国家教材委员会专家委员会审核通过
2019

普通高中教科书

信息技术

选择性必修 3

数据管理与分析

Shuju Guanli yu Fenxi

广东教育出版社

普 通 高 中 教 科 书

信息技术 选择性必修 3

数据管理与分析

Shuju Guanli yu Fenxi

徐福荫　主编

广东教育出版社

· 广州 ·

图书在版编目(CIP)数据

信息技术. 选择性必修3：数据管理与分析 / 徐福荫主编. —广州：广东教育出版社，2019.12（2021.1重印）
普通高中教科书
ISBN 978-7-5548-3030-7

Ⅰ. ①信… Ⅱ. ①徐… Ⅲ. ①计算机课—高中—教材
Ⅳ. ①G634.671

中国版本图书馆CIP数据核字（2019）第202794号

编写单位　广东教育出版社
主　　编　徐福荫
副 主 编　朱光明　黄国洪
本册主编　周云虹　王兴芳
核心编写人员（以姓氏笔画为序）
丁　辉　邓毅怡　梁爱梅　彭丽欧

责任编辑　严洪超
责任技编　杨启承　陈　瑾
装帧设计　何　维

信息技术 选择性必修3 数据管理与分析
XINXI JISHU XUANZEXING BIXIU 3 SHUJU GUANLI YU FENXI
广 东 教 育 出 版 社 出 版
（广州市环市东路472号12-15楼）
邮政编码：510075
网址：http://www.gjs.cn
广东新华发行集团股份有限公司发行
广东新华印刷有限公司南海分公司印刷
（佛山市南海区盐步河东中心路）
890毫米×1240毫米　16开本　9.5印张　190 000字
2019年12月第1版　2021年1月第3次印刷
ISBN 978-7-5548-3030-7
定价：11.13元
批准文号：粤发改价格［2017］434号　举报电话：12315

前　言

信息技术作为当今先进生产力的代表，已经成为我国经济发展的重要支柱和网络强国的战略支撑。信息技术涵盖了获取、表示、传输、存储和加工信息在内的各种技术。自电子计算机问世以来，信息技术沿着以计算机为核心、到以互联网为核心、再到以数据为核心的发展脉络，深刻影响着社会的经济结构和生产方式，加快了全球范围内的知识更新和技术创新，推动了社会信息化、智能化的建设与发展，催生出现实空间与虚拟空间并存的信息社会，并逐步构建出智慧社会。

数据管理与分析技术已经广泛应用于人们的日常生活与学习中，成为解决问题的重要方式。有效地管理与分析数据可帮助人们获取有价值的信息，为决策形成提供重要依据。本教科书是针对数据管理技术与数据分析方法的应用而设置的选择性必修模块。

通过本教科书的学习，同学们能了解数据管理与分析技术，能根据需求分析，形成解决方案；能选择一种数据库工具对数据进行管理，从给定数据中提取有用信息并应用于解决实际问题中；在活动过程中形成对数据特征、数据价值、数据管理思想与分析方法的认识。

本教科书按“数据需求分析”“数据管理”“数据分析”三部分内容展开，围绕信息技术学科核心素养设计了“中学生膳食和运动习惯的数据管理与分析调查”“中学生体质健康数据管理系统的需求分析与数据建模”“中学生体质健康数据管理系统的数据管理”“中学生体质健康数据管理系统的数据分析”“体验电子商务数据的管理与分析新技术应用”项目范例，教师围绕“情境→主题→规划→探究→实施→成果→评价”的项目范例主线开展教学活动，帮助同学们掌握本教科书的基础知识、方法和技能，增强信息意识、发展计算思维、提高数字化学习与创新能力，树立正确的信息社会价值观和责任感，从而促进同学们的信息素养提升。

本教科书要求同学们对现实世界中的真实性问题进行自主、协作、探究学习。同学们围绕“项目选题→项目规划→方案交流→探究活动→成果交流→活动评价”的项目学习主线开展学习活动，体验“做中学、学中创、创中

乐”的项目学习理念和“从实践入手、先学后教、先练后讲”的项目学习策略，将知识建构、技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，从而促进信息意识、计算思维、数字化学习与创新、信息社会责任的信息技术学科核心素养达成。

本教科书设置了“项目范例”“项目选题”“项目规划”“方案交流”“探究活动”“项目实施”“成果交流”“活动评价”等学习栏目，指导同学们开展项目学习活动。其中，“项目范例”是教师通过“情境”“主题”“规划”“探究”“实施”“成果”“评价”等活动，引导同学们了解开展项目学习活动的全过程；“项目选题”是同学们从真实世界选择自己感兴趣的项目主题；“项目规划”是同学们根据项目选题，制订自己的项目方案；“方案交流”是同学们展示交流自己设计的项目方案，师生共同探究、完善其方案；“探究活动”是同学们通过“问题”“观察”“分析”“阅读”“思考”“交流”“实践”“实验”“体验”“调查”“讨论”“拓展”等活动，获取知识和技能的过程；“项目实施”是同学们运用在项目学习过程中所获得的知识和技能来完成项目方案；“成果交流”是教师组织同学们展示交流项目成果，共享创造、分享快乐；“活动评价”是教师组织同学们开展项目评价活动。

本教科书各章首页的导言，叙述了本章的学习目的与方式、学习目标与内容，让同学们对整章有个总体认识。每章设置了“本章扼要回顾”，通过知识结构图把每章的主要内容及它们之间的关系描述出来，有助于同学们建立自己的知识结构体系。每章结尾的“本章学业评价”设计了基于学业质量水平的测试题，并通过本章的项目活动评价，让同学们综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。此外，本教科书为同学们提供了配套学习资源包，里面含有中学生体质健康数据管理系统、MariaDB数据与分析的各Python程序设计的源代码等，为同学们提供数据采集、管理、分析和可视化表达所需的实验数据和环境。当然，同学们还可以自己收集素材，让自己的项目学习作品更有特色。

CONTENTS

目录

第一章
数据管理与分析应用概述

随着互联网技术、多媒体技术与通信技术的迅猛发展，数据呈现爆炸式增长，数据管理与分析技术已经广泛应用于人们的生活中，成为信息社会中解决问题的重要方式。

本章将通过“数据管理与分析调查”项目，进行自主、协作、探究学习，让同学们认识到数据是一种重要的资源；通过科学管理与分析数据，可以使数据实现其应有价值；感受数据管理与分析技术的重要性，从而将知识建构、技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，促进信息技术学科核心素养达成，完成项目学习目标。

- 数据管理与分析技术
- 数据管理与分析的重要性及应用价值

项目范例　中学生膳食和运动习惯的数据管理与分析调查

情境

为了促进学生体质健康发展，激励学生积极进行身体锻炼，根据教育部《国家学生体质健康标准（2014年修订）》，国家要求各地区和各学校开展相关调研工作，并做好数据管理与分析工作。

为了认识中学生的膳食和运动习惯，促进学生体质健康发展，为学校食堂和学生膳食个人习惯等方面做出相应的改善建议和措施，某中学进行了一次全校“中学生膳食和运动习惯”的调查活动。

主题

中学生膳食和运动习惯的数据管理与分析调查

规划

根据项目范例的主题，在小组中组织讨论，利用思维导图工具，制订项目范例的学习规划，如图1–1所示。

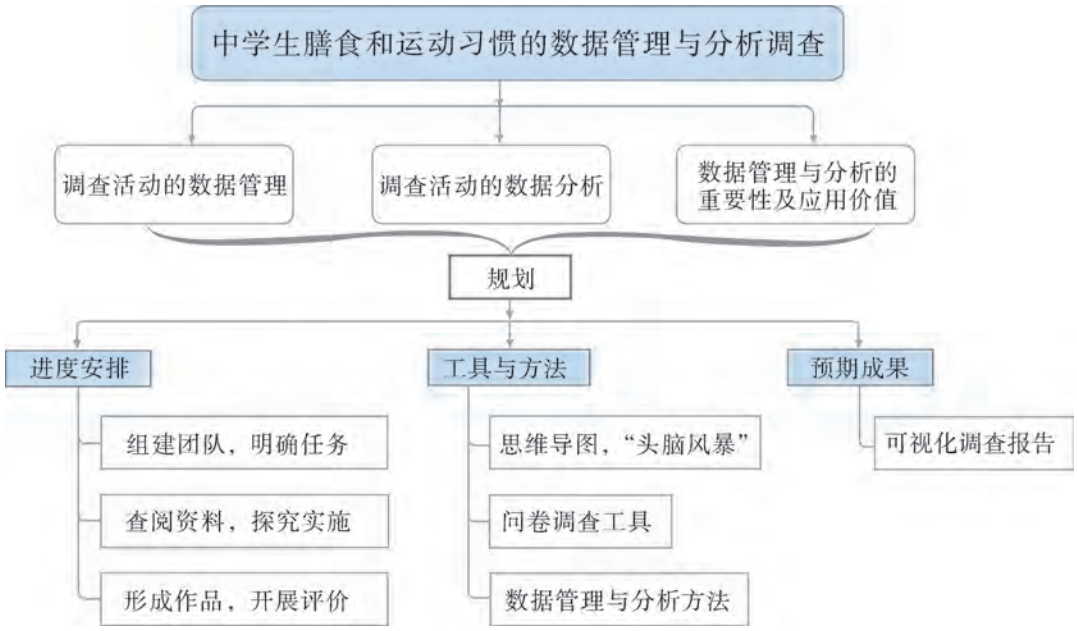


图1–1　“中学生膳食和运动习惯的数据管理与分析调查”项目学习规划

探究

根据项目学习规划的安排，通过调查、案例分析、文献阅读和网上资料搜索，开展“中学生膳食和运动习惯的数据管理与分析调查”项目学习探究活动，如表1-1所示。

表1-1　“中学生膳食和运动习惯的数据管理与分析调查”项目学习探究活动

探究活动	学习内容		知识技能
调查活动的数据管理	分析数据管理技术。	正确描述数据。	认识到数据是一种重要的资源。
		认识数据管理技术及发展。	
调查活动的数据分析	体验数据分析技术。	了解数据分析技术及特征。	
		体验数据分析技术的基本步骤和方法。	
数据管理与分析的重要性及应用价值	体验数据管理与分析的重要性。	体验数据管理技术的重要性。	感受数据管理与分析技术的重要性。
		体验数据分析技术的重要性。	
	认识数据的应用价值。	数据的预测性价值。 数据的挖掘性价值。 数据的分析性价值。	认识到通过科学管理与分析数据，可以使数据实现其应有价值。

实施

实施项目学习各项探究活动，进一步认识中学生膳食和运动习惯的数据管理与分析调查。

成果

在小组开展项目范例学习过程中，利用思维导图工具梳理小组成员在“头脑风暴”活动中的观点，建立观点结构图，运用多媒体创作工具（如演示文稿、在线编辑工具等），综合加工和表达，形成项目范例可视化学习成果，并通过各种分享平台发布，共享创造、分享快乐。例如，运用在线编辑工具制作的“中学生膳食和运动习惯的数据管理与分析调查”可视化报告，可以在教科书的配套学习资源包中查看，其目录截图如图1-2所示。

图1–2　“中学生膳食和运动习惯的数据管理与分析调查”可视化报告的目录截图

评价

根据教科书附录2的“项目活动评价表”，对项目范例的学习过程和学习成果在小组或班级上进行交流，开展项目学习活动评价。

项目选题

同学们以3～6人组成一个小组，选择下面一个参考主题，或者自拟一个感兴趣的主题，开展项目学习。

1. 中学生早餐营养搭配的数据管理与分析调查
2. 校园歌手大赛成绩的数据管理与分析调查
3. 图书馆图书借阅的数据管理与分析调查

项目规划

各小组根据项目选题，参照项目范例的样式，利用思维导图工具，制订相应的项目方案。

方案交流

各小组将完成的方案在全班进行展示交流，师生共同探讨、完善相应的项目方案。

1.1　数据管理与分析技术

1.1.1　数据管理技术与方法

在开展“中学生膳食和运动习惯的数据管理与分析调查”项目时，我们可以直接利用发放问卷的形式采集数据，也可以利用网络工具设计问卷来直接采集和管理数据。而对于项目调查活动中所涉及的各种数据，可以利用表格数据分析工具（如Excel，SPSS等），进行统计、分析及研究，并借助数据库技术、大数据技术去学习数据管理与分析的基础知识。

探究活动

同学们结合“中学生膳食和运动习惯的数据管理与分析调查”项目活动问卷需求，通过获取有关的资料，请尝试分别对膳食种类信息、膳食喜好信息、运动种类信息、中学生运动喜好信息等信息进行数据的描述，认识数据管理技术。

1. 数据的描述

“中学生膳食和运动习惯的数据管理与分析调查”项目包括制订问卷、发放问卷、收集问卷、处理问卷、数据整理、数据分析、撰写调查报告、修改提交报告等过程。在开始制订问卷时，就需要将各种数据规范化，要对数据进行合理解释和描述，有效地进行拆解和组合，从而适于数据的处理和分析，达到数据有效管理的目标。

（1）认识数据。

数据是现实世界客观事物的符号记录，是信息的载体，是计算机加工的对象。在计算机科学中，数据是对所有输入计算机并被计算机识别、存储和处理的符号的总称，是联系现实世界和计算机世界的途径。在大数据时代，数据不仅是信息的载体，也是人们提取信息做出决策的重要依据，成为人们认识和理解现实世界客观事物的重要资源。如图1–3所示是国家统计局网站上发布的权威数据。

图1–3　国家统计局网站上发布的权威数据

（2）解释数据。

数据是形成信息和知识的源泉，是计算机程序加工的“原料”。一般来说，数据主要包括结构化数据（structured data）、半结构化数据（semi-structured data）、非结构化数据（unstructured data）。合理解释数据，首先要对数据进行选择或将数据转化为结构化数据，其次要将数据融入相应的背景进行解读，对数据做出合理解释，转化为有意义的信息。

因此，数据和信息都是可解释的。如图1-4所示，单纯性数据37.8，可以是毫无意义的，但是添加一定背景，就如同为数据赋予了骨架。例如，一名叫小睿的两岁儿童，用体温计测量的腋下体温为37.8 ℃，此时37.8就转化为有意义的信息，即说明小睿为低烧状态，应该先采取一定的降温手段，再去深入探讨导致这次低烧的原因及预防方法。

图1-4　合理解释数据

（3）描述数据。

在日常生活中，人们通常直接用自然语言对人或事物进行描述，自然语言可以是中文、地方方言或者外文等。在计算机中，为了存储和处理这些人或事物，就要抽出对这些感兴趣的特征组成一条记录来描述。学生记录就是描述学生的数据，这样的数据是有结构的，是记录计算机中表示和存储信息的一种格式表达。

例如，在生活中可以这样来描述一名学生的个人数据信息：彭睿同学，男，学号是20190506873，2002年12月12日出生，出生在广东省广州市，2016年入学，高二（5）班，家庭成员有父亲、母亲、爷爷、奶奶，籍贯是河南省洛阳市，在学校住宿，学习成绩优秀，不懂广州本地方言等。

通过认识、解释和描述数据，结合“中学生膳食和运动习惯的数据管理与分析调查”项目活动实际需求，在计算机中应描述为：

（彭睿，男，20190506873，20021212，广东省广州市，2016，高二年级，5班）

2. 数据管理技术及发展

数据管理是指对数据的采集、分类、组织、编码、存储、查询和维护等活动，从而实现数据的规范化和结构化。以数据库为代表的数据管理技术已经历近半个世纪的大发展。数据管理技术已经从第一代的层次与网状数据库系统、第二代的关系数据库系统，发展到新一代数据库，继而发展到大数据管理技术，人们在不断努力开发适合最新需求的数据库管理系统，如图1-5所示是数据管理技术发展的主要历程示意图。

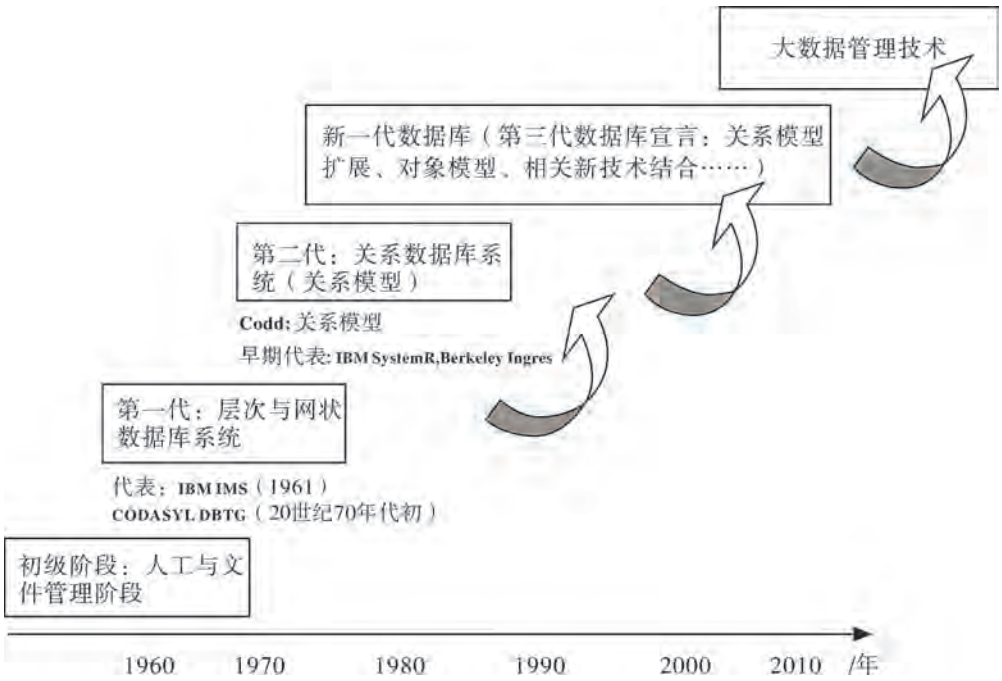


图1–5 数据管理技术发展历程示意图

3. 数据管理方法

数据管理方法有五大类：人工管理、文件系统管理、数据库系统管理、新一代数据库和大数据管理技术。

（1）人工管理。

20世纪50年代中期以前，计算机刚刚诞生不久，硬件和软件的发展水平都比较低，计算机主要用于科学计算，数据量少，数据结构简单，用户一般用机器指令编写程序，通过纸带输入程序和数据，如图1–6所示。

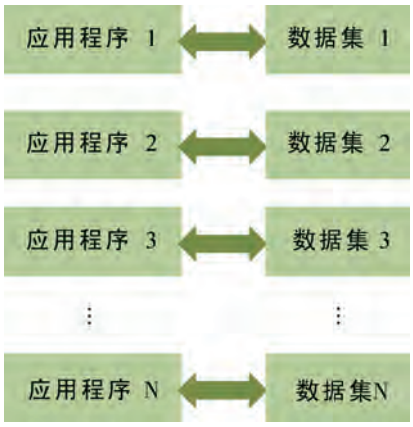


图1–6 人工管理阶段应用程序与数据之间的对应关系

这个时期数据管理处于人工管理阶段，其主要特点有：

第一，没有专门的软件用来管理数据，管理数据需要依赖应用程序本身来处理。

第二，数据和程序是紧密联系的，一组数据只能对应一个应用程序，而数据又不能共享。

第三，数据通常包含在程序中，不具有独立性，一旦数据的结构发生变化，应用程序就要作相应的修改。

（2）文件系统管理。

20世纪50年代后期至60年代中期，数据管理进入了文件系统阶段。在文件系统中，数据可按其内容、结构和用途组织成若干个独立的文件，应用程序可以通过操作系统从文件中读写数据，如图1-7所示。

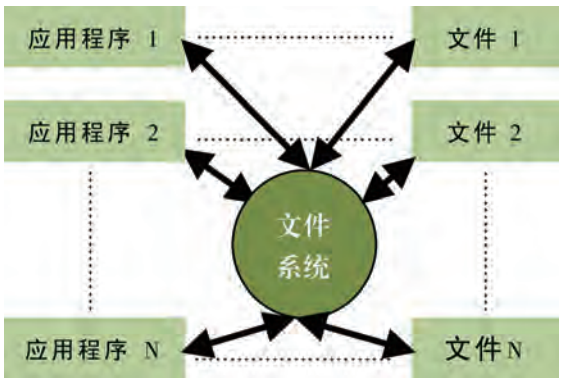


图1-7　文件系统阶段应用程序与数据之间的对应关系

在文件系统中，文件可以与程序分离，有利于长期保存，与人工管理相比，取得了长足的进步，但仍然存在以下问题：

第一，数据独立性差。在文件系统中，数据文件是按照应用程序的具体要求建立的，程序改变，将引起文件结构改变，因此程序与数据之间仍缺乏数据独立性。

第二，数据冗余度大。在文件系统中，文件一般为某一用户或用户组所有，文件仍然是面向应用的，因此数据共享性差，冗余度大。同时由于数据重复存储，各自管理，容易产生数据的不一致性。

第三，数据的安全性和完整性难以保障。文件之间相互独立，缺乏集中管理，数据的完整性和安全性等无法得到保证。

（3）数据库系统管理。

数据库（Database，DB）是按照数据结构来组织、存储和管理数据的仓库。数据库系统（Database System，DBS）克服了文件系统的缺陷并提供了对数据更高级、更有效的管理，如图1-8所示。这个阶段的程序和数据的联系通过数据库管理系统（Database Management System, DBMS）来实现。

图1-8　数据库系统阶段应用程序与数据之间的对应关系

数据库管理系统是对数据库进行管理的通用软件系统，是数据库系统的核心。它具有对数据库中的数据资源进行统一管理和控制的功能。在数据库系统中，用户对数据库进行的各种操作都是通过数据库管理系统实现的，因而使数据库中的数据具有较大的独立性。

数据库应用系统则是指针对某个特定目标，建立在数据库管理系统之上的计算机应用系统。简单地说，它是指引入数据库技术后的应用软件，由数据库、数据库管理系统、应用程序和用户等组成。因此，数据库、数据库管理系统和数据库应用系统是三个不同的概念，数据库强调的是数据，数据库管理系统强调的是对数据库进行存储与管理的系统软件，而数据库应用系统强调的是面向具体应用软件。

与人工管理和文件系统相比，数据库系统主要有以下特点：

第一，数据结构化。这是数据库系统与文件系统的根本区别。数据库系统中的数据按照某一特定的数据模型组织，具有特定的统一的结构。例如，在“中学生膳食和运动习惯的数据管理与分析调查”项目活动关系数据库里，数据库中的数据组成多个二维表形式，其中学生个人信息就是一个二维表，而二维表由若干记录组成，个人信息（彭睿，男，20190506873，20021212，广东省广州市，2016，高二年级，5班）就是一条记录，而每个记录又由若干属性项组成。

第二，数据共享。数据库中的数据是可以被多个应用程序共享的，这和文件系统不同。数据库中的数据可以通过数据库管理系统为多个用户所共享，冗余度小。

第三，数据具有较高的独立性。在数据库系统中，数据通过DBMS管理，使用户或应用程序在操作数据时，并不需要了解数据库中的数据是如何存储的，只需要以简单的逻辑结构来操作数据。

第四，数据的安全性得到保证。在数据库系统中，数据的安全性和完整性由DBMS统一管理和控制。

总的来说，如果说从人工管理到文件系统，是计算机领域质的飞跃，那么从文件系统到数据库系统，则标志着数据管理技术质的飞跃。

（4）新一代数据库。

数据库新技术是一个不断发展的范畴，在数据模型的改进、与相关技术融合以及面向应用领域等方面都在不断改进与发展。

①数据模型的改进。

相对于传统的数据库而言，集成了新的技术、工具与机制的有：

面向对象数据系统（OODBS）；

时态数据库系统（TDBS）；

实时数据库系统（RTDBS）；

主动数据库系统（ADBS）。

②数据库与相关技术结合。

比较有代表性的有：

分布式数据库；

Web数据库。

③面向应用领域。

④非结构化数据库。

（5）大数据管理技术。

随着网络技术的发展，非结构化数据的数量日趋增大。这时，主要用于管理结构化数据的关系数据库的局限性越来越明显。这就催生了数据管理技术进入新一代的数据库。如iBase数据库是一种面向最终用户的非结构化数据库，Hbase是一个适合非结构化数据存储的数据库。

例如，在制订“中学生膳食和运动习惯的数据管理与分析调查”项目活动中，需要采集不同学生、班级、群体等结构化或半结构化数据，还要采集各种半结构化或非结构化数据，如文本、图像、音频、视频等数据。在面对如此多且杂乱无章的数据文件时，要根据不同的数据采用不同的数据管理技术进行处理。

在现代信息社会里，我们既可以采用现代技术化的数据管理技术，也不排除人工管理技术来管理数据，只有将不同的数据管理技术有机结合起来，才能使数据管理更加高效，特别是根据特定的需求和目的来建立对应功能的数据库管理系统，更能实现数据管理的智能化和便利化。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，剖析调查活动的数据管理技术。

1. 对调查信息进行数据描述。
2. 认识数据管理技术及其发展历史。

1.1.2　数据分析技术与方法

通过开展“中学生膳食和运动习惯的数据管理与分析调查”项目活动，根据调查得出的系列化数据，我们可以利用表格工具（如Excel，SPSS等）或专业数据分析软件对数据进行汇总和分析，从而深入认识目前中学生的膳食和运动喜好情况与原因，为区域主管部门、学校、家庭、学生个人习惯等方面做出相应的改善建议和措施。

探究活动

在“中学生膳食和运动习惯的数据管理与分析调查”项目中，问卷内容主要包括性别、年龄、身高、体重、膳食中的肉类和蔬菜比、膳食摄入量、各种食物种类摄入程度、每天平均运动时长、主要运动方式、最喜欢的运动方式等项目数据，随机抽取全校各个年

级，让同学们现场网络限时答卷。同学们自主探究与小组研讨，寻求网络检索和老师等帮助，小组汇报下列内容。

（1）根据本次调查活动项目的目的，同学们应该从什么方面对数据进行分析？

（2）结合本次调查项目不同的数据分析类型，同学们可以采用什么数据分析技术或工具？

1. 数据分析技术

通过合理的数据管理，同学们可以得出规范化和结构化的数据。随着信息社会的发展，大数据时代的到来，数据呈现大量化、多样化、快速化、价值密度低的特征。为了提取有用信息和形成结论，进而对数据加以详细研究和概括，总结出所研究对象的内在规律，需要对数据进行分析。

一般来说，数据分析是指用适当的统计分析方法对采集来的数据进行分析，将这些大量的数据进行汇总，并做成可以被人们消化和理解的资料，从中提取有用和有价值的信息。数据分析主要分为描述性数据分析、探索性数据分析、验证性数据分析等，如图1–9所示。数据分析常常是以数和量的形式展现，通过实验、观察、调查等方式获取结果。

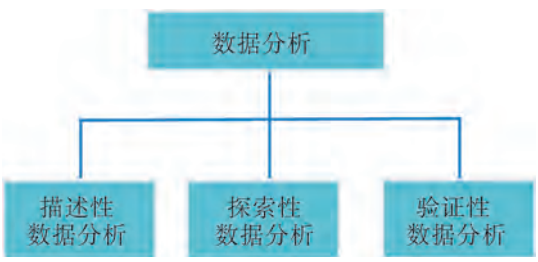


图1–9　数据分析类型

数据分析技术就是指与数据分析活动有关的技术总和，包括数据对象的描述、采集、处理、统计、分析及呈现等，在常用的数据分析中，我们会用到的工具软件主要有Excel，SPSS，Python，SAS等。在大数据时代，运用大数据的批处理、流计算、图计算及查询分析计算等功能模式，可以实现对大数据的批处理、实时分析、图结构分析、查询分析等，如常用的工具软件有MapReduce，Storm，GraphX，Dremet等。

2. 数据分析的基本步骤和方法

同学们通过调查结果采集到的数据，选择不同的数据分析方法，按照数据分析的基本步骤，得出翔实的调查数据分析结果，并进一步撰写研究调查报告。

（1）数据分析的基本步骤。

一般来说，数据分析主要包括以下四大基本步骤，如图1–10所示，它们循序渐进、缺一不可、相辅相成，无论是对小型数据分析还是对大型数据分析，都是必不可少的环节。

图1–10　数据分析的四大基本步骤

①识别需求。

识别需求是确保数据分析过程有效性的首要条件，可以为采集数据、分析数据提供清晰的目标。识别信息需求是管理者的职责，管理者应根据决策过程控制的需求提出对信息的需求。

②采集数据。

有目的地采集数据，是确保数据分析过程有效的基础。根据需求，对采集数据的内容、渠道、方法进行策划。

③分析数据。

分析数据是指对采集的数据进行加工、整理和分析，使其转化为信息。

④过程改进。

过程改进是指根据数据分析目标，改进做事的过程、方法或工具。例如，对以下问题进行分析，评估其有效性：

a. 提供决策的信息是否充分、可信，是否存在因信息不足、失准、滞后而导致决策失误的问题。

b. 采集数据的目的是否明确，采集的数据是否真实和充分。

c. 数据分析方法是否合理，是否将风险控制在可接受的范围。

d. 是否在项目实施过程中有效运用数据分析。

e. 数据分析所需资源是否得到保障。

（2）数据分析的基本方法。

数据分析具有现状分析、原因分析、预测分析三大作用，因此，数据分析的基本方法对应这三大作用进行设置。常用的数据分析方法有对比分析法、平均分析法、分组分析法、结构分析法、交叉分析法，具体如表1-2所示。

表1-2　数据分析基本方法

作用	方法	数据分析方法
现状分析	对比	对比分析法、平均分析法、综合评价分析法……
原因分析	细分	分组分析法、结构分析法、交叉分析法、杜邦分析法、漏斗图分析法、矩阵关联分析法、聚类分析法……
预测分析	预测	回归分析法、时间序列分析法、决策树分析法、神经网络分析法……

随着数据库与互联网技术等的发展和应用，数据的积累不断膨胀，数据的需求也不断更新，同时带来的数据管理与分析技术也在不断进步和更新。

最后，对数据分析除了要注意选用恰当的分析方法之外，还需注意到数据的来源，如搜索引擎抓取数据、网站的HTTP响应时间数据、网站流量来源数据等。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，剖析调查活动的数据分析。

1. 确定调查活动项目的数据分析技术。
2. 体验数据分析技术的基本步骤和方法。

1.2 数据管理与分析的重要性及应用价值

数据管理首先是要保证数据的规范性、描述性、结构性、科学性，其次是要按照实际情况有针对性地对数据进行整理，为数据分析的应用过程提取有用信息，形成结论，最后对数据加以详细研究和概括，挖掘数据背后的内在规律和内在价值，从而体现数据管理与分析的重要性。一般来说，数据管理是数据分析的基础，有效地对数据进行管理与分析可以实现数据的预测性价值、挖掘性价值和分析性价值。

1.2.1 数据管理与分析的重要性

1. 数据管理技术的重要性

（1）提高元数据管理的标准化。

元数据一般是来源于监测、勘探等数据管理技术下最原始的基础数据，具有基础性和原始性，因此，在数据管理技术中，要对大量的原始资料进行整理加工，将大量的纸质材料数字化，注重元数据采集中的标准化，依据不同的规则进行分类和提取。

例如，在“中学生膳食和运动习惯的数据管理与分析调查”项目中，依据“中国居民平衡膳食宝塔”“中国食物成分表”“中学生体质健康标准”等基础元数据，结合区域中学生的膳食特点和运动习惯整理成调查问题。将设计出的问卷初稿先进行小范围初测、个别访谈以及征求专业学者的意见，经修改、删除、完善不明确和信度较差的试题后，形成最终的调查问卷，从而能反映当前中学生膳食和运动习惯的元数据基础，并采集最初的元数据。表1-3是“中学生膳食和运动习惯的数据管理与分析调查”项目活动问卷元数据框架。

