彩，把信息更加生动地表达出来，给人留下更深刻的印象。例如，有人根据世界银行1968—2016年前十名国家的GDP数据制作了一个动态图，从中能直观地看到中国经济发展的巨大变化。

制作可视化动态效果的工具有很多，如Python的plot()函数。同学们可以上网查找一些利用Python制作的动态图表小程序，并在Python中输入并运行，体验制作动态可视化图表的方法。

微信公众号：电子课本大全

项目实施

对“地区天气数据”进行可视化并完成分析报告

一、项目活动

1. 根据半年以来所选定地区的月平均气温数据，绘制折线图，并结合数据分析这半年各地区气温的变化情况。

2. 根据所选定地区冬季空气质量等级的数据（优、良、轻度污染、中度污染、重度污染），绘制散点图。

3. 小组研讨，确定分析报告的框架（表4.2.2），然后分工合作，汇总项目前几个环节的成果，撰写一份完整的数据分析报告。

表4.2.2　分析报告框架内容及分工

基本框架	内容概述	撰写人员

二、项目检查

1. 与必修课程相比，你在数据可视化的学习上获得了哪些进步？

2. 在班上展示研究成果，并选择其中一个统计图表来分析数据。

练习提升

1. 与文字、数据表格相比，用统计图形来展示数据有哪些优势？请结合实际案例进行简要的阐述。

2. 比较不同类型统计图的特点和应用场景。

3. 数据可视化的大体步骤是什么？

4. 查阅相关资料，举例说明绘制数据统计图的常用语句（书写格式）。

5. Python中有哪些适合数据分析的扩展库？

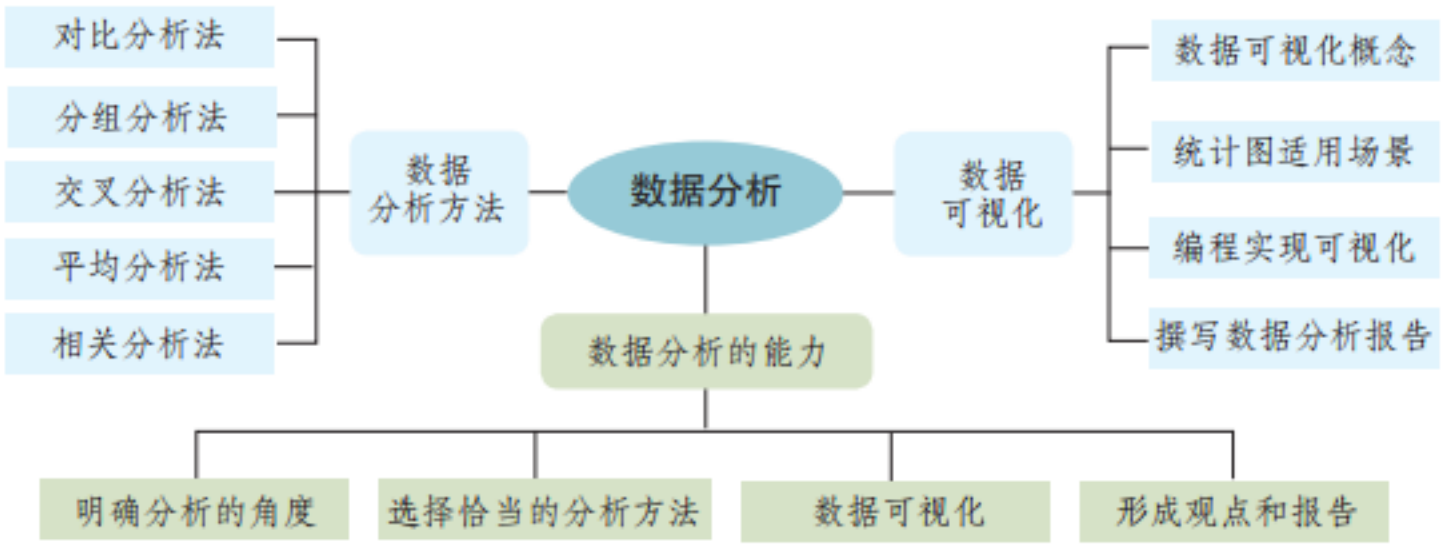
6. 描述数据分析报告的基本框架和撰写原则。

7. 尝试使用电子表格处理软件的数据透视表呈现数据。

微信公众号：电子课本大全

总结评价

1. 下图展示了本章的核心概念与关键能力，请同学们对照图中的内容进行总结。



2. 根据自己的掌握情况填写下表。

学习内容	掌握程度		
数据分析的意义	□ 不了解	□ 了解	□ 理解
数据可视化的大体步骤	□ 不了解	□ 了解	□ 理解
数据分析的主要方法	□ 不了解	□ 了解	□ 理解
简单的 Python 数据可视化程序的编写	□ 不会	□ 会	□ 熟练
数据可视化的意义	□ 不了解	□ 了解	□ 理解
数据挖掘的概念	□ 不了解	□ 了解	□ 理解
撰写数据分析报告的原则	□ 不了解	□ 了解	□ 理解
一些常见的数据分析工具	□ 不了解	□ 了解	□ 理解
能结合数据形成自己的观点，敢于表达	□ 不会	□ 会	□ 熟练

3. 回答以下几个问题，对自己的学习情况进行总结与反思。

（1）在本章的项目学习中，你在小组合作方面有什么体会？

（2）在数据分析时，编程绘制统计图形是否有难度？

（3）关于数据管理与分析这门课程，你是否对它有了更深刻的认识？

微信公众号：电子课本大全

项目评价