表1-3　“中学生膳食和运动习惯的数据管理与分析调查”项目活动问卷元数据框架

项目	维度	对应题号
第一部分	基本信息	1，2，3，4
第二部分	正餐中肉类和蔬菜的比例	5
	晚餐摄入食物总量	6
	一天的食物摄入总量	7
	正餐外的食物摄入量排序	8
	每天平均运动总时长	9
	每天的主要运动方式	10
	喜欢的运动方式	11

元数据是科学发展和基础研究的基本支撑和本源，也是国家的重要基础信息资源，在国家的宏观决策、科技创新、防灾减灾、环境保护和国民经济的各行各业发挥着不可替代的作用。加强对元数据科学的管理有利于各项工作更好地为经济建设、社会发展和人民生活提供高质量、及时周到的服务。例如，加强气象数据的开发力度，提高气象元数据的标准化格式，充分利用网络和信息技术，丰富气象服务的原始数据，拓展数据服务空间，这样便能提高对气象信息预测的准确性和针对性，才能不断适应社会经济发展日益增长的需求。

（2）加强数据管理服务的系统性。

随着现代信息技术和网络通信技术的发展，数据管理技术的系统性也变得更加突出，只有对各项数据进行系统性的管理，才能实现数据管理技术的高效。一方面，在数据管理技术中，需要对数据进行有效的信息化处理；另一方面，要依照各项数据管理技术搭建数据共享平台，增强对数据资料系统的开发能力。

例如，通过建立“气象元资料服务系统”“气象档案管理系统”“气象台历史沿革管理系统”等，从而使气象数据管理服务系统化，提升了信息化资料服务和数据深加工服务，有效地完成数据的各项数字化、系统化管理服务。

（3）优化大数据管理技术的准确性。

面对大数据时代的到来，数据管理的准确性是考验数据管理技术的重要指标之一。因此，在大数据环境下，数据管理技术须积极应对社会需求拓展各项服务领域，积极推进网络下的科学数据共享，攻克大数据管理技术的关键性阶段，开发面向服务系统的应用终端，从而让人们更加安全有效地享受大数据环境下数据管理技术的准确性服务。

2. 数据分析技术的重要性

（1）确保数据分析的完整性。

数据分析产生的分析价值建立在详尽和真实的数据层面，数据采集的完善是完善数据分析技术的一个过程，不论是数字、文本、图表等各种结构化的数据，还是各种不同形式的半结构化或非结构化数据，最后都需要通过汇总、分析，进而做出相应的规划和决策，这就需要在数据分析的前期确保数据的完整性。依靠现有的数据分析技术手段，可以确保

数据的完整性要求。

（2）提高数据决策的准确性。

数据的完整性约束可以确保数据的准确性，随着计算机技术的飞速发展以及专业化和国际化，各种数据分析技术应运而生，数据的准确性在依托于目前各项分析技术与方法手段基础上，让更多的数据分析可以直接依靠分析技术的自动化和智能化，不仅可以降低人为的不准确性因素，更能最大效率地提高数据分析对于决策的准确性。

（3）增强数据创造的价值性。

数据分析技术是增强国民各项经济价值创造力的重要手段，实施有效的数据分析技术是使数据价值增值的最佳方式。在信息化高速发展的背景下，各商业体积累了海量数据，依靠目前数据分析技术的数据仓库（DW）技术、数据挖掘技术，通过积极探讨商业经营效益的分析，可以促进商业数据的最大价值化。

在信息化和网络化不断发展的时代，特别是大数据时代的产生和发展，数据已经成为衡量效益的重要指标，同时也为其对于科学的评估提供了重要的参考资源。因此，随着社会和科技的不断发展，数据管理和分析技术的重要性也变得越来越突出，数据管理与分析技术的发展也必将引领数据时代的健康发展。

探究活动

讨论

结合数据管理与分析技术的重要性，同学们查找有关资料，结合实例来讨论影响数据管理和分析技术重要性发展的因素主要有哪些。

1.2.2 数据管理与分析的应用价值

通过项目范例的学习，我们认识到数据是一种重要的资源，并通过科学管理与分析数据，可以使数据实现其应有的价值。

1. 数据的预测性价值

（1）气象预测。

气象中的气流、风速、云层等各种数据通过系统软件的数据分析，能够比较准确地预报某区域在某时间段的气象情况。例如，针对大部分自然灾害均由气象因素引发这一现状，广东省以科技创新加强气象现代化建设，着力构建未雨绸缪式的气象趋势预测及高效有序的预警信息发布体系，更好地服务于民，使得相关部门和公众在应对恶劣天气突发事件过程中游刃有余，最大限度预防和减少突发事件可能造成的危害。如图1-11所示是广东省突发事件预警信息发布中心内的区域数值天气预报重点实验室。

图1–11　区域数值天气预报重点实验室

（2）工业预测。

工业中的生产过程、生产产品以及各种资源等丰富的数据经过有效分析，往往能优化生产工艺和流程，节约生产成本，降低能耗，增加利润。

例如，由于航空公司的自身原因、机场流量控制、机场航空管制、天气恶劣等原因经常会导致航班延误，由民航局发布的2012年民航行业的统计数据，称航班准点率为74.83%。但是某知名公司却做到了能够比航空公司更准确地预测信息。据报道，该公司已建立155处无源雷达接收站，每4.6秒接收一次雷达眼监测到的每架飞机的信息，通过这些信息可以准确了解每一架飞机在空中飞行的情况以及飞机的着陆时间。公司还建立自己的数据库，将自测的信息以及其他信息全部备案保存，经过十多年的积累，公司存储了海量的航空信息，拥有了其他任何公司都无法比拟的数据资源。

（3）商业预测。

商业活动中海量的数据通过系统软件的分析，能够准确得出某一类商品的市场行情和发展趋势，生产者、销售者可以借此制定生产、经营策略，消费者则可以选购到性价比高、自己喜爱的商品。

现代信息社会里，通过利用计算机或手机等终端进行网络购物的现象已经越来越普遍，同学们通过登录相关知名的网络购物平台，体验平台中的数据管理与分析。小组交流平台的商业预测性价值主要体现在哪些方面，有哪些效果。

2. 数据的挖掘性价值

一般来说，数据挖掘就是从无意义的数据中提取有意义的信息，指导我们在结构化数据中发现潜在的关系和规律。数据挖掘有三个阶段：把数据变得透明，让大家看到数据；可以提问题，可以形成互动做出实时分析；数据要具有某些预测功能。

例如，通过网络留言挖掘顾客的意见，顾客在博客、论坛、社交网站，甚至微博、微信朋友圈中用文字或图片记录的消费体验，对商品和服务发表的看法和评价，是一种非结构化数据。

如何把散布在网络上各种结构化、半结构化、非结构化的数据资源进行整合，从中自动挖掘有价值的信息和知识，从而上升为智慧，实现从数据到信息，再到知识，最后到智慧的转变，便是当前数据挖掘面临的巨大挑战之一。如图1-12所示是从数据到信息、到知识、最后上升为智慧的四级跳模型示意图。随着信息社会与大数据的发展，数据挖掘技术主要有决策树、聚类、时间序列、贝叶斯分类、线性回归、关联规则、类神经网络、Logistic回归等，如图1-13所示。

图1-12　数据—信息—知识—智慧的四级跳

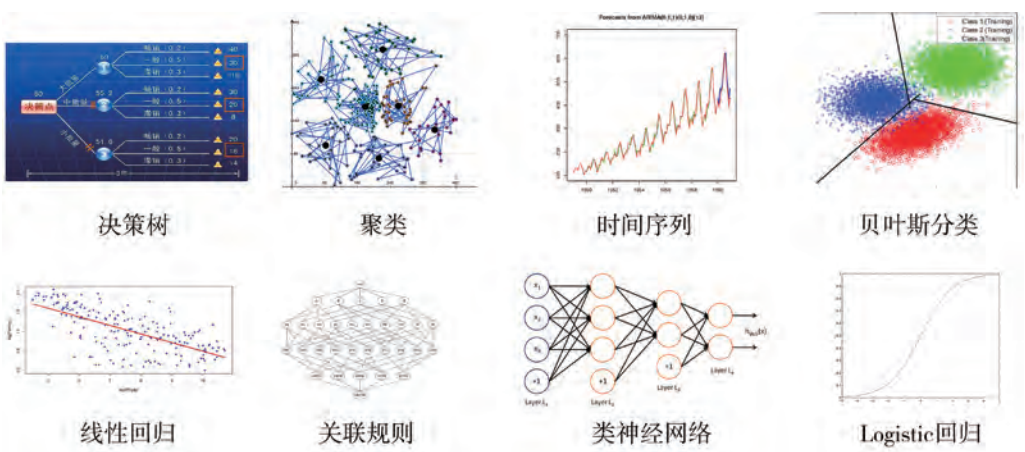


图1-13　常见的数据挖掘技术

3. 数据的分析性价值

数据管理与分析无论是在企业还是在政务部门的各种分析中，都具有数据的分析性价值，主要包括以下三种分析性价值：

（1）现状性分析。

数据的现状性分析可以提供企业或政务部门在现阶段整体的运营情况，以及各企业或政务部门的各项业务的构成，其中包括各项业务的发展和变动情况。在“中学生膳食和运动习惯的数据管理与分析调查”项目的设计之初就是想摸清区域内中学生膳食和运动习惯的现状，进而呈现数据进行现状性分析。数据的呈现也可称为数据的可视化表达，常用的呈现方式主要有柱形图、条形图、折线图、雷达图、饼状图、圆环等。一般来说，简单的数据呈现主要是采取基本图表形式进行统计分析，当对整体项目调查活动完成导出数据

后，数据分析还需要对数据进行价值分析和数据的可视化表达，根据不同的数据呈现需要采用不同的手段，如表1-4所示是常用的数据可视化工具及简介。如图1-14所示是本次调查活动中中学生对食物种类喜好程度的统计情况，用折线图和雷达图呈现。

表1-4　常用的数据可视化工具及简介

数据可视化工具	简　　介
Many Eyes	可以得到的图表有散点图、矩阵图、网络图、条形图、直方图、气泡图、线图、堆叠图、饼图、树形图、字树、标签云等。
iCharts	分免费版和商业版：私人图表、自定义模板。
Wolfram Alpha	可以输入字符串识别各种数据。
Visualize	可以得到的图表有图表、地图、示意图、仪表板等。
Data Wrangler	可以实现清洗和重新整理数据：分割、提取、填充、合并、包装、删除、推广、折叠、展开、调换不同的数据点等。

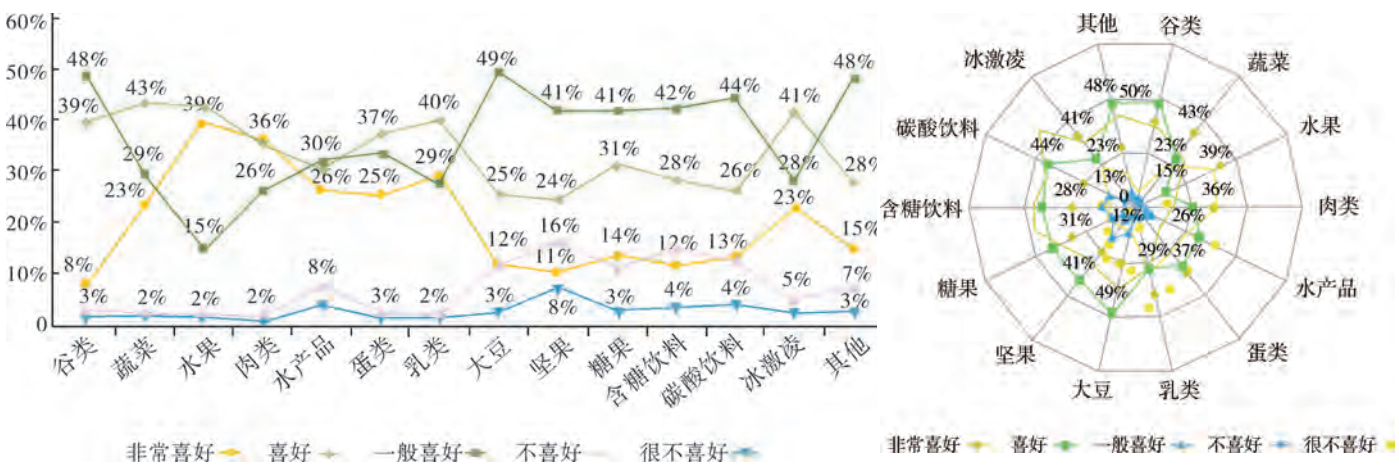


图1-14　中学生对食物种类喜好程度的数据分析的折线图和雷达图

交流

同学们针对“中学生膳食和运动习惯的数据管理与分析调查”项目的设计意图，结合问卷的框架结构，展开小组研讨，分析数据可以用哪些具体的方式来进行可视化表达，进而对数据进行分析。

（2）原因性分析。

数据的原因性分析可以确定企业或政务部门所存在的问题，认清形势，并针对原因做出相应的解决方案。

（3）发展性分析。

数据的发展性分析可以对企业或政务部门的发展趋势做出推测，便于制订运营计划和发展。

数据管理与分析是让数据产生价值的手段，而数据的运用才是数据价值的体现，在项目实践过程中，数据管理与分析将形成数据分析报告，体现数据的应用价值。

互联网的一天

互联网所涵盖的范围非常广阔，人们产生、分享和消耗的数据量很难用实体形式来衡量，那么到底有多少数据产生呢？这些数据就是一种重要的资源。

2012年“互联网的一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多；发出的社区帖子达200万个；卖出的手机为37.8万台，高于全球每天出生的婴儿数量37.1万。国际数据公司（IDC）的研究结果表明，2008年全球产生的数据量为0.49 ZB，2009年的数据量为0.8 ZB，2010年增长为1.2 ZB，2011年的数据量更是高达1.82 ZB，相当于全球每人产生200 GB以上的数据。

事实上，当我们仍然在把微博等社交平台当作抒情或者发表议论的工具时，某些敛财高手却正在挖掘这些互联网的“数据财富”，先人一步用其预判市场走势，而且取得不俗的收益。这些庞大数字，意味着什么？它意味着，一种全新的致富手段也许就摆在面前，它的价值堪比石油和黄金。

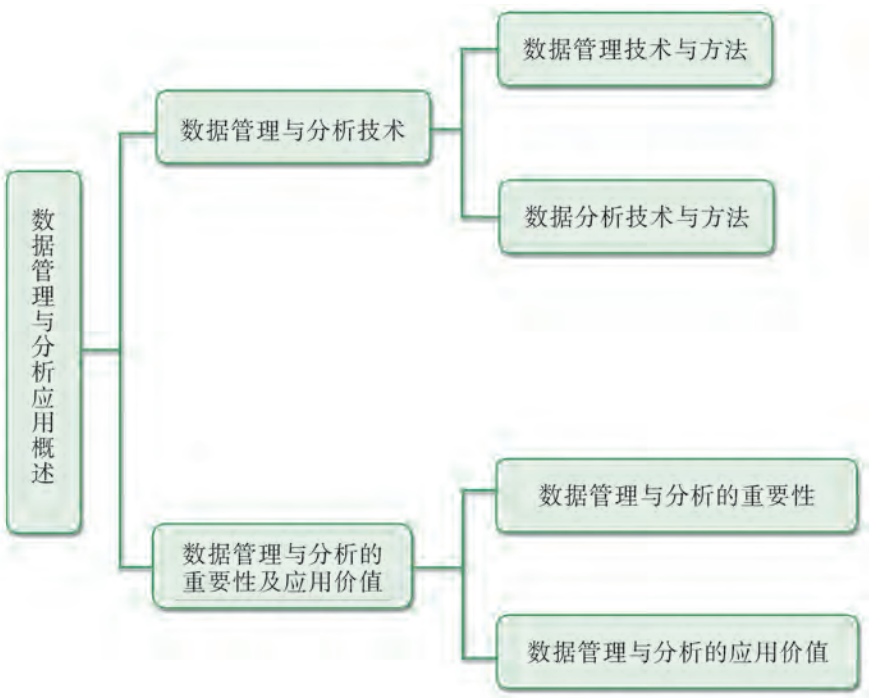
各小组根据项目选题及拟订的项目方案，结合本节所学知识，体验数据管理与分析的重要性，认识数据的应用价值，并参照项目范例的样式，撰写相应的项目成果报告。

各小组运用数字化学习工具，将所完成的项目成果，在小组或班级上进行展示与交流，共享创造、分享快乐。

各小组根据项目选题、拟订的项目方案、实施情况以及所形成的项目成果，利用教科书附录2的“项目活动评价表”，开展项目学习活动评价。

本章扼要回顾

同学们通过本章学习，根据“数据管理与分析应用概述”知识结构图，扼要回顾、总结、归纳学过的内容，建立自己的知识结构体系。



回顾与总结

本章学业评价

同学们完成下列测试题（更多的测试题可以在教科书的配套学习资源包中查看），并通过“本章扼要回顾”以及本章的项目活动评价，综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。

1. 单选题

（1）按照数据的结构类型划分，以下不属于数据类型的是（　　）。

A. 结构化数据　　B. 半结构化数据　　C. 非结构化数据　　D. 结构型数据

（2）以下不属于数据管理技术阶段的是（　　）。

A. 人工管理　　B. 文件系统管理　　C. 资源管理器管理　　D. 数据库系统管理

（3）一般来说，数据分析主要有以下项目：①采集数据；②识别需求；③过程改进；④分析数据。根据数据分析的基本步骤进行排序，正确的是（　　）。

A. ① ② ③ ④　　B. ② ① ④ ③　　C. ① ② ④ ③　　D. ② ① ③ ④

2. 思考题

在大数据时代，如何有效地使用数据管理与分析技术？通过调查谈谈在不同行业的方法和经验，从中领会管理大数据的三大价值（时间价值、管理价值和经济价值）。

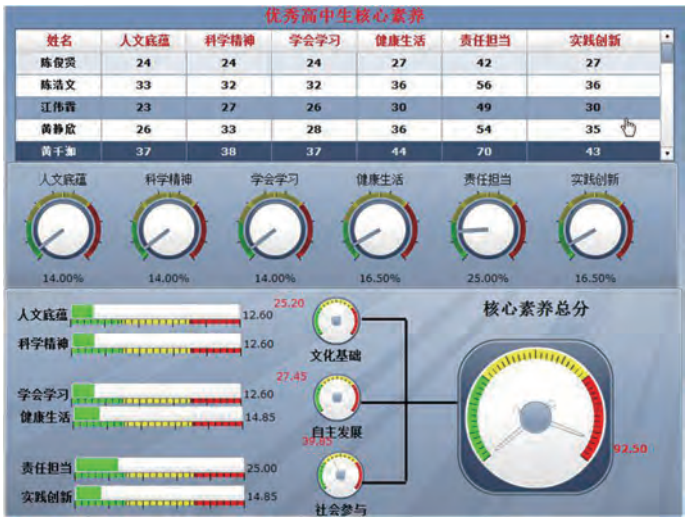
3. 情境题

2016年9月，教育部发布了《中国学生发展核心素养》报告，指出了目前中国学生要做全面发展的人，要从文化基础、自主发展、社会参与三大层面努力成为具有优秀核心素养的学生，主要包含人文底蕴、科学精神、学会学习、健康生活、责任担当、实践创新等六大核心素养，结合有关数据管理与分析的技术完成下列问题。

（1）小睿同学想通过《中国学生发展核心素养》有关项目的测定，明确自身的核心素养的现状。结合下图模型，为了更好地对数据进行管理和分析，应该如何根据核心素养的栏目来设计本次调查活动？

“中国学生发展核心素养”模型

（2）在中学生核心素养的调查活动中，小睿同学利用Xcelsius可视化软件工具得出了如下图所示的全年级数据呈现模型，试分析这些呈现方式的合理性和可行性，并对该呈现方式提出建议。



Xcelsius可视化软件分析模型

第二章
需求分析与数据建模

数据管理与分析技术已经广泛应用于人们的日常生活与学习中，成为解决问题的重要方式。正确的需求分析与数据建模是有效管理与分析数据的关键。

本章将通过“数据管理系统的需求分析与数据建模”项目，进行自主、协作、探究学习，让同学们初步了解分析业务需求、建立数据管理与分析问题整体解决方案的基本过程；尝试对既定方案进行分析、评价，发现问题并优化方案；了解数据采集途径的多样性；能利用适当的工具对数据进行采集和分类；认识噪声数据现象和成因；理解不同结构化程度数据的区别和在管理与应用上的特点；了解关系数据模型的基本概念；掌握设计较简单关系数据库的逻辑结构的方法，从而将知识建构、技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，促进信息技术学科核心素养达成，完成项目学习目标。

- 项目需求分析与解决方案
- 数据的采集与分类
- 建立关系数据模型

项目范例　中学生体质健康数据管理系统的需求分析与数据建模

情境

青少年强则国强，青少年健康事关国家和民族的未来。健康体魄是青少年为祖国和人民服务的基本前提，是中华民族旺盛生命力的体现。近年来，为了贯彻落实健康第一的指导思想，建立健全国家学生体质健康监测评价机制，激励学生积极参加身体锻炼，教育部印发了《国家学生体质健康标准（2014年修订）》（以下简称《标准》），要求各学校每学年开展覆盖本校各年级学生的《标准》测试工作，并根据学生学年总分评定等级。只有达到良好及以上的学生，方可参加评优与评奖。

为了及时跟踪和了解当前本地区中学生的体质健康情况，有必要以《标准》为依据记录该地区学生的体质健康数据，并对记录的数据进行统计分析以便做出科学的指引。为此，计划建立一个数据管理系统，希望能有效地解决以上问题。

为了完成“中学生体质健康数据管理系统的需求分析与数据建模”项目，首先需要了解作为数据管理系统使用者的用户到底有哪些业务需求，提出相应的解决方案，需要采集哪些数据，以及对这些数据如何建模，这就是需求分析与数据建模阶段。

主题

中学生体质健康数据管理系统的需求分析与数据建模

规划

根据项目范例的主题，在小组中组织讨论，利用思维导图工具，制订项目范例的学习规划，如图2–1所示。

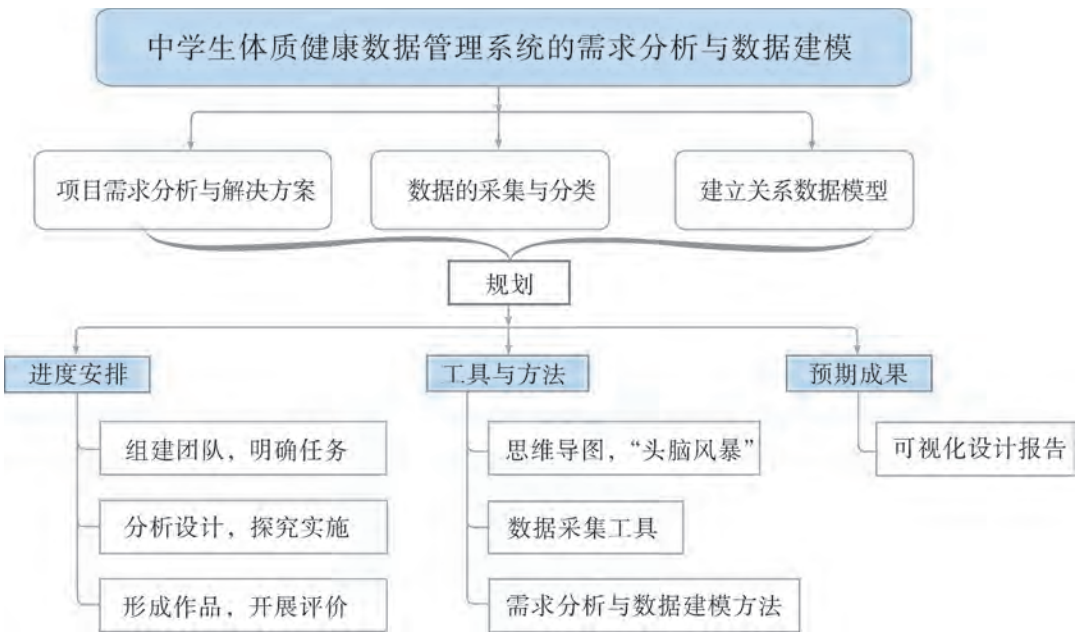


图2–1　“中学生体质健康数据管理系统的需求分析与数据建模”项目学习规划

探究

根据项目学习规划的安排，通过调查、案例分析、文献阅读和网上资料搜索，开展“中学生体质健康数据管理系统的需求分析与数据建模”项目学习探究活动，如表2-1所示。

表2-1　“中学生体质健康数据管理系统的需求分析与数据建模”项目学习探究活动

探究活动	学习内容		知识技能
项目需求分析与解决方案	项目需求分析。	分别作为系统的使用者和设计者了解业务需求。 尽可能列举与项目有关的问题。	初步了解分析业务需求的过程。
	项目解决方案。	项目解决方案的设计。 项目解决方案的评价。 项目解决方案的优化。	初步了解建立整体解决方案的过程。 尝试对方案进行分析、评价并优化。
数据的采集与分类	数据的采集。	注重元数据采集中的标准化和数据采集的途径。 采集项目数据。	了解数据采集途径的多样性和规范性。 能依据不同规划，利用适当的工具对数据进行采集和分类。 认识噪声数据现象及其成因。 理解不同结构化程度数据的区别以及在管理和应用上的特点。
	数据的分类。	数据分类的方法。 项目数据分类。	
建立关系数据模型	概念模型与E-R方法。	实体、属性、实体之间的联系相关概念。 建立实体—关系（E-R）概念模型的方法。 建立项目E-R模型。	初步了解概念模型和E-R方法。 初步掌握设计较简单数据库的逻辑结构的方法。 对采集到的数据建立关系数据模型。
	从概念模型到关系数据模型的转换。	机器世界的有关术语。 概念模型转换为关系数据模型的方法。	

实施

实施项目学习各项探究活动，进一步认识中学生体质健康数据管理系统的需求分析与数据建模。

成果

在小组开展项目范例学习过程中，利用思维导图工具梳理小组成员在“头脑风暴”活动中的观点，建立观点结构图，运用多媒体创作工具（如演示文稿、在线编辑工具等），综合加工和表达，形成项目范例可视化学习成果，并通过各种分享平台发布，共享创造、分享快乐。例如，运用在线编辑工具制作的“中学生体质健康数据管理系统的需求分析与数据建模”可视化报告，可以在教科书的配套学习资源包中查看，其目录截图如图2–2所示。

图2–2　“中学生体质健康数据管理系统的需求分析与数据建模”可视化报告的目录截图

评价

根据教科书附录2的“项目活动评价表”，对项目范例的学习过程和学习成果在小组或班级上进行交流，开展项目学习活动评价。

项目选题

同学们以3～6人组成一个小组，选择下面一个参考主题，或者自拟一个感兴趣的主题，开展项目学习。

1. 校运会管理系统的需求分析与数据建模
2. 图书馆图书借阅管理系统的需求分析与数据建模
3. 早餐营养搭配管理系统的需求分析与数据建模

各小组根据项目选题，参照项目范例的样式，利用思维导图工具，制订相应的项目方案。

各小组将完成的方案在全班进行展示交流，师生共同探讨、完善相应的项目方案。

2.1 项目需求分析与解决方案

要设计与实现一个数据管理系统，我们首先要了解用户想要这个系统做什么，用户希望这个系统实现什么功能或者解决什么问题，这就需要对用户的业务需求进行分析，只有确定了需求之后才能提出相应的解决方案。

2.1.1 项目需求分析

所谓需求分析，是指对用户的业务活动进行分析，也指对要解决的问题进行详细分析，弄清楚问题的要求，包括需要输入什么数据，要得到什么结果，最后应输出什么。在软件工程中，需求分析指的是在建立一个新的或改变一个现存的电脑系统时描写新系统的目的、范围、定义和功能时所要做的所有工作。在项目需求分析阶段，设计者通过和系统用户交流，了解业务需求，获得用户的示例表单、报表、查询、更新等活动的说明，明确系统的设计与实施目的，并编写需求分析说明文档，然后进行可行性论证，制订设计与实施的计划，这是由设计者和用户共同完成的一个过程。理解需求是在问题与其最终解决方案之间架设桥梁的第一步。

以小组为单位，跟你的同学一起通过角色替换，分别扮演设计者和用户：

（1）如果你是用户，你希望“中学生体质健康数据管理系统”应该具备哪些功能？

（2）如果你是设计者，你需要了解哪些事情？

将你们讨论的内容和得到的结论记录下来。

从用户的角度分析：

从设计者的角度分析：

需求分析，简单地说就是分析用户的具体实际需求，是设计数据库的基本和起点。需求分析的结果是否准确地反映了用户的实际需求将直接影响到后面阶段的设计。需求分析的任务是通过详细调查现实业务要处理的对象，通过对原系统工作情况的充分了解，明确不同角色的用户对功能及管理者的需求，在此基础上确定新系统的功能。对于开发一个数据管理系统项目而言，在项目需求分析阶段最有效的做法通常就是在与用户交流的时候提出一些开放性的问题，尽可能全面了解用户的业务需求。项目需求分析最重要的目标是弄清楚该系统究竟要“做什么”。

例如，对于设计一个“学生成绩管理系统”而言，通过设计者和用户的调研协商，可以确定如下需求信息：

（1）角色分类：管理员、教师、学生。

（2）管理员功能设计如下：

①班级、教师、学生信息的添加、修改、删除与查询（包括账户密码的修改）。

②课程基本信息的添加、修改、删除与查询。

③授课计划（教师授课信息）的添加、修改与删除。

（3）教师功能设计如下：

①修改个人基本资料和密码等。

②查看课程情况。

③查看、打印或导出学生名单。

④填写或修改学生成绩。

⑤查看和导出学生成绩。

⑥统计学生成绩（如成绩排序、计算平均分、计算最高分、计算合格率等）。

（4）学生功能设计如下：

①账户密码修改。

②查看个人成绩。

③查看课程开设信息。

④选课。

以小组为单位，就“中学生体质健康数据管理系统的需求分析与数据建模”的设计，分析并尽可能把可以提出的问题列举出来。

针对“中学生体质健康数据管理系统的需求分析与数据建模”的设计，我们可以考虑提出如下问题：一般有哪些人需要用到该数据管理系统？访问这些人，了解他们希望该系统实现何功能。目前我国中学生体质健康一般包括哪些数据信息？这些数据信息有何特征？通过这些数据如何反映出学生的体质健康情况？……通过对这些问题的回答，有助于我们系统地分析业务需求。

在本项目中，我们知道使用的用户主要包括学校老师、相关上级行政部门和学生。在通过与学校老师和有关部门负责人沟通了解后，得知中学生体质健康标准登记的一般流程为：学生在学校参加一系列的体质测试；学校老师负责把学生的信息及测试成绩录入系统，并根据成绩及相应的规则对学生体质健康情况进行分析和处理；学生可以通过查询及时了解自己的体质健康情况；上级部门可根据各校上报的学生成绩，对学生的体质健康情况进行跟踪和分析，以便给出正确的指导意见。图2-3为用于记录学生体质健康数据的“国家学生体质健康标准登记表”。

<table>
<tr><td>姓名</td><td colspan="3"></td><td>性别</td><td colspan="3"></td><td>民　族</td><td colspan="4"></td></tr>
<tr><td>班号</td><td colspan="3"></td><td>学号</td><td colspan="3"></td><td>出生日期</td><td colspan="4"></td></tr>
<tr><td colspan="4">一年级</td><td colspan="4">二年级</td><td colspan="4">三年级</td><td colspan="2">毕业成绩</td></tr>
<tr><td>指标（项目）</td><td>成绩</td><td>得分</td><td>等级</td><td>指标（项目）</td><td>成绩</td><td>得分</td><td>等级</td><td>指标（项目）</td><td>成绩</td><td>得分</td><td>等级</td><td>得分</td><td>等级</td></tr>
<tr><td>身高标准体重</td><td></td><td></td><td></td><td>身高标准体重</td><td></td><td></td><td></td><td>身高标准体重</td><td></td><td></td><td></td><td rowspan="6"></td><td rowspan="6"></td></tr>
<tr><td>肺活量
体重指数</td><td></td><td></td><td></td><td>肺活量
体重指数</td><td></td><td></td><td></td><td>肺活量
体重指数</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>奖励得分</td><td colspan="3"></td><td colspan="4"></td><td colspan="4"></td></tr>
<tr><td>学年总分</td><td colspan="3"></td><td colspan="4"></td><td colspan="4"></td></tr>
<tr><td>等级评定</td><td colspan="3"></td><td colspan="4"></td><td colspan="4"></td></tr>
<tr><td>体育教师签字</td><td colspan="3"></td><td colspan="4"></td><td colspan="4"></td></tr>
<tr><td>班主任签字</td><td colspan="3"></td><td colspan="4"></td><td colspan="4"></td></tr>
</table>

图2-3　国家学生体质健康标准登记表

通过与用户的访谈交流和实地调查分析，我们可以初步得知“中学生体质健康数据管理系统的需求分析与数据建模”设计与实施的目的主要包括以下几个方面的功能需求：

（1）数据录入：如录入学校信息、学生个人信息、学生测试成绩等。

（2）数据查询：如查询学生个人信息、学生体质健康指标、学校学生体质健康指标等。

（3）数据修改：如修改学校信息、学生个人信息、学生测试成绩等。

（4）统计分析报表输出：如统计学校等级信息、男女生体形差异、个人综合评分情况等。

2.1.2 项目解决方案

项目解决方案的重点是分析现存的问题，提出新系统的功能需求及相应的技术实现手段和实施保障的措施，说明用户需求是可以实现的。解决方案是系统开发人员在与用户充分交流的基础上结合自己的专业知识而提出的。

1. 项目解决方案的设计

项目解决方案的目的就是为了让用户了解该项目是可行的。作为设计者，一般会从“为什么做”“做什么”“达到什么效果”“怎么做”“如何保障质量”等方面考虑。

以小组为单位，尝试从回答“为什么做”“做什么”“达到什么效果”“怎么做”“如何保障质量”等问题入手，讨论并设计本小组的项目解决方案。

小组项目解决方案

（1）现状与问题

（2）功能需求

（3）技术实现手段

（4）保障措施

项目解决方案是系统设计者在充分了解用户的业务需求基础上提出的。解决方案的基本结构一般包括以下五个部分：现状分析与诊断、系统规划与设计、系统技术方案、系统实施方案、保障措施。

（1）现状分析与诊断。

一般从本项目所涉及的业务现状描述入手，分析当前存在的问题，并提出改进的建议，得出实施项目系统的必要性，以及需要解决的问题等，即回答了“为什么做”的问题。

（2）系统规划与设计。

根据现状分析提出的需求，从总体目标、指导思想、总体框架等方面对本项目系统进行总体规划与设计，也就是回答了“做什么”的问题。

（3）系统技术方案。

从基本功能介绍、关键问题解决方案两个层面介绍具体的技术方案。基本功能介绍是对本项目所涉及的系统产品，在标准功能基础上适当补充新增功能或用户特殊需求的功

能。关键问题解决方案是就用户特别关心的问题、用户特殊需求中有一定难度的问题等提出解决方案和建议。这一步其实就是回答了“达到什么效果”的问题。通常情况下，这一步在之前的项目需求分析阶段已经给出结果。

（4）系统实施方案。

从本项目的预期效益入手，首先分析项目实施存在的风险，接着介绍规避风险的实施保障措施，最后给出初步实施进度计划。实施规划要结合用户的实施打算，如果系统规模比较大，可以结合用户的需求适当进行目标分解，分期完成，本步骤回答了“怎么做”的问题。

（5）保障措施。

从能为用户提供的全方位服务承诺入手，阐述技术支持与服务的保障措施，让用户无后顾之忧，这就回答了“如何保障质量”的问题。

在“中学生体质健康数据管理系统的需求分析与数据建模”项目范例中，现状分析与诊断就是，由于学生的数据信息量大，传统的纸张记录方式不利于学校老师和相关部门对学生体质健康情况的跟踪和分析，有必要使用数据库系统对这些数据进行集中管理和分析。通过之前的项目需求分析，我们初步可以确定系统的基本功能包括数据的录入、查询、修改及统计和分析报表的输出等。为了操作直观和方便，我们计划选用开源软件MariaDB作为数据库管理实现的软件基础，根据数据分析阶段的需要，我们还会结合Python，SPSS，SAS等软件完成中学生体质健康数据的分析，由此得到本范例初步的项目解决方案。

中学生体质健康数据管理系统的需求分析与数据建模 项目解决方案

（一）现状分析与诊断

本项目中，使用的用户主要包括学校老师、相关上级行政部门和学生。中学生体质健康标准登记的一般流程为：学生在学校参加一系列的体质测试；学校老师负责把学生的信息及测试成绩录入系统，并根据成绩及相应的规则对学生体质健康情况进行分析和处理；学生可以通过查询及时了解自己的体质健康情况；上级部门可根据各校上报的学生成绩，对学生的体质健康情况进行跟踪和分析，以便给出正确的指导意见。传统的手工记录方式不利于及时对数据进行分析，为此需要建立一个计算机数据管理系统有效地管理学生体质健康数据。

（二）系统规划与设计

通过建立“中学生体质健康数据管理系统”数据库管理学生体质健康测试的相关数据，并能利用这些数据实现相关的查询和统计分析等功能。

（三）系统技术方案（系统基本功能）

1. 数据录入：例如录入学校信息、学生个人信息、学生测试成绩等。

2. 数据查询：例如查询学生个人信息、学生体质健康指标、学校学生体质健康指标等。

3. 数据修改：例如修改学校信息、学生个人信息、学生测试成绩等。

4. 统计分析报表输出：例如统计学校等级信息、男女生体形差异、个人综合评分情况等。

（四）系统实施方案

采用小组合作方式，通过本课程的学习，利用MariaDB数据库管理软件建立和管理“中学生体质健康数据管理系统”数据库，根据需求使用适当的数据分析软件（如Python，SPSS，SAS等）、适当的分析方法对数据库数据进行统计和分析，并根据统计分析结果撰写数据分析报告。

（五）保障措施

作为系统的设计者和开发者，需要为用户整理出详细的使用手册，在一定的时间段内密切留意用户使用系统后的反馈意见并及时作出调整和更新。

2. 项目解决方案的评价

目前，关于项目解决方案的评价还没有唯一的标准。一般来说，可以从以下方面进行评价：

（1）是否能够透视现存问题并提出有针对性的解决措施。

（2）是否针对本项目业务的特点和流程设计。

（3）能否满足基本需求、关键需求和未来变化的需要。

以小组为单位，讨论作为一个数据管理系统，除了以上项目解决方案中提到的几部分内容外，还应该考虑哪些因素，把讨论得到的结果列举出来。

与其他小组分享交流本小组的数据管理系统项目的需求分析及解决方案，各小组从第三方的角度尝试对该方案进行评价并给出改进的建议。

3. 项目解决方案的优化

优化项目解决方案，是指让所设计的项目解决方案更加有针对性，更能满足需求和未来变化的需要。

对项目解决方案进行优化，通常可以采取以下方法：

（1）重做需求分析，确认现存问题，重新提出有针对性的解决措施。

（2）重新梳理项目业务的特点和流程，根据特点和流程进行二次设计。

（3）检查项目基本需求、关键需求和未来变化的需要，改进解决方案。

拓展

几种常用数据库管理软件简介

1. 关系数据库。

（1）Oracle Database，简称Oracle，是以分布式数据库为核心的关系数据库管理系统，系统可移植性好、使用方便、功能强。

（2）SQL Server数据库是一款RMDBS数据库。SQL Server的优点为：①真正的客户服务器体系结构；②图形化用户界面，更加直观、简单；③丰富的编程接口工具，为用户进行程序设计提供更多的选择；④具有很好的伸缩性，可跨界运行；⑤对Web技术的支持，使用户能够容易地将数据库中的数据发布到Web上；⑥提供数据仓库功能。

（3）Microsoft Office Access是把数据库引擎的图形用户界面和软件开发工具结合在一起的数据库管理系统，其优势为：①存储方式单一，便于用户操作和管理；②界面友好、易操作；③集成环境、处理多种数据信息；④支持ODBC。

（4）PostgreSQL是一个开源数据库系统，可以运行在所有主流操作系统上，包括Linux、Unix和Windows。PostgreSQL是完全的事务安全性数据库，完整地支持外键、联合、视图、触发器和存储过程。PostgreSQL对很多高级开发语言均有编程接口。

2. 非关系数据库。

（1）Apache Hbase是一个分布式的、面向列的开源数据库。Hbase是Hadoop项目的子项目，它利用Hadoop MapReduce来处理Hbase中的海量数据。Hbase是一个适于非结构化数据存储的数据库。从技术上看Hbase更像是“Data store”多于“Data base”，它是一种“NoSQL”数据库。

（2）Redis是一个开源的数据结构存储系统，它可以用作数据库、缓存和消息中间件，支持多种类型的数据结构。Redis内置了复制、LUA脚本、LRU驱动事件、事务和不同级别的磁盘持久化，并通过Redis哨兵和自动分区（Cluster）提供高可用性。

（3）MongoDB是一个基于分布式文件存储的数据库，旨在为WEB应用提供可扩展的高性能数据存储解决方案。MongoDB支持的查询语言非常强大，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。

（4）图形数据库是“NoSQL”数据库的一种类型，它应用图形理论存储实体之间的关系信息。相对于关系数据库中的各种关联表，图形数据库中的关系可以通过关系能够包含属性这一功能来提供更为丰富的关系展现方式。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，完成相应的项目需求分析与解决方案。

1. 分析项目需求。
2. 讨论项目解决方案。
3. 交流项目解决方案并进行优化。

2.2 数据的采集与分类

完成了项目需求分析及提出项目解决方案后，下一个环节就要进入数据的采集与分类阶段了。

2.2.1 数据采集的途径

在数据管理系统设计中，了解现实系统的运作过程，有必要采集各种原始凭证，并弄清数据的来龙去脉等。

探究活动

讨论

以小组为单位，讨论并列举你所知道的数据采集的途径，尝试为本小组项目采集相关数据信息。

常用数据采集的途径：

数据采集的途径多种多样。在数据管理系统设计中，根据解决问题的需要，开发人员通常会使用以下几种数据采集的技术：

1. 分析文档资料

分析文档资料有助于了解一些内部信息，比如对数据库的需求是如何提出的、需要记录的数据信息类型等。例如，要为某公司设计一个小型财务数据库系统，设计者必须先熟悉该公司的财务业务流程，其中一个有效的方法就是调用该公司的财务报表文档进行分析。

2. 面谈

面谈，通过与人面对面交流来采集信息，是比较常用的一种技术。面谈需要良好的沟通技能，访问者提前准备一系列明确的问题对被访问者提问，而且要选择合适的被访问者才能保证面谈更有效。例如，国家人口普查就是通过入户面谈的方式进行数据采集。

3. 实地调查

实地调查是了解一个系统运作的最有效的技术。成功的实地调查需要调查者进行精心

准备，尽可能多地了解相关的人或业务活动。例如，为某小型超市设计一个超市管理系统，最有效的数据采集方法就是到超市实地调查，全面了解该超市的实际经营流程。

4. 研究

对应用或问题本身进行详细研究也是一种有用的数据采集方法。相关的期刊、参考书籍和互联网都是很好的信息资源，可以提供解决类似问题的方法。

5. 问卷调查

通过问卷的方式进行调查也是一种常用的数据采集技术。利用问卷调查，可以从大量的人群中采集数据信息。问卷调查表里可以包括两类问题，即自由格式的问题和固定格式的问题。自由格式的问题为回答者在作答时提供了较大的自由度，但是被调查者的答案难以列表统计。固定格式的问题需要明确作答，对于每一个问题，被调查者都要从给出的答案中选择，被调查者的答案容易列表统计。例如，若要设计一个“早餐营养管理系统”，我们就可以通过问卷调查的方式采集大量的数据。

在“中学生体质健康数据管理系统的需求分析与数据建模”项目中，我们可以通过分析“国家学生体质健康标准登记表”了解到系统需要记录的数据信息。我们还可以通过与学校老师和相关部门负责人面谈，以及到学校实地调查等方法，了解学校和相关部门对中学生体质健康数据记录和统计分析的一般流程，并采集第一手数据信息。我们还根据《国家学生体质健康标准（2014年修订）》了解到学生的体质健康水平是由学生的身体形态、身体机能和身体素质等方面综合评定的。根据年龄段，初、高中学生分别分为6组进行测试评定。在所有测试指标中，身体形态类中的身高、体重，身体机能类中的肺活量，以及身体素质类中的50米跑、坐位体前屈为各年级学生共性指标。同时，标准还给出了详细的总分计算标准以及评定等级标准，如图2-4所示。

姓名		性别		民　族	
班号		学号		出生日期	

一年级				二年级				三年级			
指标（项目）	成绩	得分	等级	指标（项目）	成绩	得分	等级	指标（项目）	成绩	得分	等级
身高标准体重				身高标准体重				身高标准体重			
肺活量体重指数				肺活量体重指数				肺活量体重指数			
奖励得分											
学年总分											
等级评定											
体育教师签字											
班主任签字											

国家学生体质健康标准（2014年修订）

一、说明

1.《国家学生体质健康标准》（以下简称《标准》）是国家学校教育工作的基础性指导文件和教育质量基本标准，是评价学生综合素质、评估学校工作和衡量各地教育发展的重要依据，是《国家体育锻炼标准》在学校的具体实施，适用于全日制普通小学、初中、普通高中、中等职业学校、普通高等学校的学生。

男生体重指数（BMI）单项评分表（单位：千克/米^2）

等级	单项得分	一年级	二年级	三年级	四年级	五年级	六年级	初一	初二	初三	高一	高二	高三	大学
正常	100	13.5~18.1	13.7~18.4	13.9~19.4	14.2~20.1	14.4~21.4	14.7~21.8	15.5~22.1	15.7~22.5	15.8~22.8	16.5~23.2	16.8~23.7	17.3~23.8	17.9~23.9
低体重	80	≤13.4	≤13.6	≤13.8	≤14.1	≤14.3	≤14.6	≤15.4	≤15.6	≤15.7	≤16.4	≤16.7	≤17.2	≤17.8
超重	80	18.2~20.3	18.5~20.4	19.5~22.1	20.2~22.6	21.5~24.1	21.9~24.5	22.2~24.9	22.6~25.2	22.9~26.0	23.3~26.3	23.8~26.5	23.9~27.3	24.0~27.9
肥胖	60	≥20.4	≥20.5	≥22.2	≥22.7	≥24.2	≥24.6	≥25.0	≥25.3	≥26.1	≥26.4	≥26.6	≥27.4	≥28.0

图2-4　“中学生体质健康数据管理系统的需求分析与数据建模”数据采集

2.2.2 数据的分类

一般来说，开始采集到的数据都是比较凌乱的，有些数据可能很关键，有些数据却无关紧要。到底哪些数据资料是我们建立数据管理系统所关心和必需的呢？这就需要根据项目的需要，对采集到的各种原始数据进行分类整理，提取有用的信息。

在“中学生体质健康数据管理系统的需求分析与数据建模”数据采集中，某学校采集的一组学生身高的数据如下（单位：cm）：165，174，175，157，15，163，173，121，166，174，355，163，185，285，85。以小组为单位，观察这组数据是否合理，并讨论如果使用了这组数据进行后续的数据分析，会有什么影响。

1. 噪声数据现象及其成因

噪声数据（Noisy data），就是无意义的数据，就是被测量的变量的随机误差或方差，是指数据中存在着错误或异常（偏离期望值）的数据。

引起噪声数据的原因有很多，比如可能是硬件故障、编程错误、语音或光学字符识别程序（OCR）中的乱码等。拼写错误、行业简称以及俚语也会阻碍机器读取，从而引起噪声数据。

噪声数据可能会影响后面数据分析的结果。因此，噪声数据处理是数据处理的一个重要环节。

2. 分类数据

以小组为单位，交流本小组采集到的数据信息是否有用，尝试对本小组项目数据信息进行分类整理，提取有用信息。

采集到的数据有不同的类别。有些数据是有结构的，可以方便地用二维表结构来表示，如数字、符号等，称为结构化数据；有些数据却不方便用二维表来表现，如所有格式的办公文档、文本、图片、图像、音频信息和视频信息等，称为非结构化数据。所谓半结构化数据，就是介于完全结构化数据和完全非结构化数据（如声音、图像文件等）之间的数据，如HTML文档就属于半结构化数据。

对于不同结构的数据，管理和调用的方式是不同的。

（1）结构化数据，是带有表头的表结构数据，数据按行和列组织，其中第一行给出列的名字，每一列代表一个不同的事实或度量，每行表示一个已知事实集合的实例或数据。大多数公共数据都是这种格式。

（2）非结构化数据，没有具体的数据模型，如各种文档、图片、音频、视频，通常可以建立一个包含“编号”“内容描述”和“内容（指向）”的表来实现与“数据”的对应。

（3）半结构化数据，数据不总是以直接可用的格式存在，这个数据以没有表头的表格形式存储，其中的值是使用了难以理解的编码，需要使用此数据附带说明文档才能解码。

在“中学生体质健康数据管理系统的需求分析与数据建模”案例中，登记表中的学生姓名、性别、出生日期、民族等个人信息是对管理和分析有用的，需要保留；学生测试的成绩也是有用的；但是体育老师签字和班主任签字等信息与分析学生体质健康没有很大联系。由于《国家学生体质健康标准（2014年修订）》不但对中学生体质健康标准给出了定义，还给出了小学生的体质健康标准的定义，所以我们只要选择与中学生标准相关的数据信息保留即可。

为了管理的方便，防止学生重名带来的混淆，我们增加了学生“学籍号”，并为每所学校和测试项目设计了“学校编号”和“项目编号”。

经过分析，最后我们可以将构建“中学生体质健康管理系统”项目所需的基本数据分为三类，均为结构化数据。

中学生体质健康数据管理系统的需求分析与数据建模 基本数据信息

（1）关于学生的信息：学籍号、姓名、性别、学校名称、年级、班别、出生日期、民族、总得分、等级。

（2）关于学校的信息：学校名称、地址、联系电话、电子邮箱。

（3）关于测试指标项目的信息：项目编号、项目名称、项目单位、项目权重。

大数据环境下的数据采集和分类

在大数据环境下还可以简单地采用以上介绍的传统数据采集和分类方法吗？显然不行。大数据环境下，数据来源非常丰富而且形式多样，大数据要处理的往往是大量的非结构化数据。大数据环境下，数据采集和分类一般包括以下方法：

（1）系统日志采集方法。

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，Cloudera的Flume，Facebook的Scribe等，这些工具均采用分布式架构，能满足每秒数百MB的日志数据采集和传输需求。

（2）网络数据采集方法：对非结构化数据的采集。

网络数据采集是指通过网络爬虫或网站公开API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，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用DPI（Deep Packet Inspection，深度包检测）或DFI（Deep/Dynamic Flow Inspection，深度/动态流检测）等带宽管理技术进行处理。

（3）其他数据采集方法。

对于企业生产经营数据或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相关方式采集数据。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，完成相应的项目数据采集和分类。

1. 选用适当的工具进行数据采集。
2. 选用适当的方法进行分类处理。

2.3 建立关系数据模型

我们知道，关系数据库是利用二维数据表来组织和存储数据以及数据之间的联系的，对应存储的是结构化数据。对于一个简单的问题，经验丰富的设计人员可能很快就能设计出其数据库的结构，但是对于一个复杂的问题，则往往难以直接设计出来。因此，具体的做法是，人们在数据库设计的过程中，首先从用户的观点建立对于现实世界数据现象的概念模型，然后再把概念模型转换为某一数据管理系统支持的数据模型，这个过程也称为数据的抽象过程，如图2-5所示。

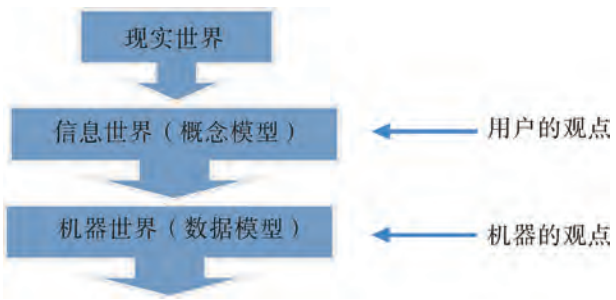


图2-5　数据的抽象过程

2.3.1 概念模型与E-R方法

概念模型是从现实世界到信息世界的第一层抽象。信息世界是现实世界在人们头脑中的反映，人的思维将现实世界的数据抽象化和概念化，并用文字符号表示出来，就形成了信息世界。

人们在研究信息世界的过程中，常常用到以下术语：

1. 实体

客观存在且可以互相区别的事物，称为实体。如一名学生、一台电脑、一本书、一场聚会。实体是信息世界的基本单位，它与现实世界中客观存在的事物相对应。

我们把拥有相同属性的实体称为同类实体，同类实体的集合称为实体集。

2. 属性

实体的特征称为属性。一个实体可以有多个特征，如姓名、性别、所在学校等都是学生的基本属性。

3. 键

能在一个实体集中唯一标识一个实体的属性称为键。键可以只包含一个属性，也可以包含多个属性。

4. 联系

在现实世界中，事物内部以及事物之间是有联系的，这些联系在信息世界中反映为实体内部的联系和实体之间的联系。实体内部的联系通常是指组成实体的各属性之间的联系，实体之间的联系通常是指不同实体集之间的联系。

实体之间的联系有三种：一对一联系、一对多联系、多对多联系。

例如，一所学校只有一个正校长，同时一个正校长只担任一所学校的正校长职务，则学校与正校长之间具有一对一联系；一个班级有若干名学生，而每个学生只在一个班级中学习，则班级和学生之间具有一对多联系；在运动会上，一个运动员可以参加多个比赛项目，一个比赛项目也可以有多个运动员参加，则运动员与比赛项目之间具有多对多联系。

探究活动

讨论

以小组为单位，讨论本小组项目中包括哪些实体。它们分别是什么？各具有哪些属性？实体之间的联系是什么？把讨论、分析的结果记录下来。

小组项目“实体—属性”记录

概念模型是数据库设计人员进行数据库设计的有力工具，也是数据库设计人员和用户之间进行交流的语言，因此概念模型应满足以下三个方面的要求：

（1）能比较真实地模拟现实世界，具有较强的表达能力，能够方便、直接地表达应用中的各种要求。

（2）简单、清晰，容易被人理解。

（3）要便于在计算机上实现。

概念模型的表示方法很多，其中最为著名、最为常用的是于1976年提出的实体—联系模型（Entity-Relationship Model），也称为实体—关系模型，简称E-R模型。

建立实体—关系模型（E-R模型）一般有四个步骤，如图2-6所示。

图2-6　建立E-R 模型的四个步骤

E-R图就是用特定的符号来描述E-R模型中实体集及实体集之间的联系，E-R图包括三个图素：

（1）实体集。用矩形框表示，框内标注实体名称。

（2）属性。用椭圆形表示，框内标注属性名称，并用连线与实体连接起来。

（3）实体之间的联系。用菱形框表示，框内标注联系名称，用连线将菱形框分别与有关实体相连，并在连线上注明联系类型，如图2-7所示。

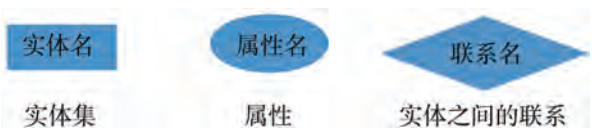


图2-7　E-R图的基本图案

可以用E-R图来表示实体集之间的三种联系，如图2-8所示。

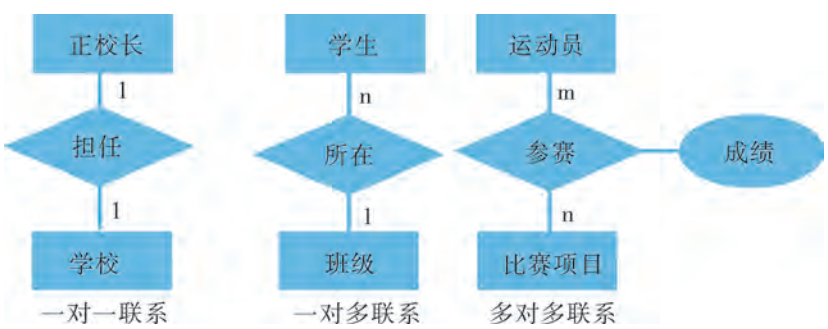


图2-8　实体集的三种联系

用E-R图表示概念模型时，人们所关心的仅仅是有哪些实体和属性，以及实体和属性之间的联系如何，而不必关心它们在计算机内是如何表示的。

实践

在“中学生体质健康数据管理系统的需求分析与数据建模”项目中，我们可以初步确定数据管理系统有三个实体，分别是学生、学校和指标项目。“学生”实体的属性有学籍号、姓名、性别、年级、班别、出生日期、民族、总得分、等级等，“学校”实体的属性有学校名称、地址、联系电话、电子邮箱等，“指标项目”实体的属性有项目编号、项目名称、项目单位、项目权重等。分析这三个实体，我们可以得到它们之间有如下联系：

（1）一所学校可以有多个学生参加测试，但一个学生却只能属于一所学校，因此它们之间应该是一对多联系。

（2）一个学生可以参加多个指标项目的测试，一个指标项目也可以由多个学生参加测试，因此它们之间应该是多对多联系。同时，作为测试结果，必定会有测试成绩，并根据标准得到该项目得分及对应等级，因此这三个属性是属于“参加”这个联系的属性。

经过以上分析和综合，可以得出“中学生体质健康数据管理系统的需求分析与数据建模”的E-R图，如图2-9所示。

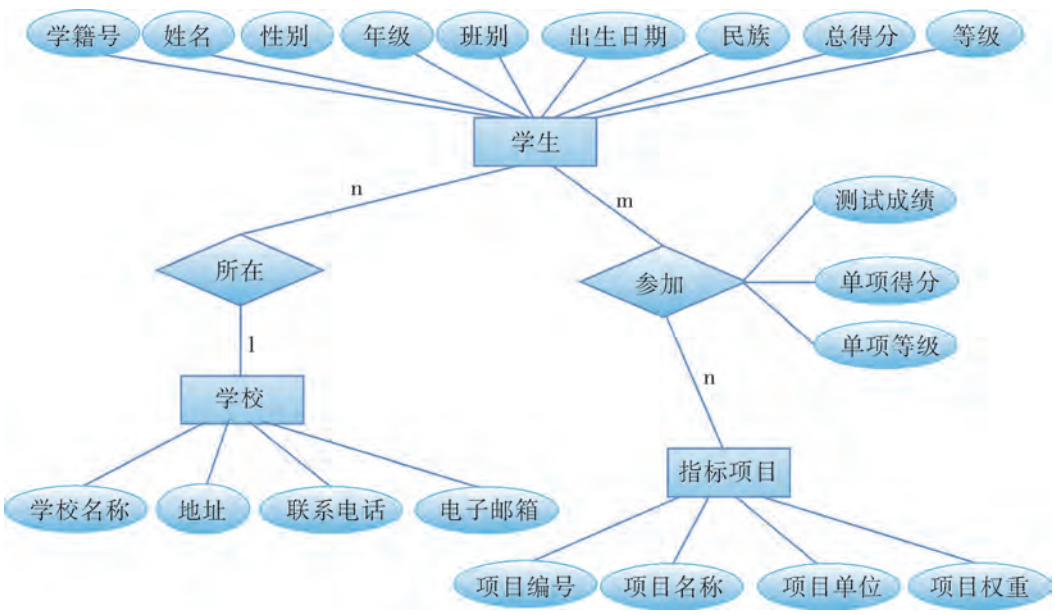


图2-9 “中学生体质健康数据管理系统的需求分析与数据建模”E-R图

2.3.2 从概念模型到关系数据模型的转换

我们已经学会如何根据采集到的数据信息建立信息世界的概念模型，并用E-R图表示出来，但这只是第一阶段的抽象过程。那么，如何将它进一步转换为机器世界中的关系数据模型呢？

机器世界又称数据世界，信息世界中的信息经过抽象和组织，以数据形式存储在计算机中，就成为机器世界。与信息世界一样，机器世界也有其用来描述数据的习惯术语，这些术语与信息世界中的术语有着对应的关系。

1. 字段

字段用来标记实体的一个属性，它是可以命名的最小信息单位。例如学生有学籍号、姓名、性别、出生日期等字段，字段与信息世界的属性相对应。

2. 记录

记录是有一定逻辑关系的字段的组合。它与信息世界中的实体相对应，一条记录可以描述一个实体。例如一个学生的记录由“学籍号、姓名、性别、出生日期”等字段组成。

3. 文件

文件是同一类记录的集合。

4. 关键字

关键字是可以唯一标识一条记录的字段，它可以是一个字段，也可以是多个字段。关键字与信息世界中的键相对应。

以小组为单位，尝试将本小组的项目概念模型转换为关系数据模型。

关系数据模型是采用二维表的形式表示实体以及实体之间的联系。将E-R模型转换为关系数据模型，一般可以分两步进行：

（1）将每个实体集转换成一个二维表。

将实体集转换成一个二维表时，实体的属性转变为二维表的字段，一个具体的实体由表中的一条记录来表示。

为了方便记录某些特殊情况，可以在每个实体中增加“备注”项。由此，我们就可以得到“中学生体质健康数据管理系统的需求分析与数据建模”项目中三个实体集的关系模型，如图2-10所示。