在完成项目活动后，请各组对项目完成情况进行评价。评价实施围绕项目主题、实施过程、分工合作、项目成果和展示交流五方面进行。根据项目评价中的评分参考，结合项目实际完成情况，确定各项评分结果，给出评分理由。同时，对项目活动进行全面梳理，指出需要进一步改进的地方。将评价内容如实填写到下列项目评价表中。

评价项	评分参考	评分（1～5分）	评分理由	待改进之处
项目主题	项目主题能反映出学科核心素养的要求（信息意识、计算思维、数字化学习与创新、信息社会责任）；主题任务与学习目标保持一致			
实施过程	项目研究计划详细，准备充分；实施过程完整，过程记录翔实，资料丰富；研究数据来源渠道多，出处明确，收集方式多样，质量高；研究方法得当，技术手段适宜			
分工合作	小组成员分工明确，态度积极，参与度高；善于提出问题，分析问题，解决问题能力强；踊跃分享观点，交流充分；能在完成自己任务的前提下，乐意帮助他组完成任务			
项目成果	项目活动成果丰富，内容具体，符合项目目标要求；研究结论清晰准确，有价值，有创新，具有指导及建设意义；项目报告或作品内容完整，论述充分，表述清楚，整齐美观			
展示交流	项目展示形式新颖，综合运用多种技术呈现成果，表现力高；语言表达清晰准确，逻辑性好			
项目总分				

微信公众号：电子课本大全

后 记

本册教科书是中国地图出版社与人民教育出版社依据教育部《普通高中信息技术课程标准（2017 年版）》，由双方共同组织团队联合编写的，经国家教材委员会 2019 年审查通过。

本册教科书的编写，集中反映了我国十余年来普通高中课程改革的成果，吸取了 2004 年版《普通高中课程标准实验教科书 信息技术》的编写经验，凝聚了参与课改实验的教育专家、学科专家、教材编写专家、教研人员和一线教师，以及教材设计装帧专家的集体智慧。本册教科书的编写人员还有李卓、张春英。为本册教科书进行装帧设计的有吕旻、李媛，摄影或提供照片的有新华社记者等。

我们感谢所有对教科书的编写、出版、试教等提供过帮助与支持的同仁和社会各界朋友。同时，我们还要感谢 2004 年版《普通高中课程标准实验教科书 信息技术》的编写人员。

本册教科书出版之前，我们通过多种渠道与教科书选用作品（包括照片、画作）的作者进行了联系，得到了他们的大力支持。对此，我们表示衷心的感谢！恳请未联系到的作者与我们联系，以便及时支付稿酬。

我们真诚地希望广大教师、学生及家长在使用本册教科书的过程中提出宝贵意见。我们将集思广益，不断修订，使教科书趋于完善。

联系方式

电　　话：010-83543863　　010-58758866

电子邮箱：sinomaps@yeah.net　　jcfk@pep.com.cn

中国地图出版社教材出版分社

人民教育出版社课程教材研究所信息技术课程教材研究开发中心

2019 年 4 月

微信公众号：电子课本大全

PUTONG GAOZHONG JIAOKESHU

XINXI JISHU



微信公众号：电子课本大全