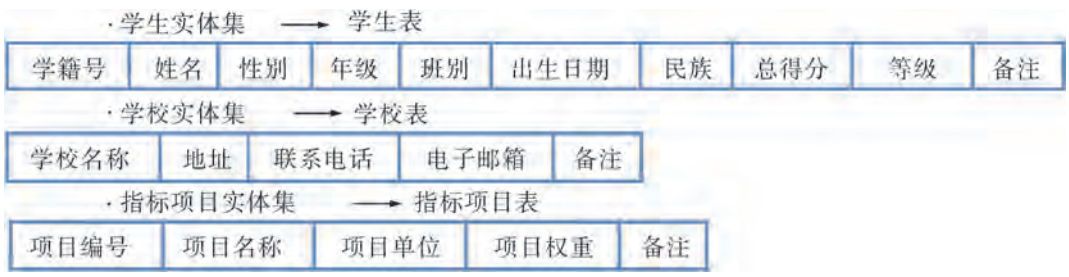


图2-10　三个实体集的关系模型

根据前面所讲的关键字的定义，我们分别确定三个表的关键字为：

“学籍号”作为“学生表”的关键字；

“学校名称”作为“学校表”的关键字；

“项目编号”作为“指标项目表”的关键字。

（2）将实体集之间的联系转换成一个二维表。

用二维表来表示实体集之间的联系，通常有以下两种方法：

方法一：定义一个新的二维表，该表除了包含联系本身的属性外，同时还包含其他实体集中的关键字属性，通过它们将这些实体集关联起来。

按照这种方法，从“中学生体质健康数据管理系统的需求分析与数据建模”项目的E-R图，我们可以得到如图2-11所示的关系数据模型。

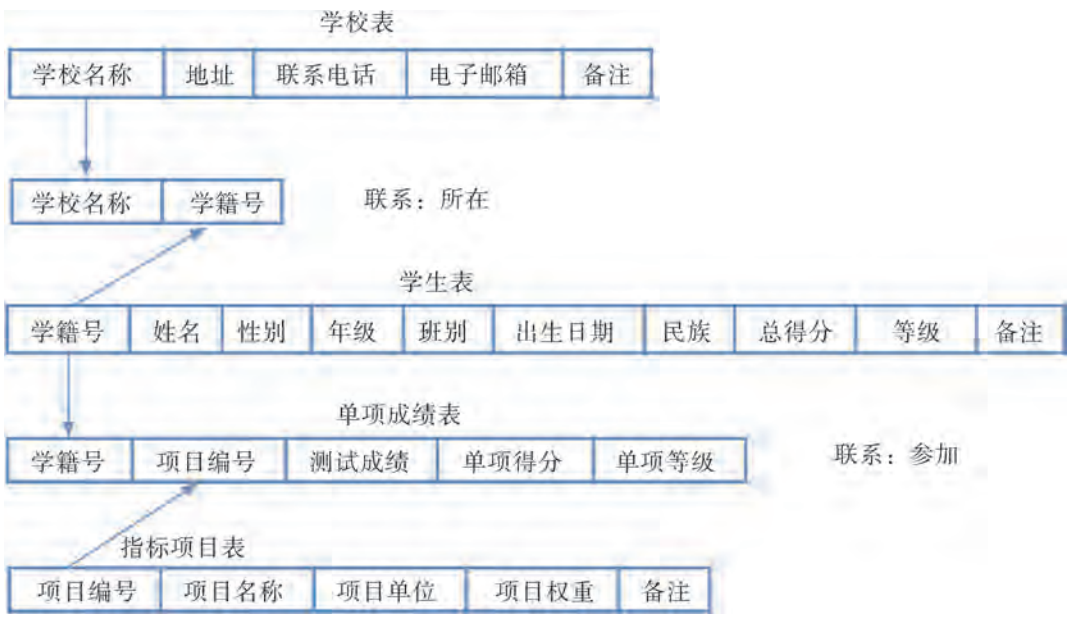


图2-11 关系数据模型

在“所在”联系表中，“学校名称”和“学籍号”分别来自“学校表”和“学生表”。

在“参加”联系表中，“学籍号”和“项目编号”分别来自“学生表”和“指标项目表”，“测试成绩”“单项得分”和“单项等级”则是联系本身的属性。

方法二：在一个表中，加入联系的属性以及另外一个表中的关键字属性，从而建立起它们之间的联系。如果我们要建立的数据库实体之间的联系比较简单，为了减少数据表的数目，则可以采用这种方式，即通过实体关系表中的字段属性建立起实体之间的联系。

在“中学生体质健康管理系统的需求分析与数据建模”项目案例中，我们可以在“学生表”中加入“学校名称”字段，就可以建立起“学生表”和“学校表”之间的联系，而不需要单独建立一个“所在”联系表。

结合以上两种方法，我们可以得到“中学生体质健康管理系统的需求分析与数据建模”项目的关系数据模型，它包含学校表、学生表、参加表和指标项目表四个表，为了便于识别，我们把“参加表”的名称改为“单项成绩表”，其表结构及关联关系如图2-12所示。

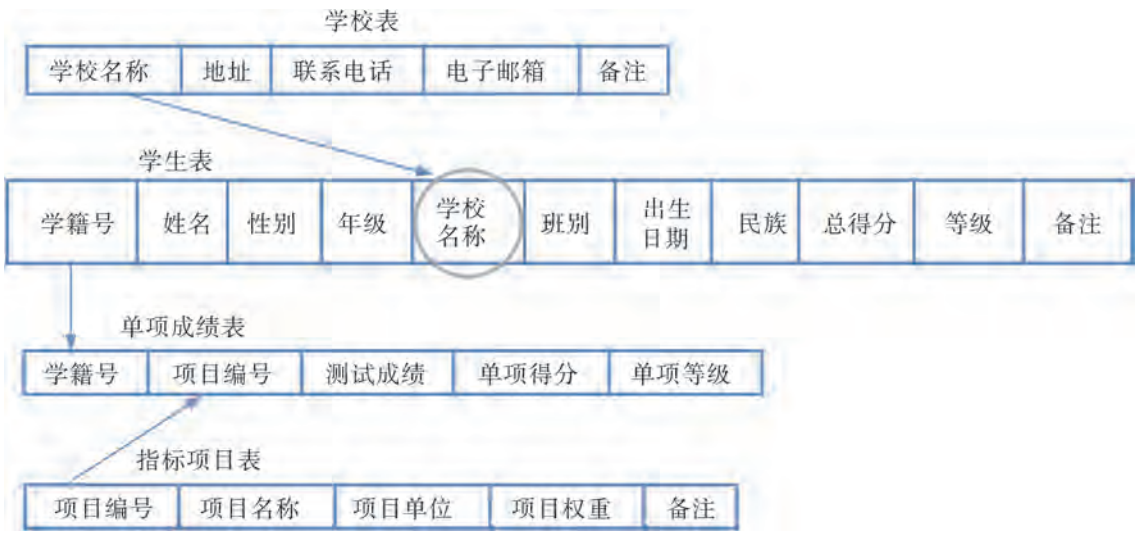


图2-12　表结构及关联关系

关系数据模型除了可以用上述的表格表示外，还可以用括号形式表示，如“学校表”可以写为：

学校表（学校名称，地址，联系电话，电子邮箱，备注）

拓展

其他数据模型简介

目前成熟地应用在数据库系统中的数据模型除了关系模型外，还有层次模型（Hierarchical Model）和网状模型（Network Model）。不同于关系模型是用“二维表”（或称为“关系”）来表示数据之间的联系，层次模型以“树结构”表示数据之间的联系，网状模型是以“图结构”来表示数据之间的联系。

1. 层次模型。

层次模型是数据库系统最早使用的一种模型，它的数据结构是一棵“有向树”。根结点在最上端，层次最高，子结点在下，逐层排列。层次模型的特征是：

（1）有且仅有一个结点没有父结点，本节点就是树的根，称为“根结点”。

（2）其他结点有且仅有一个父结点。

构成层次模型的树是由结点和连线组成的，结点表示实体集，连线表示相连两个实体之间的联系，这种联系只能是“一对多”的（“一对一”是“一对多”的特例）。通常把表示“一”的实体放在上方，作为父结点；把表示“多”的实体放在下方，作为子结点，如图2-13所示。

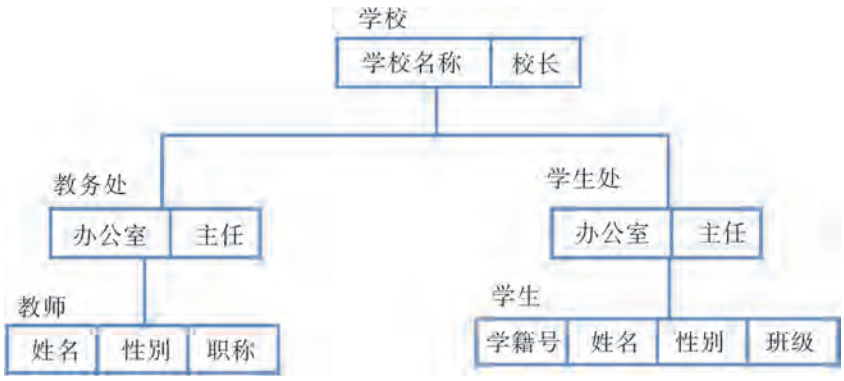


图2-13 学校行政管理的层次模型

最有影响的层次模型的数据库系统是20世纪60年代末由IBM公司推出的IMS（Information Management System）层次模型数据库系统。

2. 网状模型。

网状模型以网状结构表示实体与实体之间的联系。网中的每一个结点代表一个记录类型，联系用链接指针来实现。网状模型可以表示多个从属关系的联系，也可以表示数据间的交叉关系，即数据间的横向关系与纵向关系，它是层次模型的扩展。网状模型可以方便地表示各种类型的联系，但结构复杂，实现的算法难以规范化。其特征是：

（1）允许结点有多于一个父结点。

（2）可以有一个以上的结点没有父结点。

例如，某医院每个医生负责治疗三个病人，不同医生负责治疗的三个病人可住同一病房。它们构成了一个网状模型，如图2-14所示。

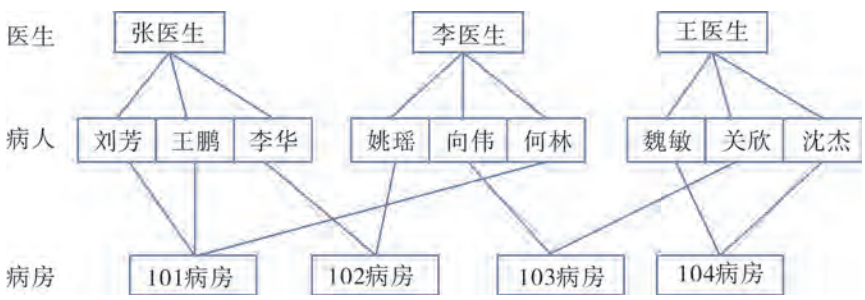


图2-14 某医院管理数据库的网状模型

一个网状模型可以理解为多个层次模型的集合，所以网状模型和层次模型本质上是一样的。从逻辑上看，它们都是基本层次关系的集合，用结点表示实体，用连线表示实体间的关系；从物理上看，它们每一个结点都是一个存储记录，用链接指针来实现记录之间的关系。当存储数据时，由于这些链接指针已经固定下来了，那么就导致数据检索时必须考虑存储路径问题；当更新数据时，涉及链接指针的调整，缺乏灵活性，系统扩张麻烦。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，建立项目实体—关系（E-R）模型并转换为相应的关系数据模型，参照项目范例的样式，撰写相应的项目成果报告。

成果交流

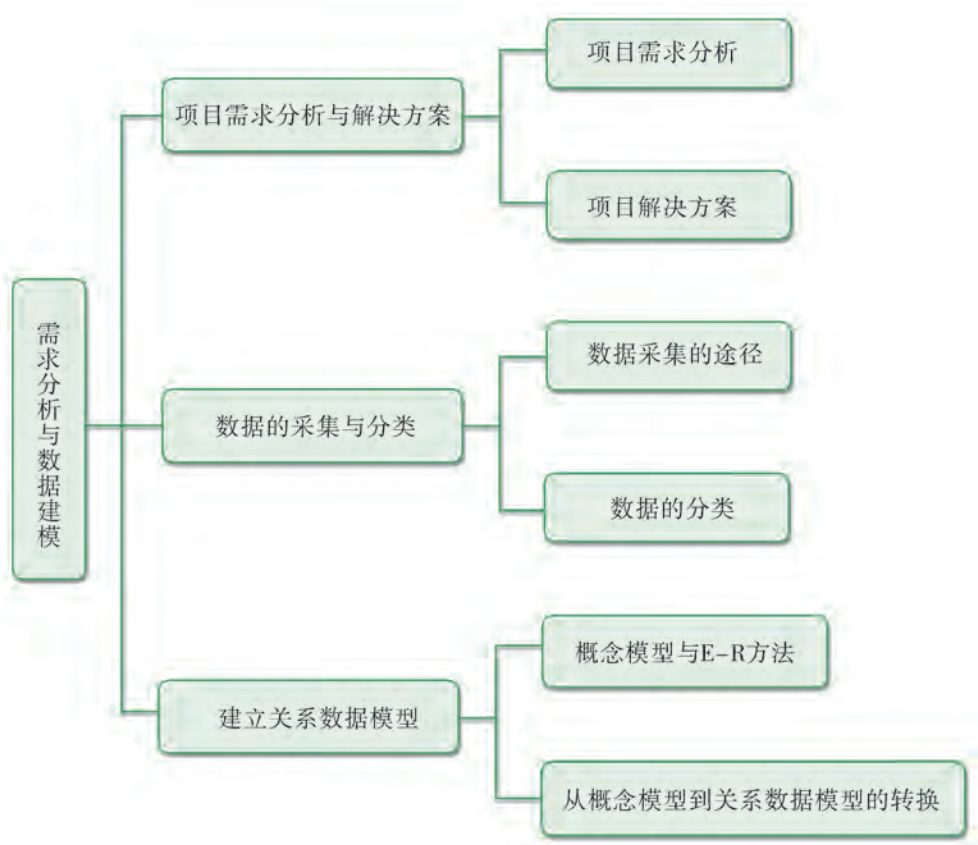
各小组运用数字化学习工具，将所完成的项目成果，在小组或班级上进行展示与交流，共享创造、分享快乐。

活动评价

各小组根据项目选题、拟订的项目方案、实施情况以及所形成的项目成果，利用教科书附录2的“项目活动评价表”，开展项目学习活动评价。

本章扼要回顾

同学们通过本章学习，根据“需求分析与数据建模”知识结构图，扼要回顾、总结、归纳学过的内容，建立自己的知识结构体系。



回顾与总结

本章学业评价

同学们完成下列测试题（更多的测试题可以在教科书的配套学习资源包中查看），并通过“本章扼要回顾”以及本章的项目活动评价，综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。

1. 单选题

（1）项目需求分析的目的是（　　）。

A. 和用户搞好关系　　B. 尽可能全面了解用户的业务需求

C. 大概了解用户的业务需求　　D. 让用户了解自己的能力

（2）以下不是常用的数据采集方法的是（　　）。

A. 面谈　　B. 实地调查　　C. 问卷调查　　D. 猜测

（3）以下不是结构化数据特征的是（　　）。

A. 可以用二维表结构来表示　　B. 没有具体的结构模型

C. 每一列代表一个不同的事实或度量　　D. 每一行表示一个实例或数据

2. 思考题

现实生活中的数据来源可以有多种途径，谈谈你对噪声数据的认识以及如何甄别数据。

3. 情境题

某市准备以“家风传承”为主题开展一次全市中小学生电脑作品比赛，鼓励全市中小学生以电脑作品的形式展现自己的家风。为更好地对收集到的作品进行登记及评奖，主办方需要为本次比赛建立一个数据库管理系统。

（1）假如你是本次数据库管理系统项目的设计人员，你认为该如何为本项目进行需求分析及设计解决方案？

（2）该项目将会产生哪些实体和关系？

第三章
数据管理

有效地管理数据可以帮助人们存储数据信息，把原本看似杂乱无章的数据转换成可供利用的数据资源，提高记录和检索信息的效率。只有对数据进行有效的管理才能发挥数据的价值与作用。

本章将通过“数据管理系统的数据管理”项目进行自主、协作、探究学习，让同学们使用数据库管理系统建立关系数据库，了解数据库基本的数据查询方法（如选择、投影、排序、统计等），能使用结构化查询语言进行简单的数据查询；结合实际案例，认识数据丢失的风险，利用实时备份与定时备份、全备份、增量备份与差异备份等多种方法进行数据备份，从而将知识建构、技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，促进信息技术学科核心素养达成，完成项目学习目标。

- 关系数据库的建立
- 数据的查询
- 数据的备份与恢复

项目范例　中学生体质健康数据管理系统的数据管理

情境

以《国家学生体质健康标准（2014年修订）》为标准记录某地区学生的体质健康数据，若以人工记录方式管理数据，数据量大容易出错。结合项目需求分析及对应的关系数据模型，为了更有效地及时跟踪和了解当前本地区中学生的体质健康情况，需要建立一个计算机数据管理系统对本地区中学生体质健康数据进行管理，这样才能更准确地记录数据并对这些数据进行统计分析，以便做出更加科学的指引。

主题

中学生体质健康数据管理系统的数据管理

规划

根据项目范例的主题，在小组中组织讨论，利用思维导图工具，制订项目范例的学习规划，如图3-1所示。

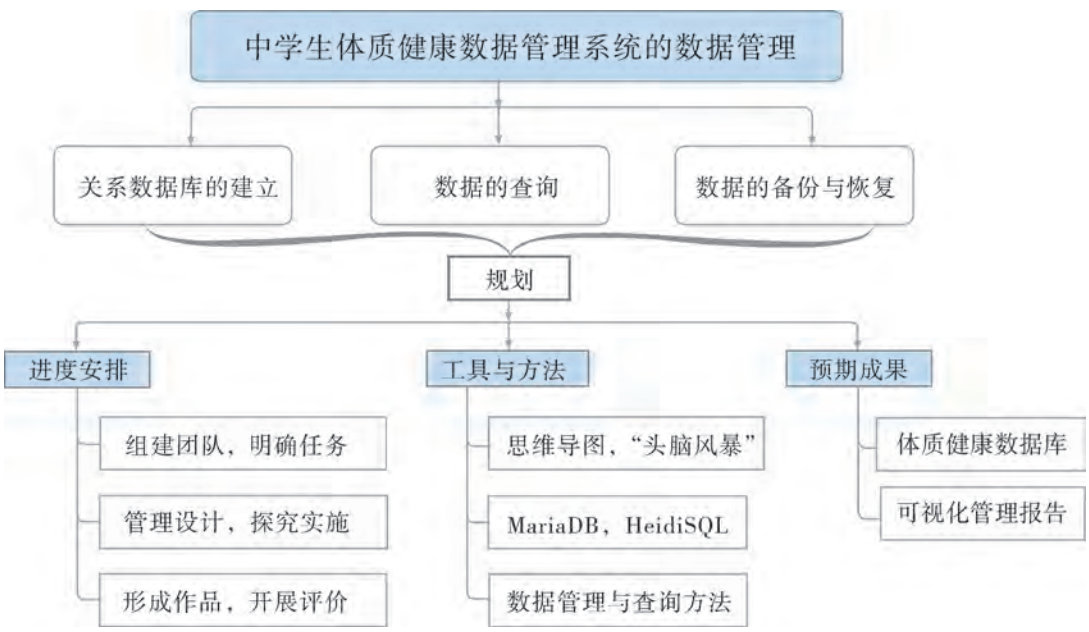


图3-1　“中学生体质健康数据管理系统的数据管理”项目学习规划

探究

根据项目学习规划的安排，通过调查、案例分析、文献阅读和网上资料搜索，开展“中学生体质健康数据管理系统的数据管理”项目学习探究活动，如表3-1所示。

表3-1“中学生体质健康数据管理系统的数据管理”项目学习探究活动

探究活动	学习内容		知识技能
关系数据库的建立	创建数据库和数据表。 修改表的结构。 建立表之间的联系。 数据库事务的处理。	在MariaDB中创建数据库和数据表。 修改表的结构完善数据库。 建立表之间的联系。 了解数据库记录中增、删、改及导入的相关操作。	使用数据库管理系统建立关系数据库。
数据的查询	数据库基本的查询方法。 使用结构化查询语言SQL查询数据。	选择查询。 投影查询。 排序查询。 统计查询。 结构化查询语言SQL的简介。 使用SQL语言查询数据。	了解数据库基本的数据查询方法（如选择、投影、排序、统计等）。 能使用结构化查询语言进行简单的数据查询。
数据的备份与恢复	数据丢失的风险及原因。 常见的数据备份与恢复方法。	数据丢失的风险及原因。 数据备份与恢复。	结合实际案例，认识数据丢失的风险。 利用实时备份与定时备份、全备份、增量备份与差异备份等多种方法进行数据备份。

实施项目学习各项探究活动，进一步剖析中学生体质健康数据管理系统的数据管理。

在小组开展项目范例学习过程中，利用思维导图工具梳理小组成员在“头脑风暴”活动中的观点，建立观点结构图，运用多媒体创作工具（如演示文稿、在线编辑工具等），综合加工和表达，形成项目范例可视化学习成果，并通过各种分享平台发布，共享创造、分享快乐。例如，运用在线编辑工具制作的“中学生体质健康数据管理系统的数据管理”可视化报告，可以在教科书的配套学习资源包中查看，其目录截图如图3-2所示。

图3–2　“中学生体质健康数据管理系统的数据管理”可视化报告的目录截图

评价

根据教科书附录2的“项目活动评价表”，对项目范例的学习过程和学习成果在小组或班级上进行交流，展开项目学习活动评价。

项目选题

同学们以3～6人组成一个小组，选择下面一个参考主题，或者自拟一个感兴趣的主题，开展项目学习。

1. 校运会管理系统的数据管理
2. 图书馆图书借阅管理系统的数据管理
3. 早餐营养搭配管理系统的数据管理

项目规划

各小组根据项目选题，参照项目范例的样式，利用思维导图工具，制订相应的项目方案。

各小组将完成的方案在全班进行展示交流，师生共同探讨、完善相应的项目方案。

3.1　关系数据库的建立

数据管理的第一步是建立数据库。我们根据已经设计好的“中学生体质健康数据管理系统”关系数据模型，建立相应的数据库，以实现对数据的有效管理。

3.1.1　创建数据库和数据表

数据库是长期储存在计算机内、有组织的、可共享的数据集合。数据库中的数据以一定的数据模型组织、描述和储存在一起，具有尽可能小的冗余度、较高的数据独立性和易扩展性的特点，并可在一定范围内为多个用户共享。这种数据集合具有如下特点：①尽可能不重复，以最优方式为某个特定组织的多种应用服务；②其数据结构独立于使用它的应用程序；③对数据的增、删、改、查由统一软件进行管理和控制。

数据表（Table）是数据库最重要的组成部分之一。数据库就像档案柜，是数据的物理存储区域。当我们使用档案柜存储资料时，会在档案柜中创建文件，然后将相关联的数据放入特定的文件中。在数据库领域中，这种文件就叫作“表”。数据表是结构化的文件，用来存储特定数据类型的数据。表可能存储客户清单、产品目录或者其他信息列表。

人们经常使用术语“数据库”来指代他们运行的数据库软件，这是错误的。数据库软件实际上称为数据库管理系统，数据库是通过DBMS创建和操作的容器。

目前数据库管理系统软件非常多，常用的有ACCESS，SQL Server，MySQL，MariaDB，Oracle，Sqlite，PostgreSQL数据库等。请同学们查阅文献，讨论这几种常用数据库系统的应用场合及优缺点，选择自己感兴趣的数据库管理软件作为项目实现的软件基础，为小组自选项目建立数据库和数据表。

目前几乎所有的主流数据库管理系统，如Oracle，Sybase，MySQL，MariaDB，Sqlite，SQL Server，PostgreSQL均支持关系数据模型。本教科书选用了开源数据库管理系统软件MariaDB（MySQL分支之一，开源免费版）作为项目范例数据库管理实现的软件基础。同时，为了操作直观和方便，教科书还选用了具有图形用户界面的HeidiSQL作为对MariaDB的操作界面。相关软件的安装程序可从教科书配套资源包中下载安装。

建立关系数据库一般包括两个步骤：

（1）创建数据库。

（2）在数据库中创建数据表。

根据已经建立的“中学生体质健康数据管理系统”的关系数据模型创建对应的数据库和数据表，具体操作如下：

1. 启动MariaDB和HeidiSQL，在HeidiSQL窗口中，创建与MariaDB的连接，在会话管理窗口中按“新建”按钮，设置用户与密码，并将Unnamed重命名为“localhost”后进行保存，如图3-3所示。

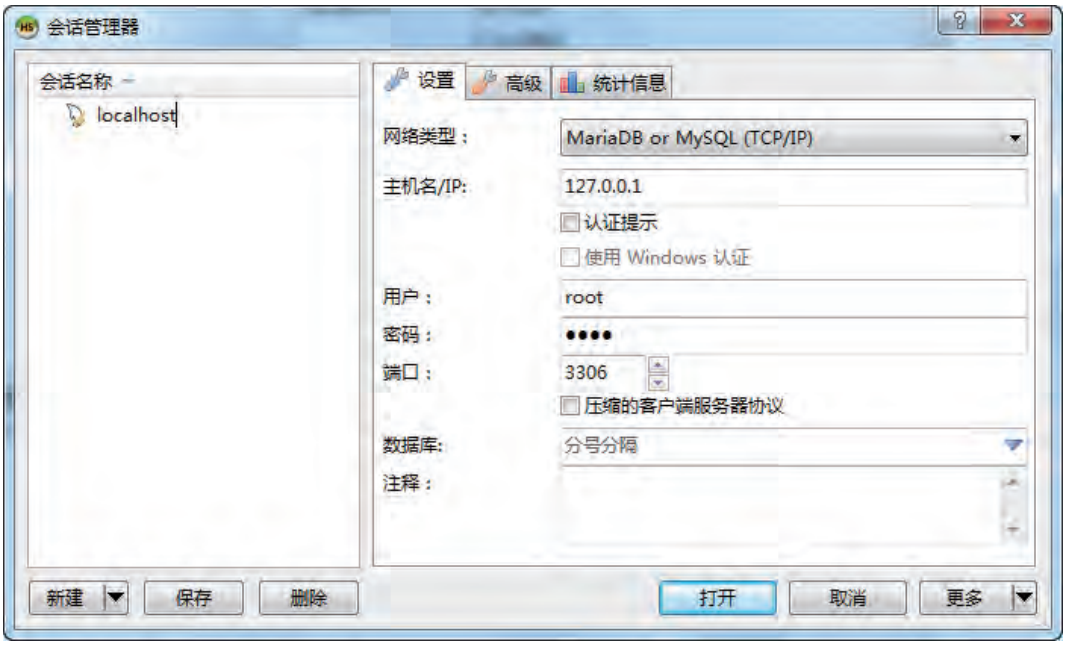


图3-3　创建数据库连接

2. 在连接“localhost”中按鼠标右键弹出下拉快捷菜单，选择“创建新的”→“数据库”，在创建数据库界面的“名称”后输入“中学生体质健康数据管理系统”，并将字符集选择为“utf8_general_ci”，如图3-4所示。

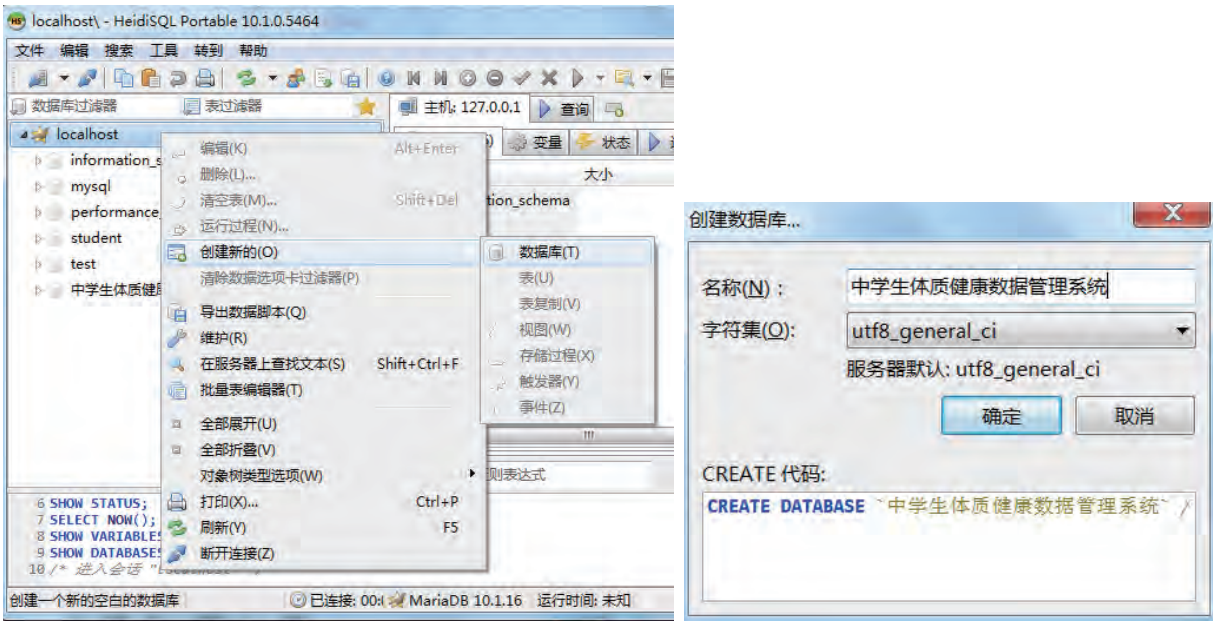


图3-4　创建空的数据库

3. 创建数据表。

创建数据表首先要定义数据表的结构，包括以下三个方面的内容：

（1）确定数据表中各个字段的名称。

（2）设置各个字段的属性，包括字段的数据类型、字段说明和对字段的约束条件等。

（3）确定数据表的主键。

例如，“学生表”中字段名、数据类型、主键设定如表3-2所示。

表3-2　“学生表”结构

字段名	数据类型	长度	是/否主键
学籍号	文本	19	是
姓名	文本	12	
性别	文本	2	
年级	文本	8	
学校名称	文本	255	
班别	文本	255	
出生日期	日期/时间		
民族	数字	长整型	
总得分	数字	长整型	
等级	文本	6	
备注	文本	255	

右键单击新建的“中学生体质健康数据管理系统”数据库，在快捷菜单中选择“创建表”→“表”。在表设计视图窗口中分别输入字段名称，并选择数据类型。在“学籍号”字段上点击右键选择“创建新索引”，选择“KEY”并保存，如图3-5所示。

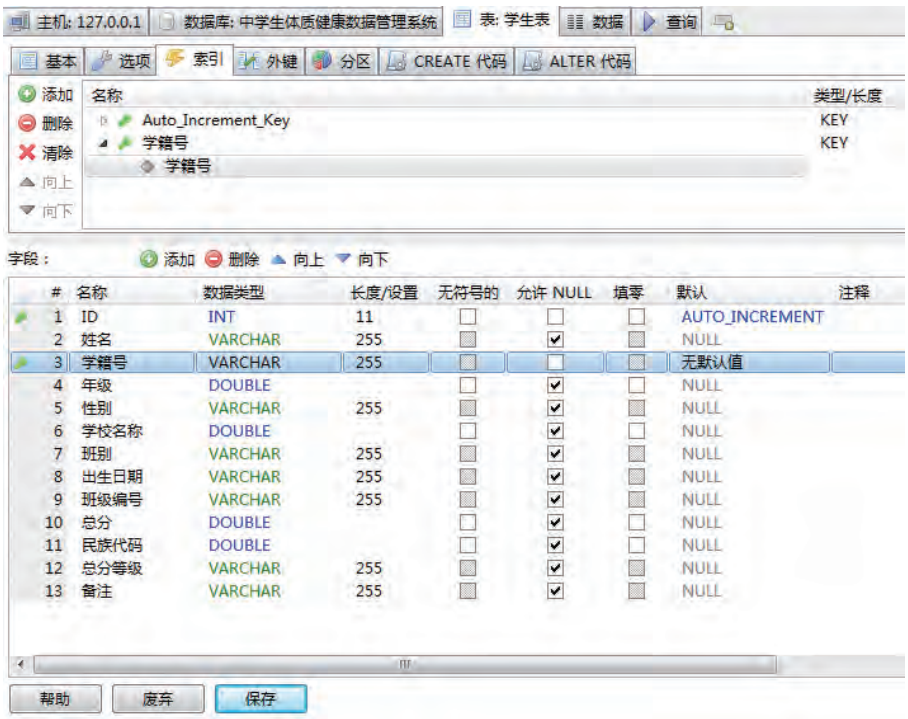


图3-5　“学生表”的表结构

3.1.2　修改表的结构

有时候我们在数据分析阶段设计出来的数据表结构并不一定完善，例如，在创建“指标项目表”时，对应的关系数据模型为：指标项目表（项目编号，测试对象，项目名称，项目单位，项目权重，备注），负责录入的同学忘记了创建“项目单位”字段，而且把“备注”字段放在了“项目名称”和“项目权重”字段之间，如图3-6所示。同学们思考如何才能完善表结构，使其与关系数据模型对应。

图3-6　不完善的“指标项目表”的表结构

一般的数据库管理系统软件都允许我们在创建好数据表后，甚至输入数据后再对表的结构进行修改。一般修改数据表的结构包括添加、删除、移动字段和改变字段类型等操作。需要注意的是，在创建完数据库和数据表后，如果要继续修改数据表，必须先打开数据库和数据表。

为了完善“中学生体质健康数据管理系统”数据库中“指标项目表”的表结构，我们需要增加“项目单位”字段。具体操作如下：

1. 在数据库窗口中，选中“指标项目表”，编辑数据表，如图3-7所示。

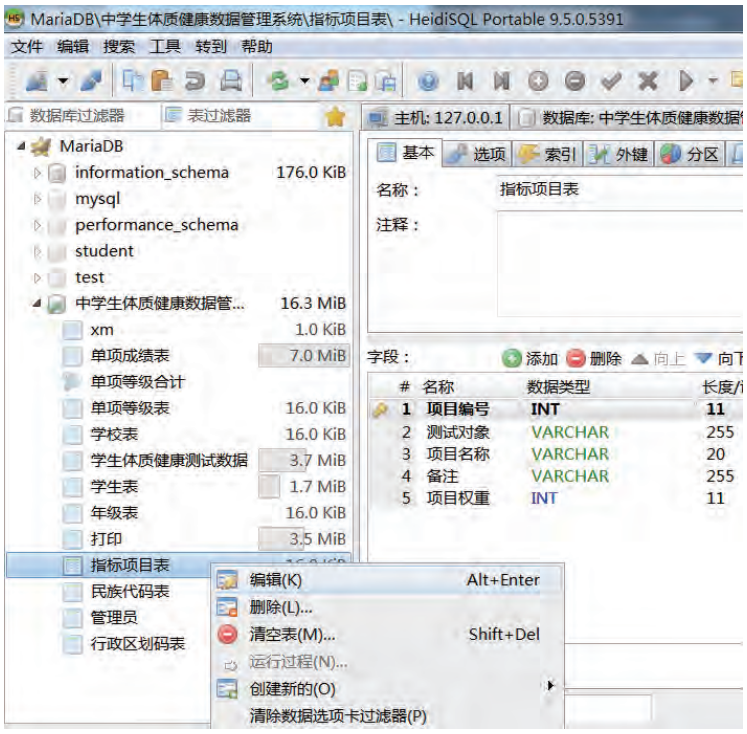


图3-7　编辑数据表

2. 选择“项目名称”字段行，按鼠标右键弹出下拉菜单，选择“添加”按钮，在该选中字段后面插入一空行，直接在空行上输入要插入的字段信息即可，如图3-8所示，完善后的“指标项目表”的表结构如图3-9所示。

图3-8　添加字段

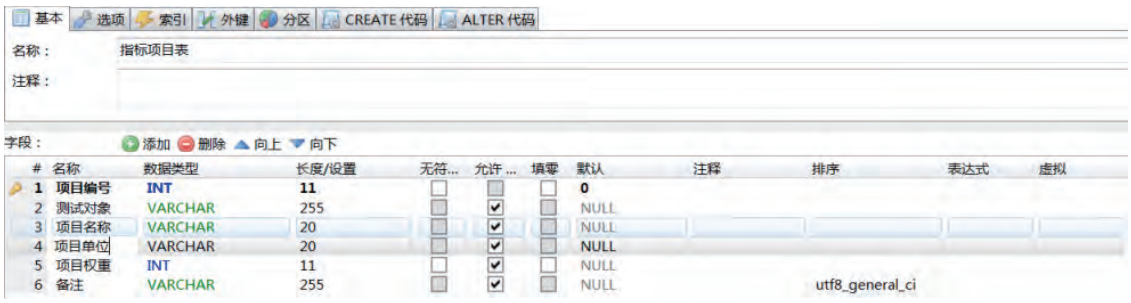


图3–9　完善后的“指标项目表”的表结构

3.1.3　建立表之间的联系

问题

数据库是多个数据表的集合，而数据表并不是相互孤立的，有些表之间是有一定联系的。如何为这些表建立联系？

在HeidiSQL中，我们可以通过添加外键的方法为数据表建立关联。

实践

为“中学生体质健康数据管理系统”中“单项成绩表”“学生表”和“指标项目表”建立关联，分别添加外键，如图3–10所示。

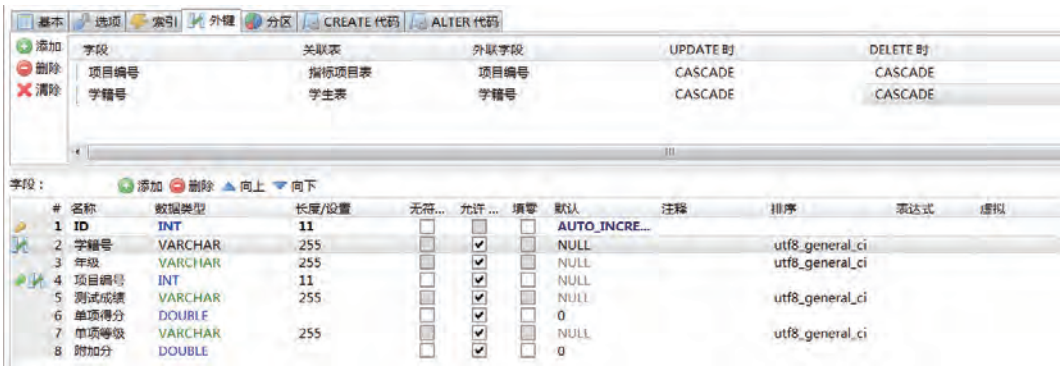


图3–10　建立关联

3.1.4　数据库事务的处理

思考

建立了数据库和数据表后，如何为数据表增加记录？对于类似“学校”“性别”和“民族”等字段的值经常会有重复，如何提高记录输入的效率？

编辑数据库包括对记录的增加、删除、修改等操作。其中记录的增加包括直接输入记录数据，也可以通过设置字段的查阅方式输入记录数据，还可以选择参照另一个数据表的输入方式输入记录数据。

交流

为“中学生体质健康数据管理系统”中的工作表添加记录，与同学们交流并根据需要对数据记录进行删除和修改。

1. 增加记录。

（1）直接在数据表中添加记录。

一般情况下，我们可以直接打开数据库中的数据表，在需要增加记录的位置选择“”按钮，输入需要增加的内容并保存，如图3-11所示。

图3-11　利用“打开表”直接添加记录数据

（2）设置自行输入的查阅方式。

有些特定的数据，比如“性别”字段只有两个值“男”和“女”，直接输入重复的工作量比较大，我们可以通过“设置字段的查阅方式”进行输入。具体操作如下：

①在“学生表”的编辑窗口修改性别字段类型，将varchar修改为enum，如图3-12所示。

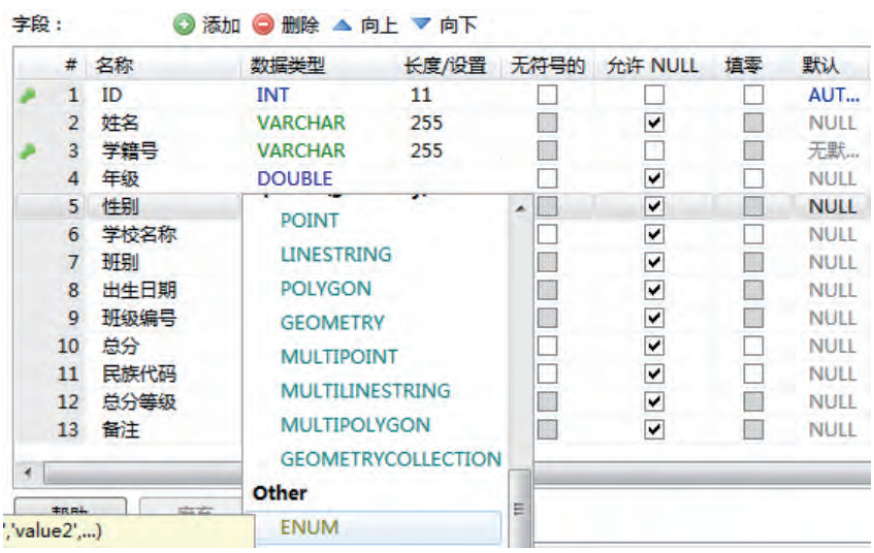


图3-12　修改“性别”字段类型

②设置“性别”字段的值为“男”“女”，如图3-13所示。

图3-13　设置“性别”字段的值

③在“学生表”的数据视图选择“性别”字段，在倒三角形上按鼠标左键出现“男”“女”，可以直接用鼠标左键选择性别，如图3-14所示。

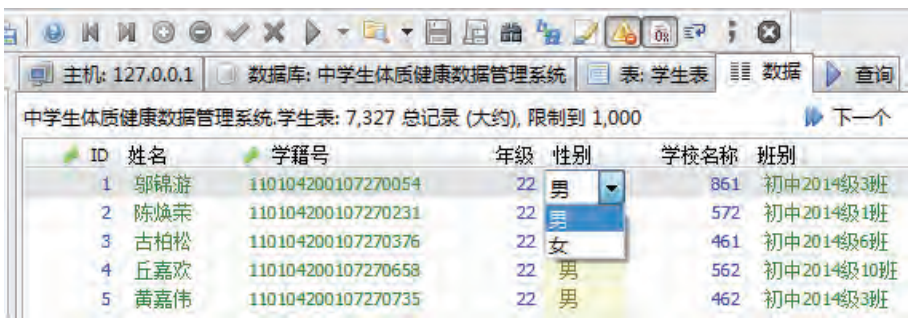


图3-14　用鼠标左键选择性别

（3）设置参照另一数据表的输入方式。

还有些字段数据的输入需要参照另外一个数据表的数据进行输入，比如“学生表”中的“民族代码”字段。为了让数据的输入更加简便不容易出错，我们可以新建立一个“民族代码表”，“学生表”中的“民族代码”字段数据的输入就参照“民族代码表”中的数据输入，如图3-15所示，“学生表.民族代码”与“民族代码表.民族”的数据类型要相同，若不同，要修改一致。

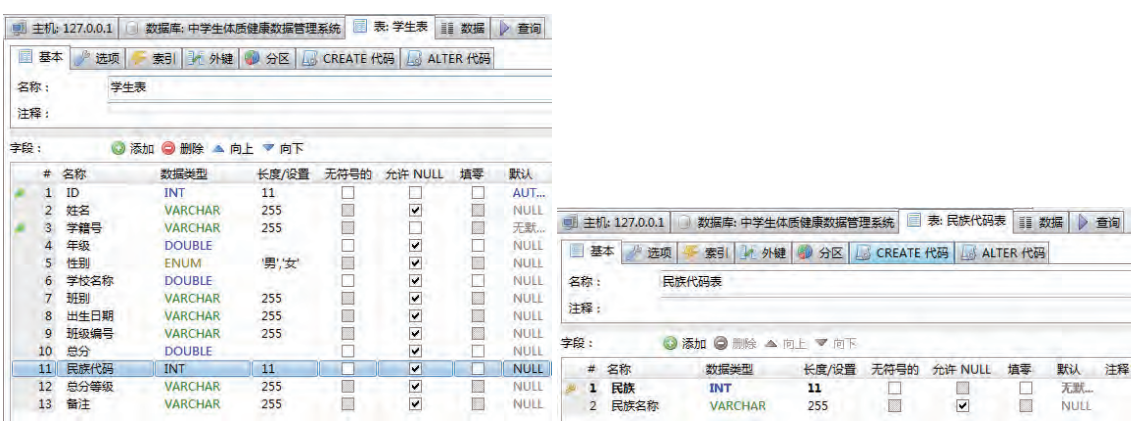


图3-15　“学生表.民族代码”与“民族代码表.民族”数据类型要相同

①在“学生表”中选择“外键”，出现如图3-16所示的“外键”对话框。

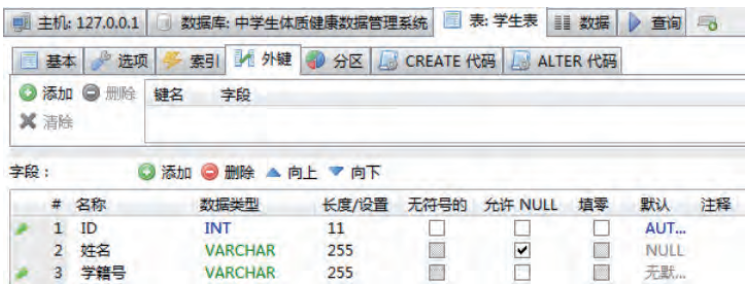


图3－16　“外键”对话框

②在“外键”中按“＋”按钮添加，点击字段下面空白处并左键选择“民族代码”，设置关联表为“民族代码表”，设置外联字段为“民族”，设置删除时和更新时均选“CASCADE”，点击“保存”按钮，结果如图3-17所示。

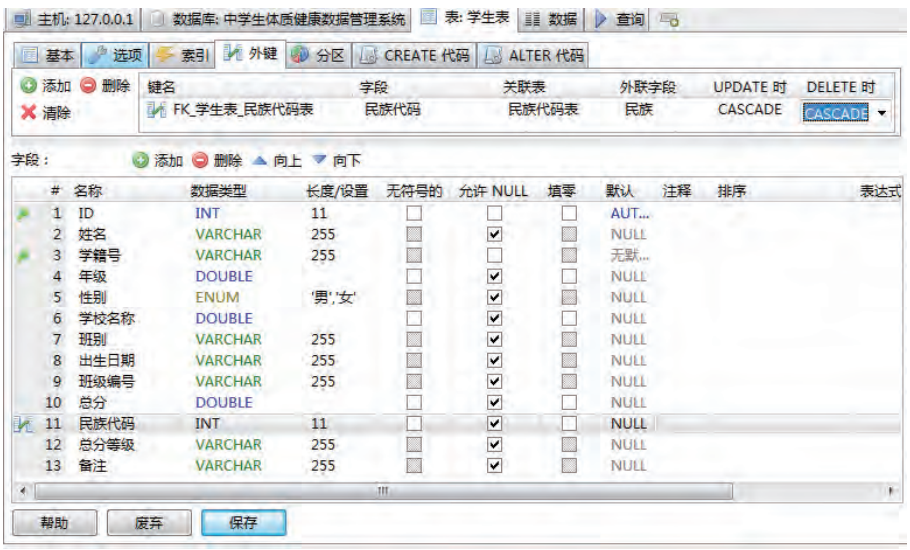


图3－17　设置“学生表”外键

③进入“学生表”的数据视图，点击“民族代码”字段的“▼”按钮，即可选择相应的代码填充，如图3-18所示。

图3-18　填充民族代码

（4）导入数据。

作为典型的开放型数据库，MariaDB支持与其他类型的数据库文件进行数据的交换和共享，同时也支持与其他Windows程序创建的数据文件进行数据交换。数据的导入就是将其他格式的数据合并在数据库中，并实现对导入数据的调用。

比如，在“中学生体质健康数据管理系统”中，我们已经把学校信息录入一个Excel表格文件中，如何把这个Excel表格中的数据导入MariaDB管理系统中呢?

为了把如图3-19所示的“学校表.xlsx”的数据导入HeidiSQL中，我们需将“学校表.xlsx”另存为csv格式的文件“1.csv”（为了便于以后读取数据，需保存在英文路径下），然后用记事本打开“1.csv”并另存为utf8格式的“1.csv”文件，才能重新导入HeidiSQL的“中学生体质健康数据管理系统”数据库中。具体数据库导入操作如下：

A	B	C	D	E	F	G
ID	学校编号	学校名称	地址	联系电话	电子邮箱	备注
1	171	海南省海	海南省海	0898-8888	171@qq.com	
2	172	广东省南	广东省南	0757-6666	172@qq.com	
3	261	广东省湛	广东省湛	0759-8899	261@qq.com	
4	364	广东省梅	广东省梅	0753-6666	364@qq.com	
5	365	湖南省长	湖南省长	0731-8888	365@qq.com	
6	461	湖北省武	湖北省武	027-88887	461@qq.com	
7	462	湖北省天	湖北省天	0728-8886	462@qq.com	
8	464	江西省南	江西省南	0791-8888	464@qq.com	
9	465	安徽省合	安徽省合	0551-8888	465@qq.com	
10	561	浙江省杭	浙江省杭	0571-8888	561@qq.com	
11	562	广东省广	广东省广	020-88888	562@qq.com	
12	565	广东省东	广东省东	0769-8888	565@qq.com	
13	572	广东省河	广东省河	0762-8888	572@qq.com	
14	661	广东省深	广东省深	0755-8888	661@qq.com	
15	662	广东省汕	广东省汕	0754-8888	662@qq.com	
16	764	广东省江	广东省江	0750-9999	700000064@qq.com	
17	861	广西桂林	广西桂林	0773-8888	861@qq.com	

图3-19　Excel表格文件“学校表.xlsx”

①在HeidiSQL中打开“中学生体质健康数据管理系统”数据库，按鼠标左键选中“学校表”，在“工具”菜单下选中“导入CSV文件”，如图3-20所示。

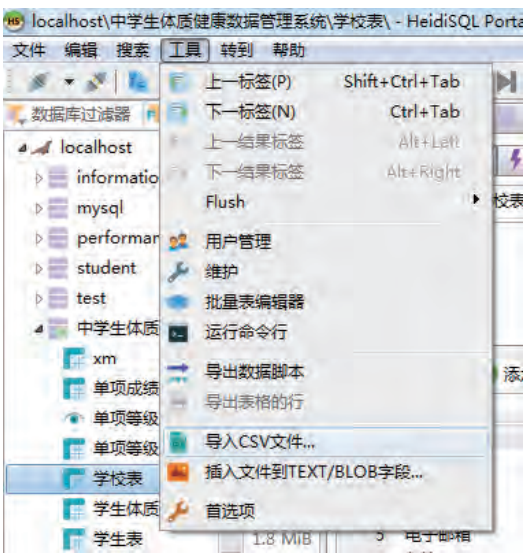


图3-20　导入CSV文件

②导入文件“1.csv”，在“导入文本文件”界面进行相关设置，然后点击“导入！”按钮，完成导入CSV文件，如图3-21所示。

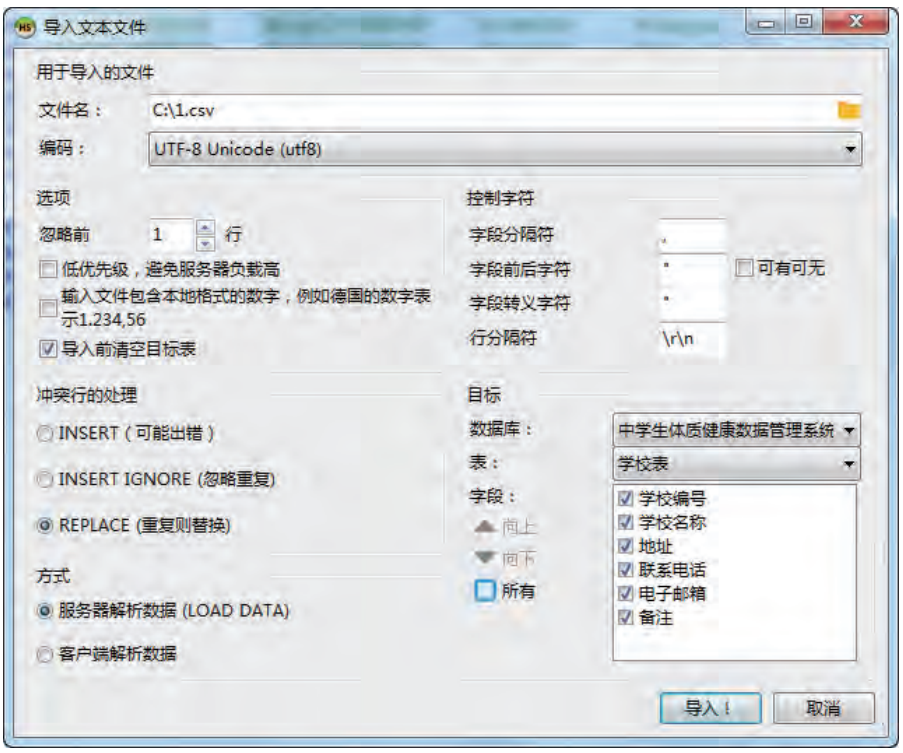


图3-21　导入扩展名为CSV的文件

2. 删除记录。

如果在管理数据过程中需要删除表中的某个记录，只需要在数据视图中选择需要删除的记录，点击工具栏上的“”按钮，然后点击“确定”按钮即可完成删除记录的操作。例如，删除“陈焕荣”的记录，如图3-22所示。

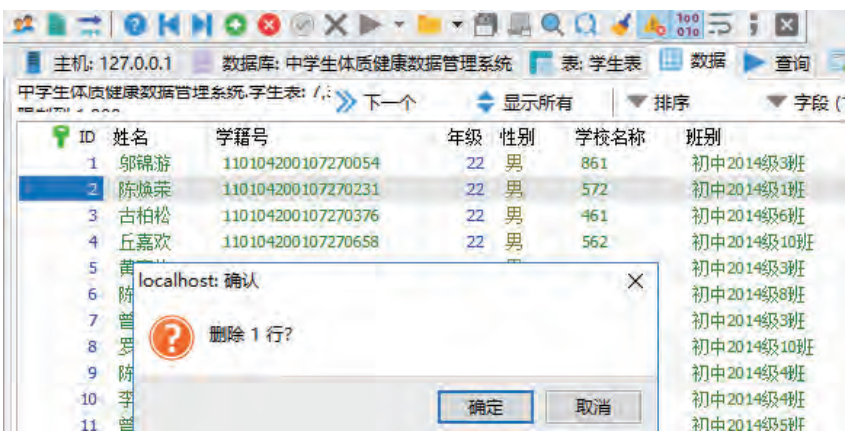


图3-22　删除记录

当然，我们也可以利用点击鼠标右键的快捷菜单来完成删除记录的操作，比如删除学生“陈焕荣”的记录，如图3-23所示。

图3–23　利用快捷菜单删除记录

3. 修改记录。

如果要修改已有的记录数据，只要在数据视图中双击要修改的数据，在弹出的对话框中输入要修改的内容，点击✓按钮，即可完成修改，如图3–24所示。

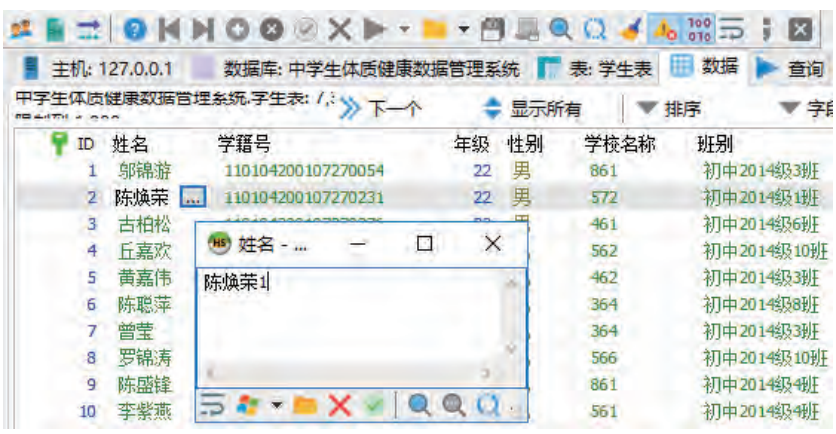


图3–24　修改记录数据

4. 插入记录。

如果要插入记录，只要点击工具栏上的⊕按钮，即可在表当前光标处插入新记录，插入完成后，点击✓按钮，如图3–25所示。

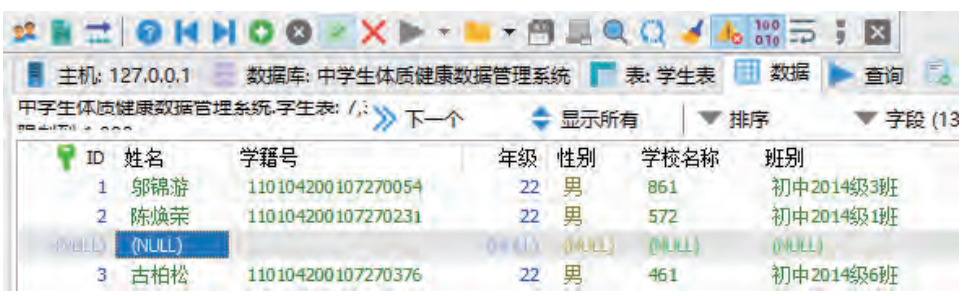


图3–25　插入新记录

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，建立项目的关系数据库。

1. 实践创建数据库和数据表的方法。
2. 实践数据库的增、删、改、导入等操作。

3.2 数据的查询

数据库里存储着大量的数据，如果能充分利用数据库管理系统提供的各种功能来检索数据及输出报表，那么我们就可以节省大量的时间和精力。从数据库中经过筛选获取满足条件数据的过程称为数据查询或查询数据库。

3.2.1 数据库基本的查询方法

探究活动

讨论

以小组为单位，通过查阅相关资料了解数据库基本的查询方法有哪些，并和小组成员分享这些方法的使用场合。

数据查询的方法有许多，包括数据的选择、投影、排序、统计等。创建查询时，我们要确定该查询涉及哪些字段，这些字段涉及哪些表，有什么约束条件以及对查询结果显示的要求等。通常可以按照如图3–26所示的顺序来创建一个查询。

图3–26 查询的一般过程

1. 选择查询

选择查询是从一个关系中找出满足给定条件的记录的操作，是从行的角度进行的运算，选出满足条件的那些记录构成原关系的一个子集。

思考

如何查看单项测试成绩获得优秀的学生记录？如何查看在50米跑步项目中获得优秀的学生记录？

实践

在HeidiSQL中一般是通过 查询 实现查询。实现选择查询的具体方法如下：

（1）在数据库中选中“单项成绩表”，选择工具栏 查询 按钮进入查询视图，从“SQL关键字”中找到“SELECT”，如图3-27所示。

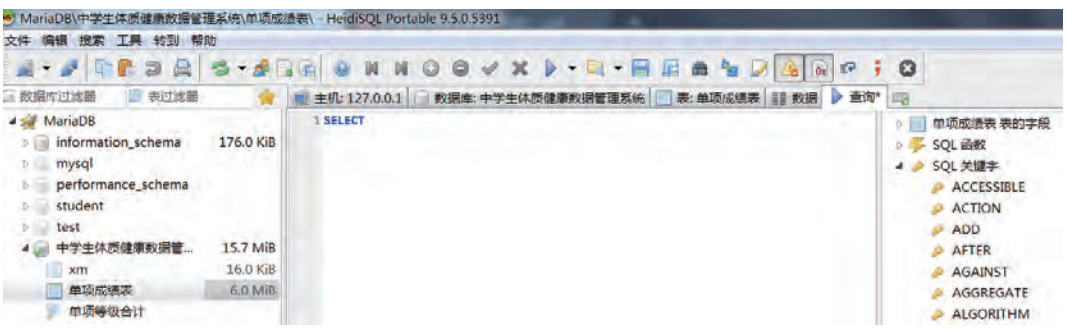
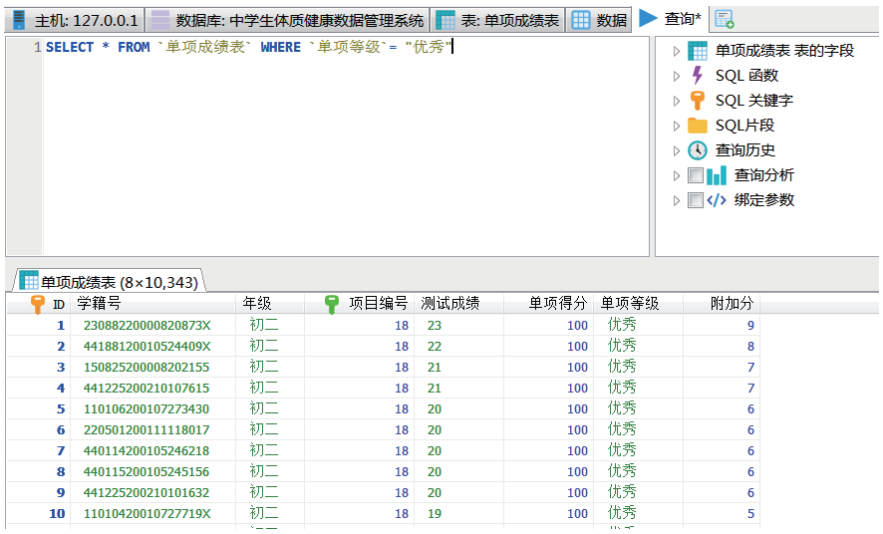


图3-27　进入查询视图

（2）在查询窗口中，分别在“单项成绩表 表的字段”“SQL函数”和“SQL关键字”等选项中选择相应的选项，添加查询条件，如图3-28所示。

图3-28　添加查询条件

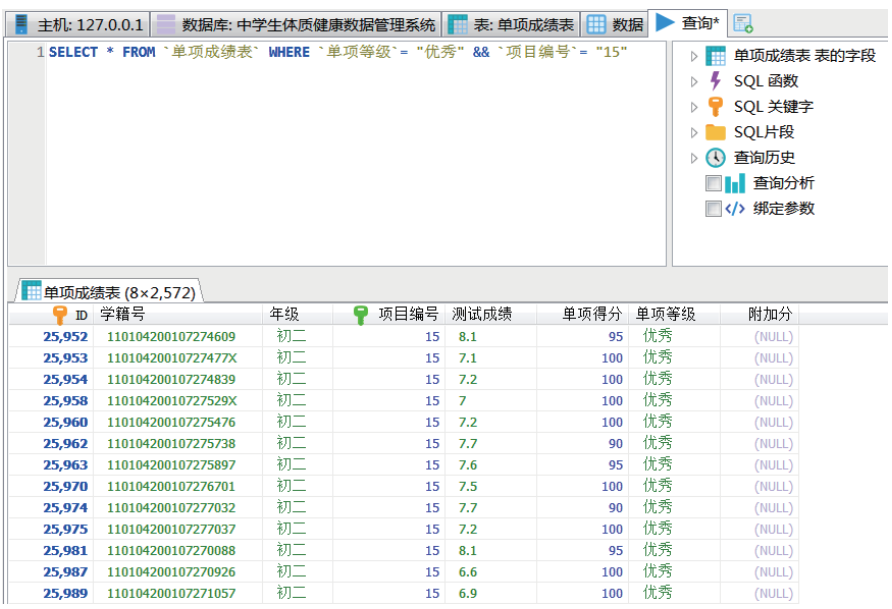
（3）按 按钮运行，查询结果如图3-29所示。



ID	学籍号	年级	项目编号	测试成绩	单项得分	单项等级	附加分
1	23088220000820873X	初二	18	23	100	优秀	9
2	441881200105244O9X	初二	18	22	100	优秀	8
3	150825200008202155	初二	18	21	100	优秀	7
4	441225200210107615	初二	18	21	100	优秀	7
5	110106200107273430	初二	18	20	100	优秀	6
6	220501200111118017	初二	18	20	100	优秀	6
7	440114200105246218	初二	18	20	100	优秀	6
8	440115200105245156	初二	18	20	100	优秀	6
9	441225200210101632	初二	18	20	100	优秀	6
10	110104200107277l9X	初二	18	19	100	优秀	5

图3-29 查询结果

如果仅仅要查看某个项目的优秀学生记录，比如50米跑步项目，由于从“指标项目表”可查到50米跑步对应的项目编号为“15”，所以需要在图3-29的基础上继续添加查询条件，其结果如图3-30所示。



ID	学籍号	年级	项目编号	测试成绩	单项得分	单项等级	附加分
25,952	110104200107274609	初二	15	8.1	95	优秀	(NULL)
25,953	11010420010727477X	初二	15	7.1	100	优秀	(NULL)
25,954	110104200107274839	初二	15	7.2	100	优秀	(NULL)
25,958	11010420010727529X	初二	15	7	100	优秀	(NULL)
25,960	110104200107275476	初二	15	7.2	100	优秀	(NULL)
25,962	110104200107275738	初二	15	7.7	90	优秀	(NULL)
25,963	110104200107275897	初二	15	7.6	95	优秀	(NULL)
25,970	110104200107276701	初二	15	7.5	100	优秀	(NULL)
25,974	110104200107277032	初二	15	7.7	90	优秀	(NULL)
25,975	110104200107277037	初二	15	7.2	100	优秀	(NULL)
25,981	110104200107270088	初二	15	8.1	95	优秀	(NULL)
25,987	110104200107270926	初二	15	6.6	100	优秀	(NULL)
25,989	110104200107271057	初二	15	6.9	100	优秀	(NULL)

图3-30 增加查询条件后的结果

“学生表”中记录了许多学生的属性信息，如果仅仅需要查询显示学生的总分等级该如何操作？

2. 投影查询

投影查询是从一个关系中选出若干指定字段的值的操作，是从列的角度进行的运算，所得到的字段个数通常比原来关系中少或排序顺序不同。

体验

在查询视图中实现投影查询，具体方法如下：

（1）打开“中学生体质健康数据管理系统”数据库，选中“学生表”，单击 查询 按钮进入查询视图，在查询窗口分别通过“SQL关键字”选择关键字“SELECT”和“FROM”，通过“学生表　表的字段”选择字段“姓名”和“总分等级”，注意字段之间要用英文的逗号间隔，最后选择“表　学生表”，如图3-31所示。

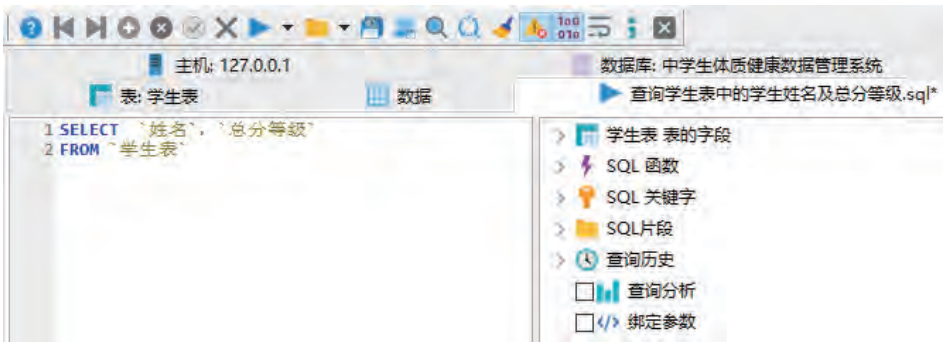


图3-31　创建查询并设置查询条件

（2）点击工具栏中的 ▶ 按钮，查询结果如图3-32所示。

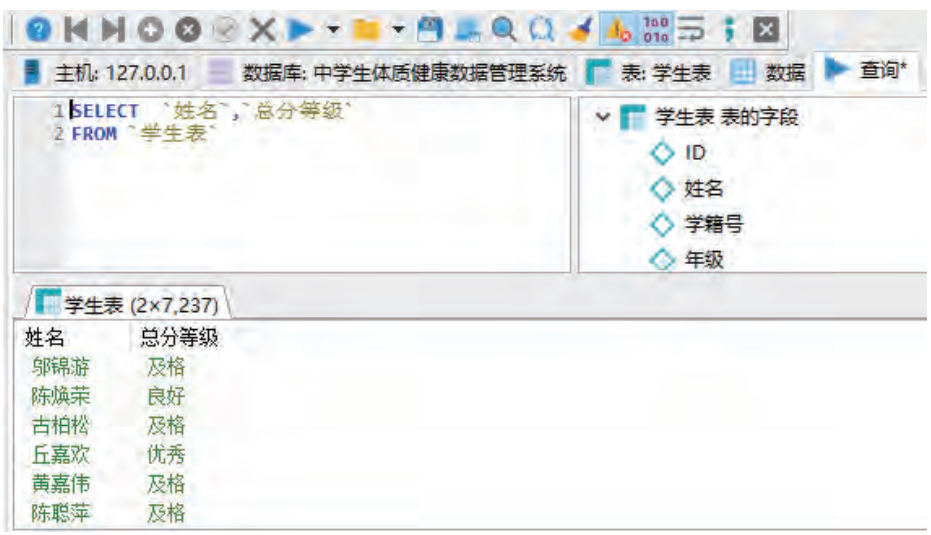


图3-32　运行查询后的结果

思考

上述实践中，通过创建查询获得了“学生表”中所有学生的总分等级。但是这些记录显示的顺序有些杂乱，能否让查询后得到的数据按某个字段（比如“总分等级”）的值进行排序呢？

3. 排序查询

查询检索得到的数据如果没有排序，数据通常按照底层表中的顺序显示。然而，如果数据随后被更新或者删除，这个顺序将会受到MariaDB如何重用回收的存储空间的影响。关系数据库设计理论认为，如果没有显示指定排序，不应该认为检索的数据顺序是有意义的。若要为查询检索到的数据添加排序，则需要在查询中添加“ORDER BY”选项。

在如图3-31所示查询条件的基础上，与同学们交流，继续从“SQL关键字”选择关键字“ORDER”和“BY”，从“学生表　表的字段”中选择字段“总分等级”，再从“SQL关键字”中选择关键字“DESC”，如图3-33所示。重新运行查询，查询结果按照“总分等级”字段值进行排序，如图3-34所示。

图3-33　为查询添加排序选项

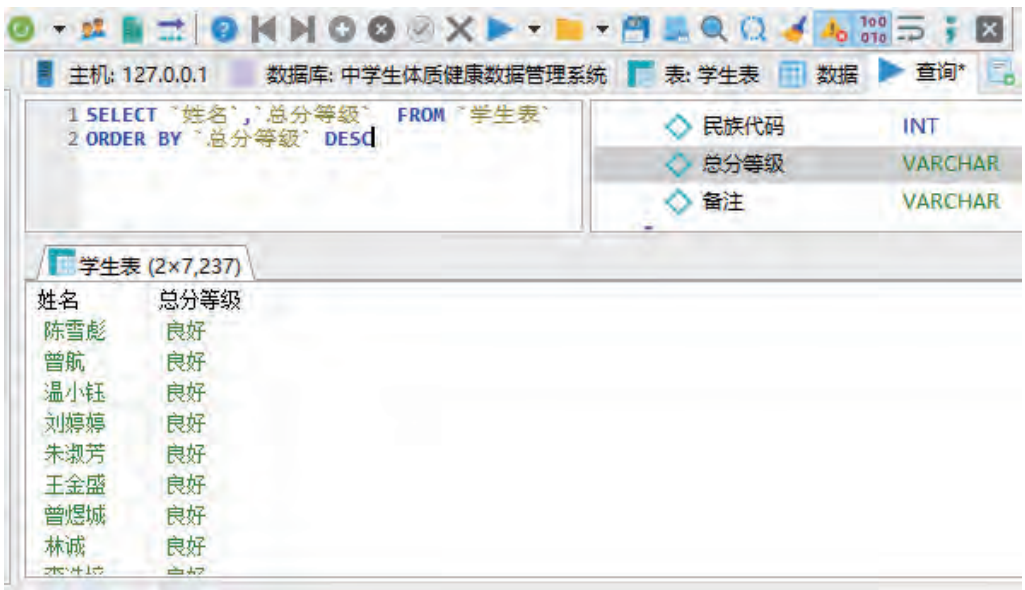


图3-34　运行排序查询后的结果

排序中的“ASC”和“DESC”分别代表什么含义？

4. 统计查询

数据录入数据库后，除了对里面的数据进行选择查询、投影查询和排序查询等操作外，还经常需要对里面的数据进行一些统计，如统计各测试指标项目的平均得分、各学校参加测试的学生人数、各学校有多少学生测试的综合成绩可以达到“优秀”等级……有时候，我们还需要将这些统计结果以报表的形式表现出来。

如何利用“中学生体质健康数据管理系统”数据库统计各测试指标项目的平均分是多少？如何利用“中学生体质健康数据管理系统”数据库中的“单项成绩表”统计出各单项等级中共有多少人（即优秀、合格等各有多少人次），他们的学籍号、测试成绩、单项得分各是多少？最后将统计结果以报表的形式输出。

数据统计的目的是将表的记录予以分组后，再加以计算。

要计算某一个班学生各测试指标项目的平均分，就是将该班所有的测试得分按指标项目分组进行统计。这里需要用到的表是“指标项目表”和“单项成绩表”，其中约束条件是“指标项目表.项目编号”=“单项成绩表.项目编号”，具体操作步骤如下：

（1）打开“中学生体质健康数据管理系统”数据库，在“查询”中新建一个新的查询。

（2）在“查询”视图中分别通过“SQL关键字”和“SQL函数”添加相应的关键字，并设置所需的表和字段，设置约束条件。

（3）在“SELECT”行的“单项得分”前增加“SQL函数”“avg”,通过“SQL关键字”“AS”设置别名为“平均分”。

（4）在“GROUP BY”行中添加“指标项目表.项目编号”，如图3-35所示。

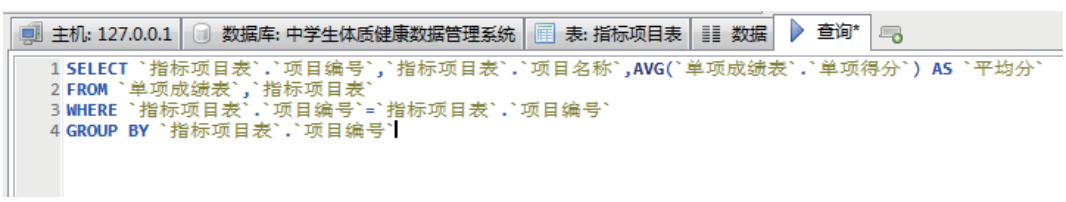
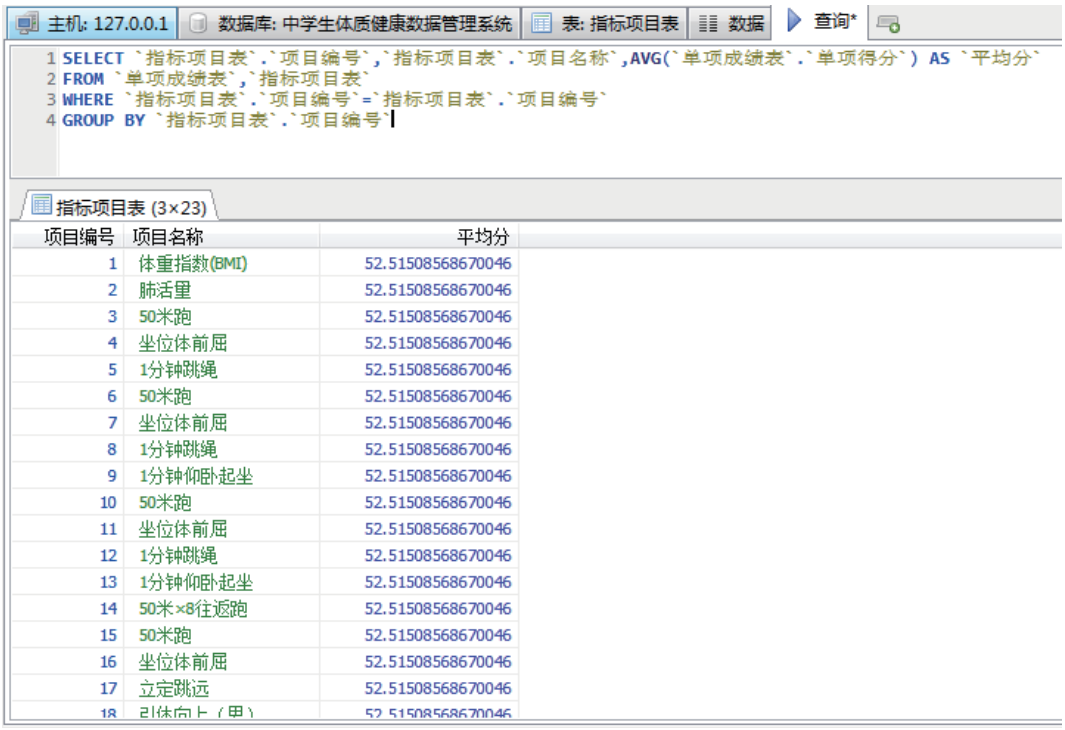


图3-35　选择汇总功能

（5）运行查询，部分结果如图3-36所示。



项目编号	项目名称	平均分
1	体重指数(BMI)	52.51508568670046
2	肺活量	52.51508568670046
3	50米跑	52.51508568670046
4	坐位体前屈	52.51508568670046
5	1分钟跳绳	52.51508568670046
6	50米跑	52.51508568670046
7	坐位体前屈	52.51508568670046
8	1分钟跳绳	52.51508568670046
9	1分钟仰卧起坐	52.51508568670046
10	50米跑	52.51508568670046
11	坐位体前屈	52.51508568670046
12	1分钟跳绳	52.51508568670046
13	1分钟仰卧起坐	52.51508568670046
14	50米×8往返跑	52.51508568670046
15	50米跑	52.51508568670046
16	坐位体前屈	52.51508568670046
17	立定跳远	52.51508568670046
18	引体向上（男）	52.51508568670046

图3-36 统计的结果

汇总功能的部分选项内容功能如下：

分组（Group by）：按某一字段对记录进行分组。

合计（Sum）：计算字段中值的总和。

平均值（Avg）：计算平均值。

最小值（Min）：搜索该字段的最小值。

最大值（Max）：搜索该字段的最大值。

计数（Count）：计算记录条数。

3.2.2 使用结构化查询语言SQL查询数据

到目前为止，我们所学习的有关数据库的查询方法都是在HeidiSQL中通过设置“SQL函数”和“SQL关键字”等方式来实现的。不同的数据库管理系统软件，其图形操作界面会有所不同。本节我们将学习使用结构化查询语言SQL实现数据库的数据查询。

1. 结构化查询语言SQL简介

结构化查询语言（Structured Query Language，简称SQL）是关系数据库的标准语言，由于它具有功能丰富、使用方便灵活、语言简洁易学等突出的优点，因而深受计算机界和计算机用户的欢迎。1986年10月，美国国家标准局（ANSI）的数据库委员会批准将SQL作

为数据库语言的美国标准，同年公布了标准SQL。此后不久，国际标准化组织（ISO）也做出了同样的决定。

SQL语言具有如下特点：

（1）数据描述、操纵、控制等功能一体化。

SQL原意为结构化查询语言，但实际具有集查询、操纵、定义和控制等四方面功能于一身的一体化特点。

①查询语言（Query Language，QL），用于查询数据。

②数据操纵语言（Data Manipulation Language，DML），用于增、删、改数据。

③数据定义语言（Data Definition Language，DDL），用于定义、撤销和修改数据库、表、视图及索引等。

④数据控制语言（Data Control Language，DCL），用于数据访问权限的控制。

（2）两种使用方式，统一的语法结构。

SQL的使用方式有两种。一种是交互式联机使用方式，另一种是嵌入某种高级语言（宿主语言）中使用。交互式使用方式适合于对系统的维护，嵌入式使用方式主要用于应用程序的开发，这两种使用方式都可以由用户根据需要灵活地选定。

（3）高度的非过程化。

使用SQL，用户只要提出“干什么”，而无须具体指明“怎么干”，如存取路径选择和具体处理操作等，均由系统自动完成。

（4）语言简洁，易学易用。

尽管SQL的功能很强，但语言十分简洁，核心功能只用了9个动词，而且语法接近英语口语，易学易用，如表3-3所示。

表3-3　SQL的核心动词

SQL功能	动词
数据查询	SELECT
数据定义	CREATE，DROP，ALTER
数据操纵	INSERT，UPDATE，DELETE
数据控制	GRANT，REVOKE

本小节，我们将探讨如何运用SQL语言进行数据的查询。

2. 使用SQL语言查询数据

讨论

以小组为单位，观察在项目实施中数据查询的图形操作界面，找出出现结构化查询语言SQL的地方，讨论并归纳结构化查询语言SQL的格式。

SQL语言提供了SELECT语句进行数据库的查询，其基本结构是由SELECT－FROM－WHERE组成的，其语法格式如下：

```
SELECT[ALL/DISTINCT]<目标列表达式>[，<目标列表达式>]
FROM＜表名或视图名＞[，<表名或视图名>]……
[WHERE<条件表达式>]
[GROUP BY<列名1>[HAVING<条件表达式>]]
[ORDER  BY＜列名 2>[ASC/DESC]];
```

（1）单表查询。

思考

如何用SQL语言实现在“学生表”查询“总分等级”为“优秀”的学生记录的姓名、性别、所在学校名称及对应的总分等级？

根据SELECT语句的特点，查询“学生表”的信息，条件表达式为总分等级 ='优秀'，同时指定了显示的字段为“姓名”“性别”“学校名称”和“总分等级”，所以可以写出对应的SELECT语句为：

```
SELECT 姓名，性别，学校名称，总分等级  FROM 学生表  WHERE  总分等级 ='优秀'
```

在HeidiSQL图形界面中创建查询窗口可以直接在光标所在位置输入以上代码，如图3-37所示。进行运行即可执行查询，结果如图3-38所示。

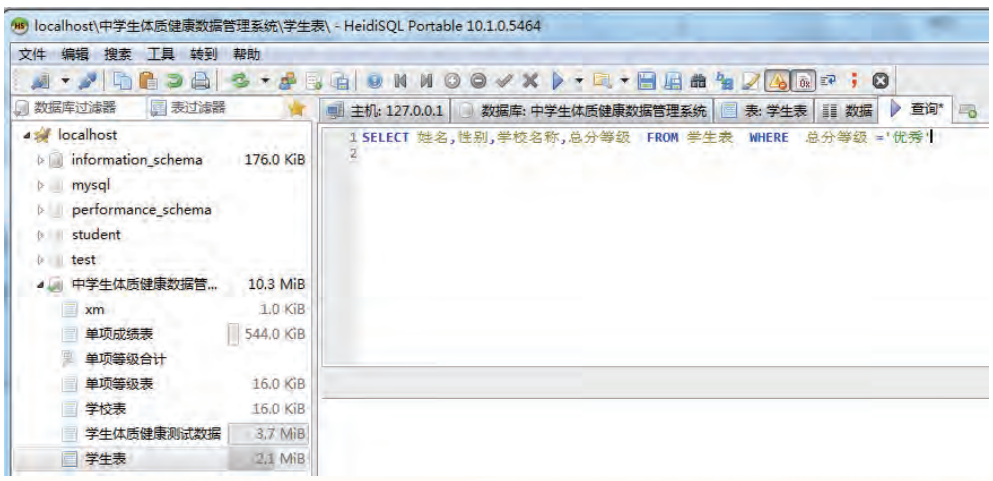


图3-37　输入SELECT查询语句

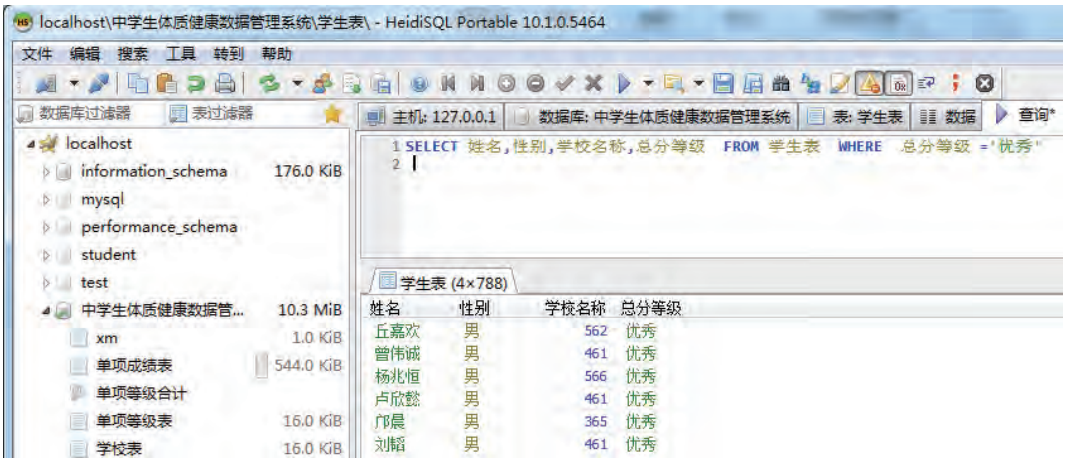


图3-38　执行SELECT查询语句结果

如果要显示查询结果记录的所有字段值呢？应如何修改SELECT语句？

以下通过若干案例帮助同学们理解SELECT查询语句。

【案例1】从单项成绩表中找出单项得分为80～89分的学生的学籍号和项目编号，对应语句为：

```
SELECT 学籍号，项目编号，单项得分
FROM 单项成绩表
WHERE 单项得分 Between 80 And 90
```

【案例2】从单项成绩表中找出单项得分为80～89分的学生的学籍号和项目编号，查询结果按项目编号降序排列，对应语句为：

```
SELECT 学籍号，项目编号，单项得分
FROM 单项成绩表
WHERE 单项得分 Between 80 And 90
ORDER BY 项目编号 DESC
```

【案例3】统计每个项目中各等级的人数，对应语句为：

```
SELECT Count(学籍号) AS 人数，项目编号，单项等级
FROM 单项成绩表
GROUP BY 项目编号，单项等级
ORDER BY 项目编号 DESC
```

案例3使用了SQL语言中的Count()函数，如前所述，Sum函数、Avg函数、Max函数和Mix函数等都是比较常用的函数，各函数的功能和数据的统计查询中汇总功能的选项一致。SQL函数经常与Select语句的Group by子句一同使用。

（2）多表查询。

多表查询是指查询过程中涉及两个或以上的表。

【案例1】查询所有学生的姓名及参加各项目相对应的单项等级，对应语句为：

```
SELECT 学生表.姓名，单项成绩表.项目编号，单项成绩表.单项等级
```

```
FROM 学生表 INNER JOIN 单项成绩表 ON 学生表.学籍号=单项成绩表.学籍号
```

【案例2】查询获得“单项等级”为“良好”的学生的姓名及其参加的项目，并按项目编号降序排列，对应语句为：

```
SELECT 学生表.姓名，单项成绩表.项目编号，单项成绩表.单项等级
FROM 学生表 INNER JOIN 单项成绩表 ON 学生表.学籍号=单项成绩表.学籍号
WHERE 单项成绩表.单项等级 = '良好' ORDER BY 单项成绩表. 项目编号 DESC
```

拓展

SQL命令还可以完成更多的操作，如改变数据库环境设置、针对某个数据库或表格授予用户存取权限、对数据库表格建立索引值等等， 若你想成为一个优秀的数据管理员，可以进一步地深入学习SQL语言。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，选用适当的方法对数据库进行查询。

1. 体验数据库基本的查询方法：选择查询、投影查询、排序查询、统计查询等。
2. 使用结构化查询语句SQL查询数据。

3.3 数据的备份与恢复

人们在进行数据库的增加、删除、修改、查询及统计等操作时，可能由于种种因素，使得原有的数据受到损害或丢失，造成的损失有时候可能无法挽回。因此，数据的备份与恢复是数据管理不可缺少的一个重要组成部分。

3.3.1 数据丢失的风险及原因

探究活动

阅读

【案例1】一份来自某公司的《全球数据保护研究报告》汇总了来自24个国家及地

区，超过3000名受访者的反馈，通过这份翔实的数据推算，2015年全球企业因数据丢失或者宕机造成的损失高达1.7万亿美元，并且自2012年起，企业数据丢失数量及造成的经济损失还处在飞速增长的状态，三年来增速超过了400%。

【案例2】2017年2月1日凌晨，某著名网站的一位工程师在维护数据时不慎删除约300GB的数据。这次事故导致服务长时间中断，网站还永久损失了部分生产数据，无法恢复。更严重的是，网站甚至损失了数据库的相关记录数据，包括项目、注释、用户账户、问题和代码段，即使是乐观地估计，本次事故也影响到约5000个项目，5000个评论和700个新用户账户。

以小组为单位，谈谈你认为数据丢失会带来的风险，以及通常会是什么原因造成数据丢失。

数据丢失的例子层出不穷。对于个人来说，可能损失的是时间；可对于很多公司来说，丢失数据就等于损失资金，甚至破产；而对于国家重要部门，丢失数据关系到国计民生，所造成的损失是无法弥补与估量的。

随着互联网的高速发展，各行各业的信息（包括财务、客户信息等重要数据）进入计算机网络系统中，计算机系统已经成为政府和各企事业单位的基础设施，因此，社会对于信息系统的依赖性逐渐增大。同时，由于对信息保护缺乏相应的理解，因系统功能不正常、人为错误、设备故障、计算机病毒、自然灾害乃至信息攻击及其他不可预测因素等所带来的系统间断、数据丢失等灾难性事故频频发生，信息系统在人为攻击和自然灾害面前的脆弱性日益显现出来。随着信息化程度的加深，数据安全越来越被人们所关注。

数据丢失的原因比较多，一般可分为人为原因、自然原因、软件原因和硬件原因。

1. 人为原因

人为原因主要是指由于使用人员的误操作造成的数据被破坏，如误格式化或误分区、误克隆、误删除或覆盖、环境的潮湿、经常不正常退出、人为地摔坏或磕碰硬盘等。

人为原因造成的数据丢失现象一般表现为操作系统丢失、无法正常启动系统、磁盘读写错误、找不到所需要的文件、文件打不开、文件打开后乱码、硬盘没有分区、提示某个硬盘分区没有格式化、硬盘被强制格式化、硬盘无法识别或发出异响等。

2. 自然原因

自然原因主要是指由于自然灾害造成的数据被破坏，如水灾、火灾、雷击、地震等造成计算机系统的破坏，导致存储数据被破坏或完全丢失，或由于操作时断电、意外电磁干扰造成数据丢失或破坏。

自然原因造成的数据丢失现象一般表现为硬盘损坏、硬盘无法识别、磁盘读写错误、找不到所需要的文件。

3. 软件原因

软件原因主要是指由于受病毒感染、零磁道损坏、硬盘逻辑锁、系统错误或瘫痪造成

文件丢失或破坏。

软件原因造成的数据丢失现象一般表现为操作系统丢失、无法正常启动系统、磁盘读写错误、找不到所需要的文件、文件打不开、文件打开后乱码、硬盘没有分区、提示某个硬盘分区没有格式化、硬盘被锁等。

4. 硬件原因

硬件原因主要是指由于计算机设备的硬件故障（包括存储介质的老化、失效）、磁盘划伤、磁头变形、磁臂断裂、磁头放大器损坏、芯片组或其他元器件损坏等造成数据丢失或破坏。

硬件原因造成的数据丢失现象一般表现为系统不认硬盘，常有一种“咔嚓咔嚓”或“哐当哐当”的磁阻撞击声，或电机不转、通电后无任何声音、磁头定位不准造成读写错误等现象。

同学们阅读《中华人民共和国网络安全法》，了解国家关于信息安全的一些策略。

3.3.2 常见的数据备份与恢复方法

以小组为单位，思考你所知道的数据备份与恢复方法。

上面提到了因人为误操作、自然灾害、软硬件损坏等带来的数据安全风险，而备份软件或命令可以通过简单的部署、操作以及定时自动备份等特性很好地规避这些问题。在系统运转正常时将系统和重要数据做一次完整备份，以后定期做增量备份，当灾难或操作失误出现时可以恢复系统和文件到指定的备份时间点，用户丢失的将只是灾难发生时到最近一次备份时间点之间的部分更新数据，以最大限度地降低用户数据丢失风险。所以，要预防数据的丢失，首先要有数据安全的意识，然后要做好数据的备份。

1. 数据备份概述

数据备份（Data backup）就是将数据加以保留，保存数据的副本到备份设备，以便在系统遭受破坏或其他特定情况下，重新加以利用进行系统恢复的一个过程。

数据恢复就是将数据恢复到事故之前的状态。数据恢复总是与备份相对应，实际上可以将其看成是备份操作的逆过程。备份是恢复的前提，恢复是备份的目的，无法恢复的备份是没有意义的。

数据备份并非简单的文件复制，多数指数据库备份，是数据库结构和数据的复制，以便在数据库遭到破坏时进行恢复。备份内容包括用户数据库和系统数据库内容。

2. 数据备份类型

备份是指对数据库事务日志进行复制，数据库备份记录了在进行备份操作时数据库中所有数据的状态。若数据库因意外而损坏，这些备份文件在数据库恢复时用于还原数据库。数据备份的类型有多种划分方式，在不同情况下应选择最合适的方式。

（1）完全备份：对服务器上的所有数据进行完全备份，包括系统文件和数据文件，并不依赖文件的存档属性来确定备份哪些文件。在备份过程中，任何现有的标记都被清除，每个文件都被标记为已备份，换言之，清除存档属性。

完全备份所需时间最长，但恢复时间最短，操作最方便。当系统中的数据量不大时，采用完全备份最可靠。

（2）增量备份：只对上一次备份后增加的和修改过的数据进行备份。增量备份过程中，只备份有标记的选中的文件和文件夹。备份后，它清除标记，即备份后清除存档属性。

由于没有重复的备份数据，既节省磁盘空间，又缩短了备份时间。但是一旦发生灾难，恢复数据则比较麻烦，因而实际应用中一般较少采用这种方式。

（3）差异备份：对上一次完全备份（而不是上次备份）之后新增加的和修改过的数据进行备份。差异备份过程中，只备份有标记的选中的文件和文件夹。它不清除标记，即备份后不标记为已备份文件，换言之，不清除存档属性。

差异备份在恢复数据时，需两份数据，一份是上一次的完全备份，另一份是最新的差异备份。

完全备份、增量备份以及差异备份之间的比较如表3-4所示。

表3-4　完全备份、增量备份以及差异备份之间的比较

类别	操作特点	优点	缺点
完全备份	备份系统中的所有数据。	恢复时间短，最可靠，操作方便。	数据量大，备份时间长，空间消耗大。
增量备份	备份上一次备份以后更新的所有数据。	数据量不大，备份时间短，空间消耗小。	恢复操作复杂，需要一个完全备份，以及其后的所有增量备份。
差异备份	备份上一次完全备份以后更新的所有数据。	优缺点介于完全备份和增量备份两者之间。	

（4）定时备份：在固定的时间间隔进行数据备份的方式，不能保证数据零丢失。

（5）实时备份：在任意时间间隔进行数据备份的方式，可以保证数据零丢失。

对于操作系统和应用程序代码，可在每次系统更新或安装新软件时做一次完全备份。

对于日常更新量很大，但总体数据量不是很大的关键应用数据库，可在每天用户使用量较小的时段安排做完全备份。

对于总体数据量很大，但日常更新量相对较小的关键应用数据库，可每隔一周或更长时间做一次完全备份，而每隔一个较短的时间（如每天）做一次增量备份。

时间做一次完全备份，而每隔一个较短的时间（如每天）做一次增量备份。

体验一次数据备份与恢复操作。

我们的数据库刚建立完善，因此这次体验选择完全备份。利用HeidiSQL实现MariaDB软件中数据备份与恢复功能。

1. 备份。

（1）选择“工具”菜单中的“导出数据脚本”，如图3-39所示。

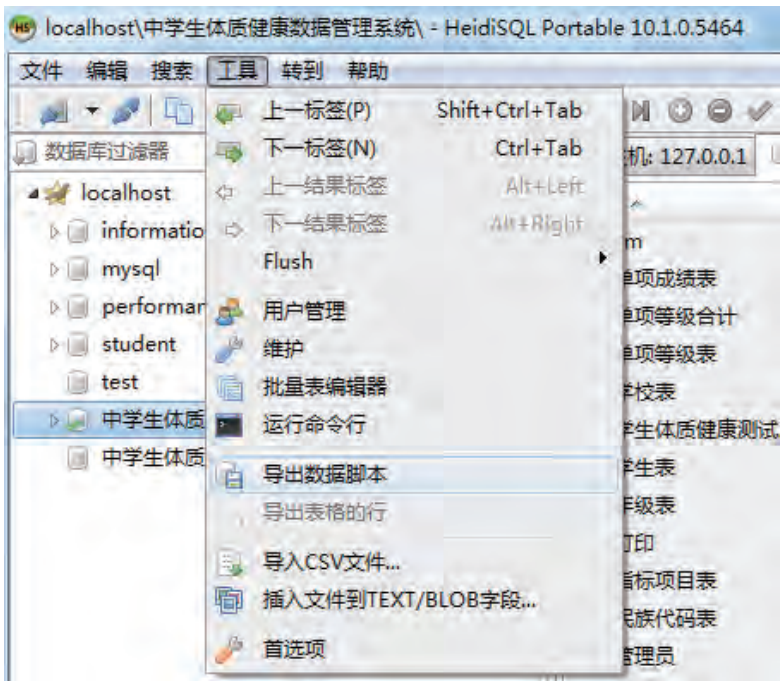


图3-39 导出数据脚本

（2）在“表工具”窗口中，选择需要备份的数据库，按照如图3-40所示设置参数，然后点击“导出”按钮，即可完成数据库的备份，如图3-41所示。

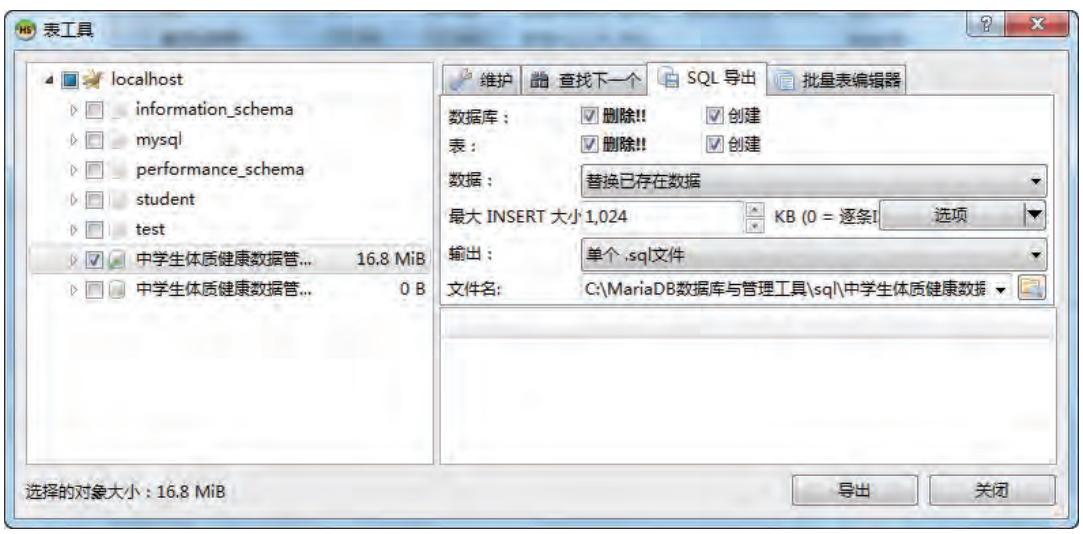


图3-40 进行备份时的参数设置

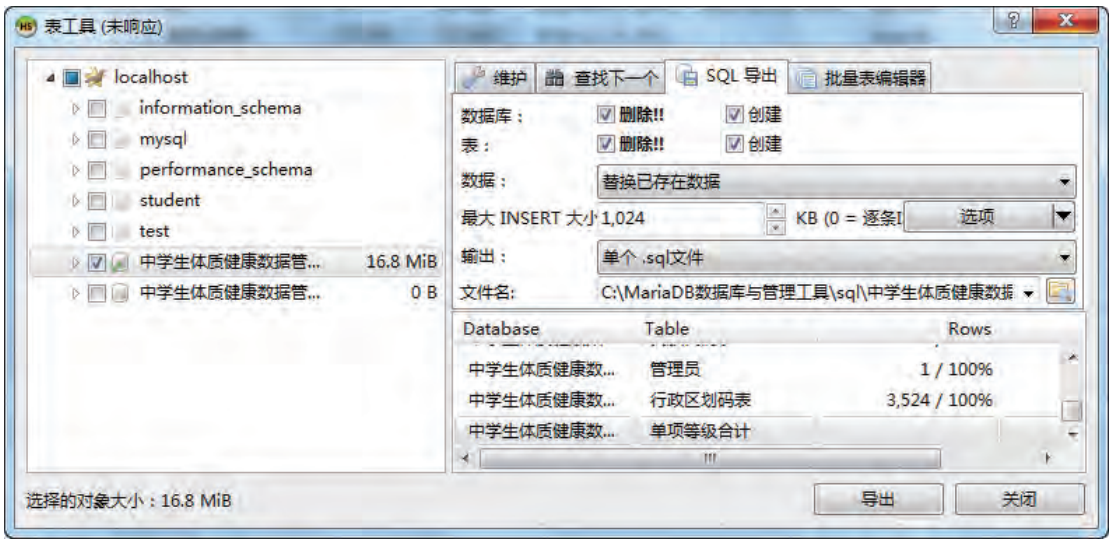


图3–41　完成备份

2. 还原。

在HeidiSQL中通过运行SQL文件来完成数据库的还原操作。具体操作步骤如下：

（1）选择需要还原的数据库，然后在“文件”菜单中选择“运行SQL文件”，如图3–42所示。

（2）在“打开”窗口中选择要运行的SQL文件，如图3–43所示。点击“打开”按钮即可开始运行还原，如图3–44所示。

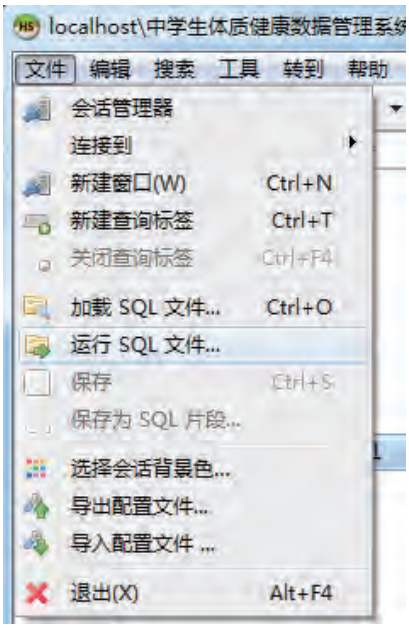


图3–42　还原备份

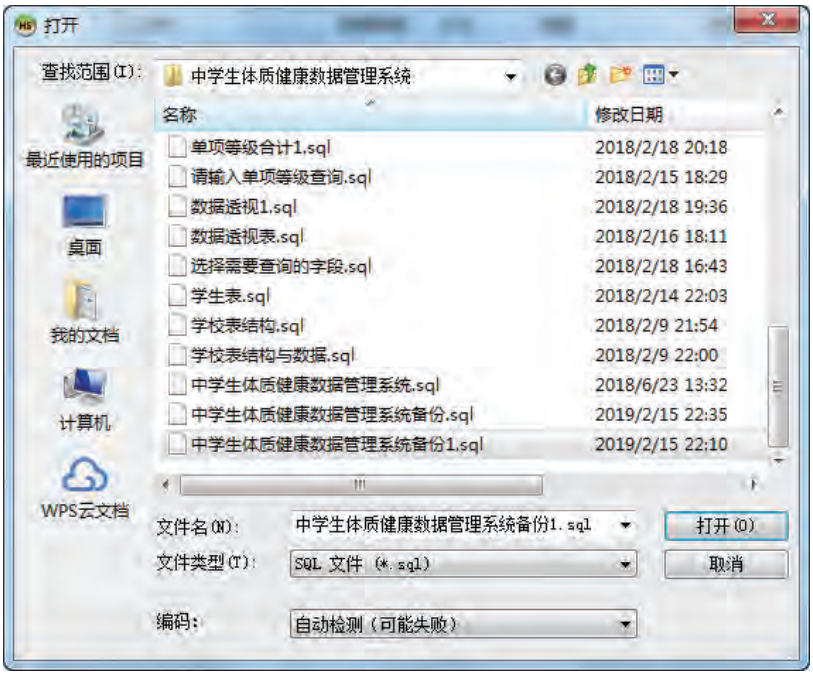


图3–43　还原对象选择

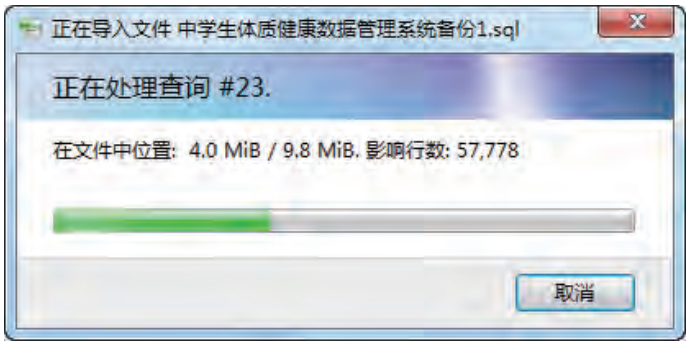


图3–44　进行还原备份

拓展

云存储属于数据备份吗？请同学们了解云存储的相关内容。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，认识数据丢失的风险，实践数据备份，并参照项目范例的样式，撰写相应的项目成果报告。

成果交流

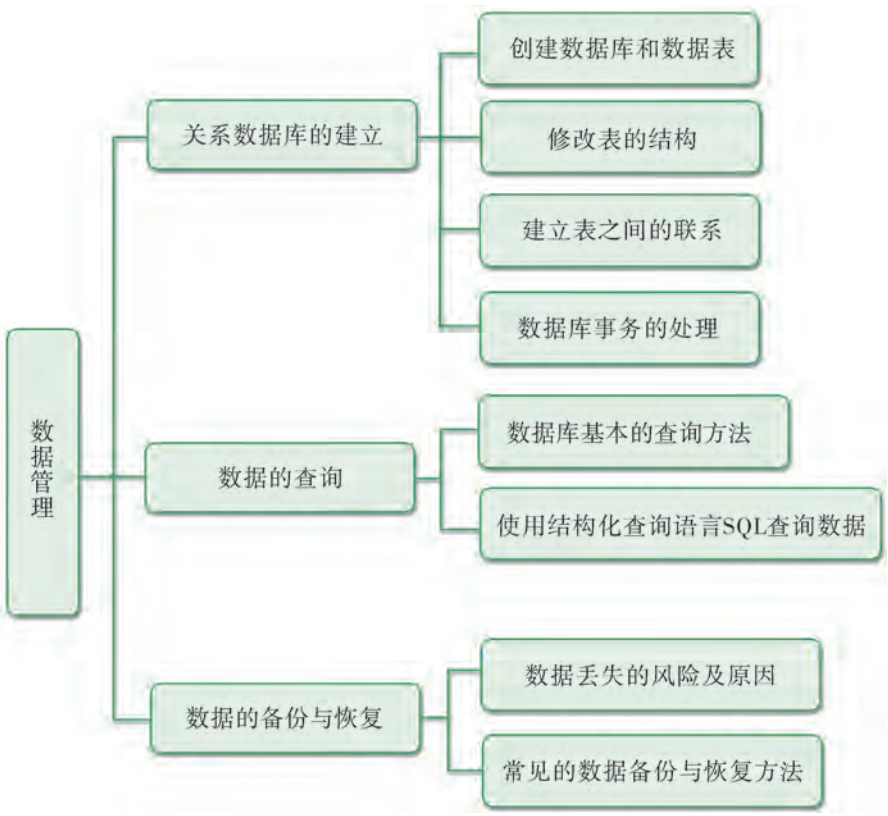
各小组运用数字化学习工具，将所完成的项目成果，在小组或班级上进行展示与交流，共享创造、分享快乐。

活动评价

各小组根据项目选题、拟订的项目方案、实施情况以及所形成的项目成果，利用教科书附录2的“项目活动评价表”，开展项目学习活动评价。

本章扼要回顾

同学们通过本章学习，根据“数据管理”知识结构图，扼要回顾、总结、归纳学过的内容，建立自己的知识结构体系。



回顾与总结

本章学业评价

同学们完成下列测试题（更多的测试题可以在教科书的配套学习资源包中查看），并通过“本章扼要回顾”以及本章的项目活动评价，综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。

1. 单选题

（1）在数据库数据查询中，投影查询是从一个关系中选出若干指定（　　）的值的操作。

A. 记录　　B. 字段　　C. 表　　D. 主键

（2）若要为查询检索到的数据添加排序，则需要在查询中添加（　　）选项。

A. GROUP BY　　B. WHERE　　C. ORDER BY　　D. FROM

（3）以下不属于数据库备份的方法是（　　）。

A. 完全备份　　B. 差异备份　　C. 文件复制　　D. 增量备份

2. 思考题

根据本章所学的知识，谈谈你对数据丢失风险的认识，以及有什么积极的措施可以尽量降低数据丢失所带来的损失。

3. 情境题

某公司的数据管理系统包括雇员、部门、项目、资产等方面的数据信息，以下是其中的四个数据表：

部门（部门名称，办公室门牌，联系电话）；

员工（员工编号，姓名，所属部门，联系电话，联系邮箱）；

工程项目（项目编号，项目名称，负责部门，开始时间，完成时间）；

项目分配（项目编号，员工编号，工作时长）。

（1）如果要查询在“财务部”工作，且电话号码为“13266666666”的员工信息，如何用SQL进行查询？

（2）如果想确定在某个工程项目中工作时间超过50小时的所有员工的姓名，如何用SQL进行查询？

第四章
数据分析

有效地分析数据可以帮助人们从数据资源中提取隐藏在数据背后有价值的信息，发现数据的意义，为人们做出判断、形成决策提供依据。

本章将通过“数据管理系统的数据分析”项目，进行自主、协作、探究学习，让同学们认识噪声数据现象及其对分析数据、获取有价值的信息、形成决策可能产生的影响；使用结构化查询语言进行数据查询；根据需要，选择恰当的方法进行数据提取；了解常用的数据分析方法，例如，对比分析法、分组分析法、平均分析法和相关分析法等；在实践中选用适当的数据分析工具，分析、呈现并解释数据，了解数据分析的常用方法和工具，从而将知识建构、技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，促进信息技术学科核心素养达成，完成项目学习目标。

- 数据分析概述
- 数据处理
- 描述性分析
- 数据的可视化表达

项目范例 中学生体质健康数据管理系统的数据分析

情境

对“中学生体质健康数据管理系统”的数据进行统计、分析，能够为学生提供个性化的身体健康诊断，使学生能够在准确地了解自己体质健康状况的基础上进行锻炼，还可为各级政府机关、教育行政部门、学校提供翔实的统计和分析数据，使之了解学生的体质健康状况，及时采取科学的干预措施。通过描述性分析，可以得到学生身体形态、身体机能和身体素质状况，从而综合评价学生的体质健康水平；通过相关分析和回归分析，得到学生身高、体重、体育活动时间与体质的相关关系并预测学生个体和群体的体质健康发展趋势；通过聚类分析对学生进行分组，为学生提高体能测试成绩的训练方案制订提供有效的依据；最后我们将会体验大数据分析解决问题的基本思路。

主题

中学生体质健康数据管理系统的数据分析

规划

根据项目范例的主题，在小组中组织讨论，利用思维导图工具，制订项目范例的学习规划，如图4-1所示。

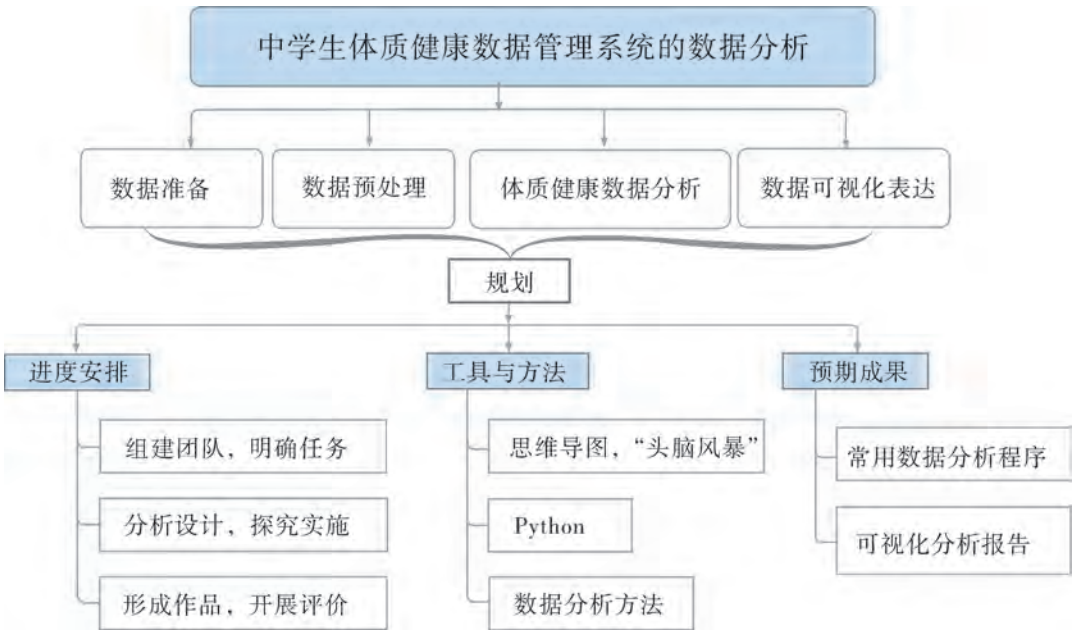


图4-1 “中学生体质健康数据管理系统的数据分析”项目学习规划

探究

根据项目学习规划的安排，通过调查、案例分析、文献阅读和网上资料搜索，开展“中学生体质健康数据管理系统的数据分析”项目学习探究活动，如表4-1所示。

表4-1　“中学生体质健康数据管理系统的数据分析”项目学习探究活动

探究活动	学习内容		知识技能
数据准备	数据的导入和导出。	明确分析任务，设计分析思路。	在实践中选用适当的数据分析工具，呈现数据。
数据预处理	数据合并、计算、分组。	抽取管理系统数据进行清洗。	认识噪声数据的现象和成因。 能利用适当的工具对数据进行采集和分类。
体质健康数据分析	常用的数据分析方法。	数据分析与建模。	了解常用的数据分析方法。
数据可视化表达	数据可视化表达常用图表。	组织数据，展现数据。	在实践中选用适当的数据分析工具，分析、呈现并解释数据。
	回归分析的实现。	结果预测与推论。	
	聚类分析的实现。		

实施

实施项目学习各项探究活动，进一步认识中学生体质健康数据管理系统的数据分析。

成果

在小组开展项目范例学习的过程中，利用思维导图工具梳理小组成员在“头脑风暴”活动中的观点，建立观点结构图，运用多媒体创作工具（如演示文稿、在线编辑工具等），综合加工和表达，形成项目范例可视化学习成果，并通过各种分享平台发布，共享创造、分享快乐。例如，运用在线编辑工具制作的“中学生体质健康数据管理系统的数据分析”可视化报告，可以在教科书的配套学习资源包中查看，其目录截图如图4-2所示。

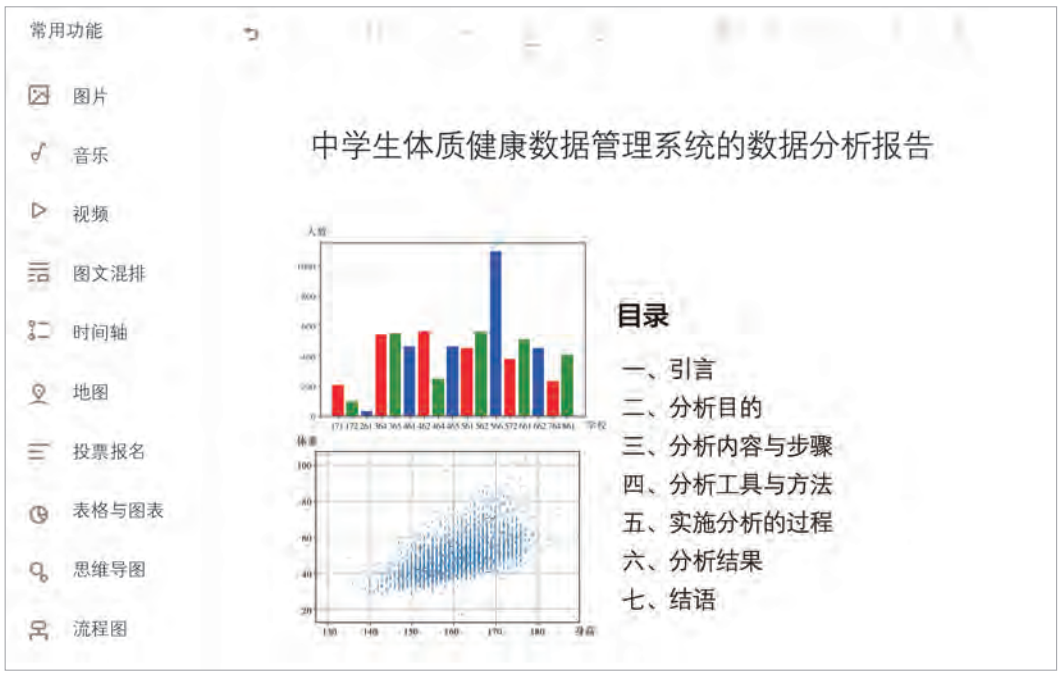


图4–2　“中学生体质健康数据管理系统的数据分析”可视化报告的目录截图

评价

根据教科书附录2的“项目活动评价表”，对项目范例的学习过程和学习成果在小组或班级上进行交流，开展项目学习活动评价。

项目选题

同学们以3～6人组成一个小组，选择下面一个参考主题，或者自拟一个感兴趣的主题，开展项目学习。

1. 图书馆图书借阅管理系统的数据分析
2. 社团活动信息管理系统的数据分析
3. 学生“一卡通”管理系统的数据分析

项目规划

各小组根据项目选题，参照项目范例的样式，利用思维导图工具，制订相应的项目方案。

方案交流

各小组将完成的方案在全班进行展示交流，师生共同探讨、完善相应的项目方案。

4.1 数据分析概述

生活中，数据无处不在。“量化自我”的生活方式记录心率、饮食习惯、作息习惯，智能汽车记录行驶习惯，智能家居记录生活习惯。互联网是一个包罗万象的知识谱系，包括新闻、电影、音乐等专业数据库，看似杂乱无章的社交信息，交叉引用的百科全书，都包含着大量的数据信息。

从大量的数据中发掘有用的信息，揭示隐含其中的内在规律，指导科学的推断和决策，需要对纷繁复杂的数据进行分析。数据分析是运用数据分析的工具和方法，根据研究的目的，对数据进行深层次挖掘和分析，找出内在的联系和变化，从而揭示事物的本质状态，预测事物的发展趋势。

4.1.1 数据分析的方法

数据分析涉及的学科领域和方法众多。常见的数据分析方法从现状、原因和预测三大方面开展，数据分析目的不同，选用的分析方法也不一样。如图4-3所示是常用数据分析的一些具体方法。

图4-3　常用数据分析方法

探究活动

讨论

在“中学生体质健康数据管理系统的数据分析”项目中，利用如图4-3所示的方法，进行现状、原因和预测三个方面的具体数据分析，如图4-4所示。以小组为单位进行讨

论，确定学生身体形态、身体机能和身体素质状况的描述的分析方法，学生饮食习惯、体育活动时间与体质关系的分析方法，以及预测学生个体和群体体质健康发展趋势的分析方法。

图4–4 “中学生体质健康数据管理系统的数据分析”三大方面

对学生身体形态、身体机能和身体素质状况的描述属于现状分析。现状分析方法包括对比分析法、平均分析法和综合评价分析法。通过相关分析和回归分析，可以得到学生饮食习惯、体育活动时间与体质的相关关系，并预测学生个体和群体的体质健康发展趋势。原因分析和预测分析中较为常用的是分组分析、交叉分析、关联分析、聚类分析和回归分析。

4.1.2 数据分析的工具

数据分析需要处理大量的数据，进行复杂的运算，因此数据分析软件的使用是必不可少的。数据分析的工具数量众多，根据分析数据层次结构的不同，常用数据分析软件可分为四类，如图4–5所示。

图4–5 常用数据分析工具

本章将学习使用Python的Pandas数据分析包进行数据分析，还将使用Numpy科学计算模块和Matplotlib绘图模块。

体验

同学们上网查找“中学生体质健康数据分析图”，如图4-6所示。

图4-6　中学生体质健康数据分析图

4.1.3 数据导入

数据存在的形式多样，有文件和数据库形式。在进行数据分析前需要从数据库或者现有的数据文件中提取符合要求的数据。

（1）导入TXT文件：read_table（file,names=[列名1，列名2，...],sep="",...）。

（2）导入CSV文件：read_csv（file,names=[列名1，列名2，...],sep="",...）。

导入TXT文件和CSV文件的参数说明：file为文件路径和文件名；names为列的名称，默认文件中第一行为列名；sep为分隔符，默认为空，表示整列导入。

（3）导入Excel文件：read_excel（file,sheetname,header=0）。

参数说明：file为文件路径和文件名；sheetname为表格的名称，如Sheet1；header为列名，默认为0，文件的第一行作为列名。

（4）导入MySQL库：read_sql（sql,con=数据库）。

从第三章建立的“中学生体质健康数据管理系统”中导出数据文件，示例代码（本教科书以Python3版本为例，不同版本的函数参数可能会有所不同）如下：

```
import pandas as pd
import numpy as np
import MySQLdb
conn = MySQLdb.connect
    （host='localhost',    #数据库地址
   user='root',          #登录名
   passwd='',     #访问密码，此处无密码
   db='adidas',          #访问的数据库
  prot=5029,           #访问的端口
charset='UTF8'）        #编码格式
data = pd.read_sql（'select * from user',onn=connection）     #user是adidas库中的表
  conn.close（）    #调用完关闭数据库
```

4.1.4 数据导出

数据采集完成后，经过处理，一般存储为Excel表或生成CSV文件，作为二次分析数据使用。Python自带CSV的处理库，CSV文件相比Excel表使用更简单，而且不需要引入第三方库。

（1）导出CSV文件：to_csv（file_path,sep=",",index=True,header=True）。

（2）导出Excel文件：to_excel（file_path,index=True,header=True）。

参数说明：file_path为文件路径；sep为分隔符，默认为空；index，header默认为True，导出索引和列名。

（3）导出到MySQL库：to_sql（tableName,con=数据库链接）。

“中学生体质健康数据管理系统的数据分析”项目需要通过数据统计、分析描述全区学生体质健康整体水平，分析出学校间，男、女生之间是否存在显著差异以及差异成因，分析影响身体素质的相关因素。

以小组为单位，讨论开展“中学生体质健康数据管理系统的数据分析”项目的工作流程，并就以下问题作为提纲制订数据分析方案，分别在小组和全班进行交流。

（1）本小组项目分析的业务需求是什么？

（2）从哪些角度分析数据才能达到要求？

（3）需要用到哪些数据？

体验

导入教科书配套学习资源包“第四章\课本素材\test4-1.xlsx”文件，尝试将“标准分”与“附加分”相加得到“总分”，生成Excel文件并保存。

参考程序代码如下，运行结果如图4-7所示。

```
import pandas as pd
import numpy as np
import xlrd
fileNameStr=(r'd:\第四章\课本素材\test4-1.xlsx')
xls=pd.ExcelFile(fileNameStr)
Df=xls.parse('Sheet1')
total=Df.标准分+Df.附加分
Df['总分']=total
Df.to_excel(r'd:\第四章\课本素材\test4-1-1.xls')
```

```
In [6]: Df
```

Out[6]:

	学校编号	学籍号	姓名	性别	出生日期	身高	体重	肺活量评分	50米跑评分	立定跳远评分	体前屈评分	标准分	附加分
0	561	330801200008209865	俞从丹	女	2000/08/20	160.0	40.0	100	80	85	95	88.2	0
1	561	44050120020815946X	鞠流如	女	2002/08/15	151.0	43.0	80	80	90	80	86.8	0
2	561	371482200008207450	字弘丽	男	2000/08/20	157.0	48.0	80	90	95	80	90.5	2
3	561	44040320010609579X	上官含秀	男	2001/06/09	157.0	42.0	78	85	100	100	91.7	0
4	561	23070720000820289X	友慧雅	男	2000/08/20	156.0	53.0	100	100	90	74	94.2	3
5	561	440501200208157472	车伟懋	男	2002/08/15	160.0	47.0	85	85	95	74	90.2	3

```
In [7]: total=Df.标准分+Df.附加分
        Df['总分']=total
        Df
```

Out[7]:

	学校编号	学籍号	姓名	性别	出生日期	身高	体重	肺活量评分	50米跑评分	立定跳远评分	体前屈评分	标准分	附加分	总分
0	561	330801200008209865	俞从丹	女	2000/08/20	160.0	40.0	100	80	85	95	88.2	0	88.2
1	561	44050120020815946X	鞠流如	女	2002/08/15	151.0	43.0	80	80	90	80	86.8	0	86.8
2	561	371482200008207450	字弘丽	男	2000/08/20	157.0	48.0	80	90	95	80	90.5	2	92.5
3	561	44040320010609579X	上官含秀	男	2001/06/09	157.0	42.0	78	85	100	100	91.7	0	91.7
4	561	23070720000820289X	友慧雅	男	2000/08/20	156.0	53.0	100	100	90	74	94.2	3	97.2
5	561	440501200208157472	车伟懋	男	2002/08/15	160.0	47.0	85	85	95	74	90.2	3	93.2

图4-7　导入文件处理后生成Excel文件

程序运行后，查看“d：\第四章\课本素材”文件夹下生成的“test4-1-1.xls”文件。

项目实施

各小组根据项目选题及拟订的项目方案，并结合本节所学知识，了解相关的数据分析方法与工具，完成项目分析的数据准备工作。

1. 讨论并体验数据分析方法与工具。
2. 体验数据导入与导出的操作。

4.2 数据处理

在数据采集过程中，由于数据的设备可能出现故障，数据输入以及数据传输的过程中可能出现错误，存储介质有可能出现损坏等，导致需要用于数据分析的数据可能不完整、包含错误值或者数据内涵不一致等情况。在数据分析前需要对数据进行处理，剔除其中噪声、恢复数据的完整性和一致性后才能进行数据分析。

4.2.1 数据清洗

1. 重复数据的处理

数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录。合并、清除是处理重复数据的基本方法。

使用duplicated( )可以获取哪些是重复的元素，使用drop_duplicates( )能够删除重复元素。

2. 缺失数据的处理

缺失值是数据中经常出现的问题，也是任何数据采集过程中可能出现的问题，如阅卷中无回答、回答错误、录入错误等现象都会导致缺失数据。缺失值会影响分析工作的进行，还会导致分析的偏差。

缺失值的处理包括两个步骤，即缺失数据的识别和缺失值处理。Python中缺失值通常以NaN表示，可以使用函数isnull( )判断缺失值是否存在。缺失值处理常用的方法有删除法、替换法、插补法等。

（1）删除法。

删除法是最简单的缺失值处理方法，根据数据处理的不同角度可分为删除观测样本、删除变量两种。在Python中可通过dropna( )函数移除所有含有缺失数据的行，这属于以减少样本

量来换取信息完整性的方法，适用于缺失值所占比例较小的情况；删除变量适用于该变量中有较多缺失且对研究目标影响不大的情况，通过dropna(axis=1)来实现整个变量的删除。

（2）替换法。

变量按属性可分为数值型和非数值型，二者的处理办法不同：如果缺失值所在变量为数值型，一般用该变量在其他所有对象的取值的均值来替换变量的缺失值；如果为非数值型变量，则使用该变量其他全部有效观测值的中位数或者众数进行替换。

（3）插补法。

删除法虽然简单易行，但会带来数据资源浪费和改变数据结构的问题，因此在条件允许的情况下，找到缺失值的替代值来进行插补，尽可能还原真实数据是更好的方法。简单的插补法可以采取前后的数据值、变量均值、中位数等其中之一来代替缺失值。

缺失值处理用到的主要工具为 Numpy 库和 Pandas库中的有关函数，表4-2中列出处理缺失数据的相关函数。

表4-2　处理缺失数据的相关函数

函数名称	使用说明
isnull( )	是缺失值返回True，否则返回False。
isnull( ).sum( )	返回每列包含的缺失值的个数。
dropna( )	删除含有缺失值的行。
dropna(axis=1)	删除含有缺失值的列。
dropna(how='all')	删除全是缺失值的行。
dropna(thresh=4)	保留至少有4个缺失值的行。
fillna('?')	使用“?”替代缺失值。
fillna(method='pad')	用前一个数据值替代缺失值。
fillna(method='bfill')	用后一个数据值替代缺失值。
fillna(df.mean( ))	用平均数替代缺失值。

探究活动

实践

导入教科书配套学习资源包“第四章\课本素材\test4-2.xlsx”文件，寻找其中的缺失值，并使用中位数替代身高字段中的缺失值，关键程序代码如下：

```
fileNameStr=(r' 第四章\课本素材\test4-2.xlsx')
xls=pd.ExcelFile(fileNameStr)
Df=xls.parse('Sheet1')
Df
Df.fillna(Df.median())
```

3. 噪声数据的处理

噪声数据是指数据中存在着错误或异常（偏离期望值）的数据。数据分析工具都有寻找噪声数据的函数，函数通过寻找数据集中与其他观测值及均值差距最大的点作为异常值。

在进行噪声数据检查后，不可以直接使用删除方式处理异常值，因为有可能孤立点的数据正是实验要找出的异常数据。在实际操作中常用分箱（binning）、回归（regression）、聚类（clustering）、计算机与人工检查相结合等方法“光滑”数据，去掉数据中的噪声。

（1）分箱。

分箱是指通过对数据进行排序，利用数据“近邻”（即周围的值）来光滑有序数据值的一种局部光滑方法。在分箱方法中，可以使用箱均值、箱中位数或箱边界等进行光滑。对于用箱均值或箱中位数光滑，可以使用平均值或中位数代替箱中的噪声数据；而对于箱边界平滑，将给定箱中最大值或最小值视为箱边界，箱中的噪声数据被替换为最近边界值。

Python中cut函数使用数值区间将数值分箱，用在长度相等的桶；qcut使用分位数将数值分箱，用在大小相等的桶。

（2）回归。

通过回归函数拟合数据来光滑数据。回归包括线性回归和多元回归。线性回归涉及找出拟合两个变量的“最佳”直线，使得一个属性数据预测另一个数据。如图4-8所示，建立线性回归方程y=ax+b，通过体重预测得到身高的值。

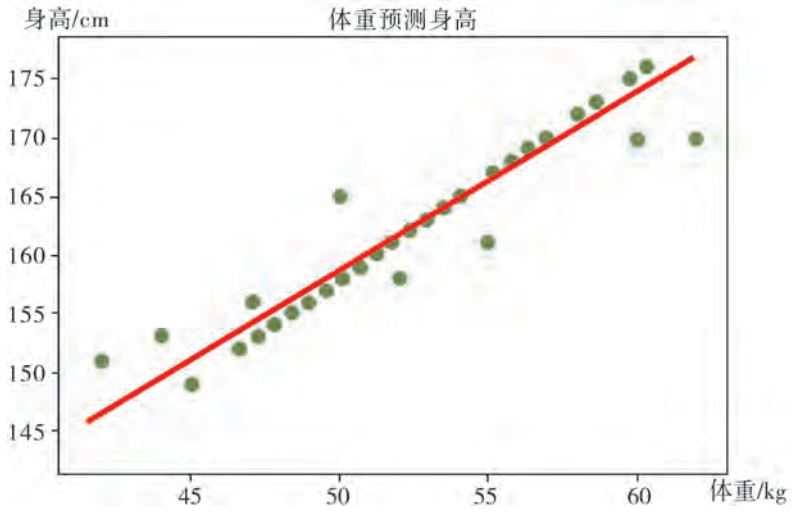


图4-8 线性回归处理噪声数据

（3）聚类。

通过聚类识别噪声数据后，考察噪声在各个属性上的值与其期望之间的距离以判定引起噪声的属性，利用所属分类中噪声属性上的值对噪声数据进行矫正，如图4-9所示，将数据组织成3簇，落在簇集合之外的值视为离群点，可以使用与属性相近的数据进行平滑。

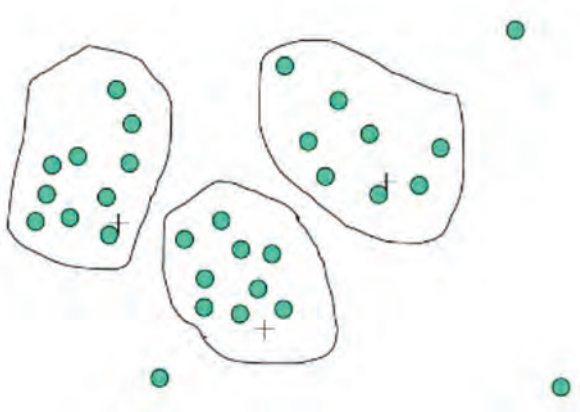


图4-9　聚类处理噪声数据

4.2.2　数据的合并

1. 纵向合并

数据的纵向合并是指将多个结构相同的数据框合并成一个数据框，如图4-10所示，将df1，df2，df3三个数据框合并成Result：

```
Result=pandas.concat([df1,df2,df3])
```

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

Result

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

图4-10　纵向合并数据

2. 横向合并

数据的横向合并是指不同结构的数据框，按照一定的条件进行合并，如图4-11所示，将df1与s1合并成Result，并保留索引。

```
Result=pandas.concat([df1, s1], axis=1)
```

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

s1

	X
0	X0
1	X1
2	X2
3	X3

Result

	A	B	C	D	X
0	A0	B0	C0	D0	X0
1	A1	B1	C1	D1	X1
2	A2	B2	C2	D2	X2
3	A3	B3	C3	D3	X3

图4-11 横向合并数据

观察

观察比较语句1和语句2在合并数据结果上的差异，了解concat函数参数的作用。

语句1：Result=pd.concat([df1,df4],axis=1)，执行结果如图4-12所示。

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	B	D	F
0	A0	B0	C0	D0	NaN	NaN	NaN
1	A1	B1	C1	D1	NaN	NaN	NaN
2	A2	B2	C2	D2	B2	D2	F2
3	A3	B3	C3	D3	B3	D3	F3
6	NaN	NaN	NaN	NaN	B6	D6	F6
7	NaN	NaN	NaN	NaN	B7	D7	F7

图4-12 语句1执行结果示意图

语句2：Result=pd.concat([df1,df4],axis=1,join='inner')，执行结果如图4-13所示。

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	B	D	F
2	A2	B2	C2	D2	B2	D2	F2
3	A3	B3	C3	D3	B3	D3	F3

图4-13 语句2执行结果示意图

讨论

在小组中讨论交流，谈谈concat函数与merge函数在合并数据使用上的异同，函数参数如表4-3所示。

1. pd.concat(objs,axis=0,join='outer',join_axes=None,ignore_index=False,keys=None,levels=None,names=None,verify_integrity=False,copy=True)

表4-3　concat函数参数说明

参数	使用说明
objs	参与连接的列表或字典，且列表或字典里的对象是pandas数据类型。
axis	指明连接的轴向，0表示纵轴，1表示横轴，缺省默认为0。
join	'Inner'为交集，'outer'为并集，缺省默认为 'outer'。
join_axes	层次化索引，某个轴向有多个索引，不执行交并集。
ignore_index	不保留连接轴上的索引，产生一组新索引range。
keys	与连接对象有关的值，用于形成连接轴向上的层次化索引。
levels	指定用作层次化索引各级别上的索引。
names	设置keys或levels时，用于创建分层级别的名称。
verify_integrity	检查结果对象新轴上的重复情况，默认False，允许重复。

2. pd.merge(left,right,how='inner',on=None,left_on=None,right_on=None,left_index=False,right_index=False,sort=True,suffixes=('_x','_y'),copy=True,indicator=False)

示例：

```
Result=pd.merge(left,right,on='key')
```

执行结果如图4-14所示。

left

	key	A	B
0	K0	A0	B0
1	K1	A1	B1
2	K2	A2	B2
3	K3	A3	B3

right

	key	C	D
0	K0	C0	D0
1	K1	C1	D1
2	K2	C2	D2
3	K3	C3	D3

Result

	key	A	B	C	D
0	K0	A0	B0	C0	D0
1	K1	A1	B1	C1	D1
2	K2	A2	B2	C2	D2
3	K3	A3	B3	C3	D3

图4-14　执行结果示意图

1. 打开教科书配套学习资源包“第四章\课本素材\4.2纵向合并文件”文件夹，尝试使用多种方法将文件夹下5个学校的数据文件合并为1个文件。

2. 打开教科书配套学习资源包“第四章\课本素材\4.2横向合并文件”文件夹，尝试使用多种方法将文件夹下2个数据文件合并为1个文件。

3. 导入教科书配套学习资源包“第四章\课本素材\test4-3.xlsx”文件，将“引体向上

（男）评分”和“1分钟仰卧起坐（女）评分”两项合并成“体能评分”，将“1000米跑（男）评分”和“800米跑（女）评分”两项合并成“长跑评分”。

4.2.3 数据的计算

数据计算是对原有的字段计算或者转换，形成数据分析所需要的新数据字段。

根据《国家学生体质健康标准》单项指标与权重说明，学生的标准分由各单项成绩加权计算得出，体重指数（BMI）、肺活量、50米跑、坐位体前屈、立定跳远、引体向上（男）/1分钟仰卧起坐（女）、1000米跑（男）/800米跑（女）权重分别为0.15，0.15，0.2，0.1，0.1，0.1，0.2。总分则由标准分与附加分相加得到。

导入教科书配套学习资源包“第四章\课本素材\test4-4.xlsx”文件，根据单项指标成绩计算学生的标准分和总分，关键程序代码如下：

```
total=0.15*Df['体重指数评分']+ 0.15*Df['肺活量评分']+ 0.2*Df['50米跑评分']+
      0.1*Df['坐位体前屈评分']+ 0.1*Df['立定跳远评分']+ 0.1*Df['体能评分']+
      0.2*Df['长跑评分']
Df['标准分']=total
result=Df['标准分']+Df['附加分']
Df['总分']=result
```

4.2.4 数据分组

数据分组是根据数据分析对象的特征，把数据划分为不同的区间来进行研究，以揭示其内在的联系和规律性，数据分组语句如下：

```
cut(series,bins,right=True,labels=NULL)
```

参数说明：series为需要分组的数据，bins为分组的依据，right为分组时右边是否闭合，labels为分组的自定义标签，默认不定义。

导入教科书配套学习资源包“第四章\课本素材\test4-5.xlsx”文件，根据总分进行分组。分组依据：总分90分及以上为优秀；总分80~89分为良好；总分60~79分为合格；总分60分以下为不合格。

项目实施

各小组根据项目选题及拟订的项目方案，并结合本节所学知识，完成数据预处理工作。

1. 实践与体验数据清洗、合并、计算、分组等方法与工具。
2. 选择适当的方法对项目所需数据进行预处理。

4.3 描述性分析

描述性分析重点描述数据的现状特征，主要用于衡量数据的集中趋势、离散趋势以及对数据进行探索性分析，是基础的统计分析手段。通过描述性分析统计可以掌握数据结构，理解变量与变量之间的关系。

4.3.1 基本统计

统计分析往往是从了解数据的基本特征开始的。基本统计常用指标主要包括平均数、众数、中位数、标准差、方差等，描述数据的集中趋势和离散趋势，常用描述性统计分析函数如表4-4所示。

表4-4　常用描述性统计分析函数

函数	使用说明
describe( )	针对series或各dataframe列计算汇总统计。
count( )	计数。
sum( )	列值总和。
mean( )	求列的平均数。
median( )	求列的中位数。
mad( )	根据平均值计算平均绝对离差。
var( )	样本值的方差。
std( )	样本值的标准差。
max( )	列的最大值。

（续表）

函数	使用说明
min( )	列的最小值。
argmin、argmax	获取到最小值和最大值的索引位置。
idxmin、idxmax	获取到最小值和最大值idea索引值。

探究活动

实践

导入教科书配套学习资源包“第四章\课本素材\test4-6.xlsx”文件，对数据名列及总分进行描述性统计分析，关键程序代码如下：

```
pd.options.display.float_format = '{:,.2f}'.format  #保留小数点后两位
Df.describe()
Df.总分.describe()
```

程序执行结果如图4-15所示。

	身高	体重	体重评分	肺活量评分	50米跑评分	立定跳远	立定跳远评分	坐位体前屈	坐位体前屈评分	长跑评分	男女体能评分	标准分	附加分	总分
count	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00	7,237.00
mean	161.09	48.46	95.99	79.27	83.83	192.66	77.23	10.74	74.84	84.91	60.00	81.46	0.49	81.96
std	6.93	8.91	9.43	13.02	12.27	26.23	13.54	6.06	13.27	14.87	24.44	7.59	1.33	8.15
min	130.00	16.80	60.00	0.00	0.00	100.00	0.00	-16.00	0.00	0.00	0.00	29.50	0.00	29.50
25%	156.00	42.10	100.00	72.00	76.00	173.00	70.00	7.00	68.00	78.00	50.00	77.20	0.00	77.30
50%	161.00	47.00	100.00	78.00	80.00	190.00	78.00	11.00	76.00	85.00	68.00	82.30	0.00	82.50
75%	166.00	53.00	100.00	85.00	95.00	210.00	85.00	15.00	80.00	95.00	76.00	86.70	0.00	87.20
max	187.00	103.00	100.00	100.00	100.00	274.00	100.00	33.00	100.00	100.00	100.00	99.00	15.00	108.60

图4-15 基本统计分析

数据结果说明：在图4-15中，count为数据总数；mean为列平均值；std为列标准差；min为列最小值；25%为列的后四分位数；50%为列的中位数；75%为列的前四分位数。从图中可得到，总分（综合评分）平均分为81.96，中值为82.50，从标准差（标准差为8.15）角度考虑学生之间有一定差异，最低分为29.5，最高分（加上附加分）108.6。从百分位数上看，学生获得的成绩较高，体质健康水平表现较好，3/4的学生综合评价得分高于或等于77.3，半数以上学生高于或等于82.50，3/4的学生高于或等于87.20。

4.3.2　平均值分析法

平均值分析法是指运用计算平均数的方法，反映总体在一段时间、地点条件下，某一数量特征的一般水平。

在项目数据分析中，统计各学校综合评价项目的平均值，通过对平均值的对比分析，更能反映学校学生群体的体质健康水平差异。

导入教科书配套学习资源包“第四章\课本素材\test4-6.xlsx”文件，统计各所学校的综合评价（总分）平均分，比较各学校平均分值情况并初步评价各学校学生体质健康总体情况，关键语句如下：

```
Df.groupby(by=['学校编号'])['总分'].agg({'人数':np.size,'平均分':np.mean,'
                标准差':np.std,'众数':np.median,'最高分':np.max,'最低分':np.min})
```

程序执行结果部分截图如图4-16所示。

学校编号	人数	平均分	标准差	众数	最高分	最低分
171	207.00	86.09	3.07	85.70	96.00	75.80
172	92.00	82.79	5.76	83.00	101.50	72.60
261	29.00	80.39	3.42	81.60	85.60	74.60
364	541.00	84.29	8.59	84.50	104.00	55.50
365	550.00	84.94	7.30	85.20	104.70	60.20
461	462.00	81.25	8.27	82.30	102.00	43.40
462	566.00	76.47	7.29	77.50	95.00	35.60
464	250.00	80.99	8.72	81.30	105.20	52.50
465	460.00	82.66	9.26	83.65	108.00	29.50
561	448.00	83.48	6.79	83.20	106.70	63.90
562	560.00	85.46	4.97	85.50	100.60	71.30
566	1,096.00	80.76	7.45	81.40	103.80	60.00
572	381.00	76.66	10.02	77.60	106.00	30.80
661	505.00	83.03	8.23	83.10	108.60	42.80
662	448.00	81.65	5.88	81.90	97.70	57.40
764	234.00	85.46	7.76	86.20	104.10	59.20
861	408.00	79.60	9.12	80.35	103.00	44.40

图4-16　各学校综合评价平均分描述性分析

从执行结果图4-16可以得到，全区各学校中，综合评价项目得分校平均分最高是编号为171的学校，得分86.09分；最低是编号为462的学校，平均分只有76.47分。平均分最低的与最高的学校相差接近10分。

思考并展开讨论：是什么原因造成编号为172的学校平均分较低？需要从哪个角度进行分析？

平均指标既可用于同一现象在不同地区、不同部门间的横向比较，也可用于同一现象在不同时间的纵向对比。

4.3.3 分组分析法

在数据分析项目中，为比较不同群体之间的差异，如不同年龄学生身体素质的差异，男女生平均身高或体重差异，班与班、校与校之间的学生身体形态差异等，需要对学生数据进行分组统计分析。分组分析法是根据数据分析对象的特征，按照一定的指标，把数据分析对象划分为不同的部分和类型，以对比分析各组之间差异性的一种分析方法。

导入教科书配套学习资源包“第四章\课本素材\test4-6.xlsx”文件，统计各所学校男女生在综合评价（总分）平均分上的差异，关键语句如下：

```
Df.groupby(by=['学校编号','性别']) ['总分'].agg({'人数':np.size,'平均分':np.mean,
                '标准差':np.std,'众数':np.median,'最高分':np.max,
                '最低分':np.min})
```

执行结果部分截图如图4-17所示。

		人数	平均分	标准差	众数	最高分	最低分
学校编号	性别						
171	女	95.00	85.46	2.81	85.50	95.50	75.80
	男	112.00	86.63	3.19	86.10	96.00	78.80
172	女	47.00	82.87	6.35	82.80	101.50	72.60
	男	45.00	82.70	5.14	83.30	95.90	73.40
261	女	14.00	81.10	3.30	81.75	85.60	75.40
	男	15.00	79.72	3.51	81.00	85.30	74.60
364	女	257.00	84.41	8.90	85.00	102.00	55.80
	男	284.00	84.18	8.30	84.15	104.00	55.50
365	女	254.00	85.15	7.25	85.20	104.70	61.30
	男	296.00	84.76	7.36	85.10	103.80	60.20
461	女	233.00	80.64	8.13	81.60	97.70	57.70
	男	229.00	81.87	8.38	82.70	102.00	43.40

图4-17　各学校按性别分组综合评价（总分）平均分统计分析

对数据进行分组的目的是为了便于对比，把总体中具有不同性质的对象区分开，把性质相同的对象合并在一组，保持各组内对象属性的一致性、组与组之间属性的差异性，以便进一步运用各种数据分析方法来解构内在的数量关系，因此分组法必须与对比法结合运用。

分组分析法的关键是确定组数与组距。结合本项目分析可以按年龄、性别、某个测试项目（如身高、体重、50米跑等）的得分或等级进行分组，对比分析不同分组的学生在身高、体重或体质健康水平上是否存在明显差异。

4.3.4　对比分析法

对比分析法也称比较分析法，是按照特定的指标系对客观事物加以比较，以达到认识事物的本质和规律并做出正确的判断或评价的数据分析方法。

对比分析法通常是把两个或以上相互联系的指标数据进行比较，从数量上展示和说明研究对象规模的大小、水平的高低、速度的快慢，以及各种关系是否协调。因此对比分析法多与其他分析方法结合使用。

根据对比的对象和方式不同，一般分为横向对比和纵向对比。横向对比是指在同一时间条件下，对不同总体指标的比较。例如，不同学校、不同班别、不同性别之间的比较。纵向对比是指在同一总体条件下，对不同时期数据进行的比较。例如，本年度与上年度比较、应届与往届的比较、实验前与实验后的比较。

4.3.5　交叉分析法

交叉分析通常用于分析两个或两个以上分组变量之间的关系，以交叉表形式进行变量间关系的对比分析。交叉分析使用的分析函数如下：

```
pivot_table(values,index,columns,aggfunc,fill_value)
```

参数说明：values为数据透视表中的值；index为数据透视表中的行；columns为数据透视表中的列；aggfunc为统计函数；fill_value为NA值的统一替换。

在“中学生体质健康数据管理系统的数据分析”项目中需要对群体间的差异进行分析。如男、女生（不同性别）在体质健康水平上是否存在明显差异？统计分析男、女生在综合评价项目“优秀”“良好”“合格”“不合格”等级的分布情况。统计时，“性别”和“综合评价（总分）”都是表中的两个字段变量，因此涉及“性别”与“综合评价（总分）”两个变量的二维交叉，反映不同的性别总分等级的分布情况，关键语句如下：

```
bins=[0,60,80,90,max(Df.总分)+1]
labels=['不合格','合格','良好','优秀']
```

```
totallev=pd.cut(Df.总分,bins,labels=labels)
Df['总分等级']=totallev
Df.pivot_table(values=['总分'],index=['总分等级'],columns=['性别'],
                aggfunc=[np.size,np.mean,np.var])
```

	size		mean		var	
	总分		总分		总分	
性别	女	男	女	男	女	男
总分等级						
不合格	36	42	54	53	50	45
合格	1,340	1,273	74	74	22	21
良好	1,771	1,752	85	85	7	8
优秀	487	536	94	94	10	10

图4–18 交叉分析结果

从执行结果图4–18可以看出，性别成为纵向量，总分等级为横向量。男生获得优秀等级人数比女生多，女生不合格人数比男生稍少；各等级男、女生平均分接近；在不合格等级中女生的差异要比男生大。

交叉分析法通常用于分析两个变量之间的关系，例如各个大型网站浏览和年龄之间的关系。实际使用中，通常把两个交叉变量推广到行变量和列变量之间的关系，这样行变量可能有多个变量，列变量也可能有多个变量。交叉分析的主要作用，是从多个角度细分数据，从中发现数据变化的具体原因。

4.3.6 相关分析

相关分析是研究现象之间是否存在某种依存关系，并对具有依存关系的现象探讨其相关方向以及相关程度。如子女的身高与其父母身高的关系。从遗传角度看，父母身高较高，其子女的身高一般也比较高。但实际情况不完全相符，因为子女的身高除了由父母的身高这一因素所决定外，还受饮食、运动和睡眠等其他因素影响。应用相关分析可以研究饮食、运动和睡眠等其他因素与身高是否存在某种相互依存关系即相关关系，以及相关关系的强弱程度。相关关系不同于因果关系，相关性表示两个变量同时变化，而因果关系是一个变量导致另一个变量变化。

1. 相关关系的类型

（1）按相关程度分类。

完全相关：一种现象的数量变化完全由另一种现象的数量变化所确定。

不完全相关：两个现象之间的关系介于完全相关和不相关之间。

不相关：两个现象彼此互不影响，其数量变化各自独立。

（2）按相关的方向分类。

正相关：两个现象的变化方向相同。

负相关：两个现象的变化方向相反。

（3）按相关的形式分类。

线性相关：两种相关现象之间的关系大致呈线性关系。

非线性相关：两种相关现象之间的关系并不表现为直线关系，而是近似于某种曲线方程的关系。

2. 相关系数

相关关系分为线性相关和非线性相关，线性相关也称为直线相关，非线性相关也就是曲线相关。

线性相关是最常用的一种，即当一个连续变量发生变动时，另一个连续变量相应地呈线性关系变动，用皮尔逊（Pearson）相关系数r度量。

皮尔逊相关系数r是反映连续变量之间线性相关强度的度量指标。它的取值范围介于+1与−1之间，如图4−19所示。

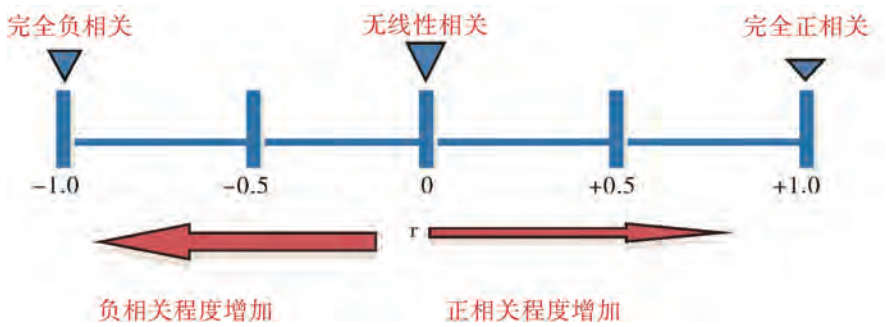


图4−19　皮尔逊相关系数指标

r的正、负号可以反映相关的方向，当r>0时表示正相关，r<0时表示负相关。r的大小反映相关的程度，当r=0时表示两个变量之间不存在线性相关关系。

相关程度	高度相关	中度相关	低度相关
相关系数 |r| 取值范围	0.8≤ |r| <1	0.3≤ |r| <0.8	0≤ |r| <0.3

3. 相关分析函数

```
DataFrame.corr( )　#计算每列两两之间的相关度
Series.corr(other)　#计算原序列与传入序列之间的相关度
```

1. 计算学生身高与体重的相关程度，关键程序语句如下：

```
Df ['身高'].corr(Df ['体重'])
```

2. 计算学生各项目得分之间的相关程度，程序语句如下：

```
Df.loc[：,['身高','体重','肺活量评分','50米跑评分','长跑评分',
        '男女体能评分','立定跳远评分','坐位体前屈评分']].corr( )
```

	身高	体重	肺活量评分	50米跑评分	长跑评分	男女体能评分	立定跳远评分	坐位体前屈评分
身高	1.00	0.57	0.26	0.25	0.04	-0.13	0.17	-0.00
体重	0.57	1.00	0.34	0.06	-0.13	-0.16	-0.04	0.05
肺活量评分	0.26	0.34	1.00	0.15	0.15	0.06	0.16	0.08
50米跑评分	0.25	0.06	0.15	1.00	0.28	0.15	0.48	0.13
长跑评分	0.04	-0.13	0.15	0.28	1.00	0.44	0.29	0.09
男女体能评分	-0.13	-0.16	0.06	0.15	0.44	1.00	0.29	0.20
立定跳远评分	0.17	-0.04	0.16	0.48	0.29	0.29	1.00	0.17
坐位体前屈评分	-0.00	0.05	0.08	0.13	0.09	0.20	0.17	1.00

图4–20 测试项目相关分析

从执行结果图4–20可以看出，身高与体重相关系数为0.57，有中度相关关系。观察上述结果，请找出相关程度较高的3组变量。

4.3.7 常用的数据分析方法对比

在统计学领域，将数据分析划分为描述性统计分析、探索性数据分析以及验证性数据分析。其中，探索性数据分析侧重于在数据之中发现新的特征，而验证性数据分析则侧重于验证已有假设的真伪证明。不同的分析目标，使用不同的分析方法，表4–5为常用的数据分析方法对比。

表4–5 常用的数据分析方法对比

方法	说明
对比分析法	前后对比，不同时期对比，班级、学校、地区对比，活动效果对比。
分组分析法	分组的目的在于便于对比，分组法必须与对比法结合起来。
结构分析法	某部分数值占总体的比率，如体质测试各项占有率。
平均分析法	算术平均、调和平均、几何平均、众数与中位数。
交叉分析法	常见的二维交叉表。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，完成项目数据分析。

1. 体验数据描述性分析的各种方法。
2. 选择合适的数据分析方法开展数据分析。

4.4 数据的可视化表达

在撰写数据分析报告时，借助于图形化手段，可以清晰有效地传达与沟通信息，深入地理解和洞察数据。Matplotlib是Python常用的可视化程序库。

4.4.1 常用图形的绘制

1. 饼形图

饼形图又称圆形图，是以扇形的面积来指代某种类型的频率，能够直观地反映个体与总体的比例关系。

绘制饼形图的方法：pie(x,labels,colors,explode,autopct)，其参数说明如表4-6 所示。

表4-6　pie函数参数使用说明

参数	使用说明
x	绘图的序列，即每一块的比例。
labels	饼形图的各部分标签，即每一块饼形图外侧显示的说明文字。
colors	饼形图的各部分颜色，使用RGB颜色。
explode	需要突出的块状序列。
autopct	饼形图占比的显示格式。例如'%.2f'，保留两位小数。

探究活动

实践

绘制饼形图表示各校人数比例，关键程序代码如下：

```
gb=Df.groupby(by=['学校编号'],as_index=False)['学校编号'].agg({'人数':np.size})
font={'family':'Simhei'}
plt.rc('font',**font)
plt.pie(gb['人数'],labels=gb['学校编号'],autopct='%.1f%%')
plt.show()
```

执行结果如图4-21所示。

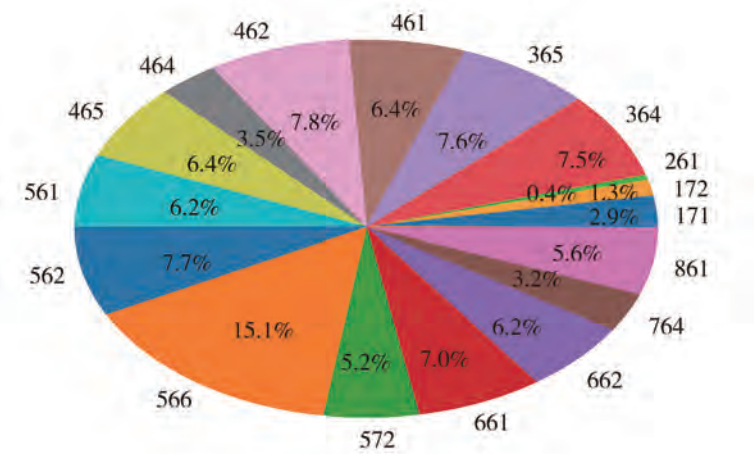


图4-21　各校人数百分比饼形图

2. 柱状图

柱状图用于显示一段时间内的数据变化或显示各项之间的比较情况，是以柱的高度来指代某种类型的频数，根据数据大小绘制的统计图，用来比较两个或两个以上的数据。

绘制柱状图的方法主要有以下两种：

```
bar(left,height,width,color)
barh(bottom, width,height, color)
```

参数说明如表4-7 所示。

表4-7　柱状函数参数使用说明

参数	使用说明
left	X轴的位置序列，一般采用arange函数产生一个序列。
height	Y轴的数值序列，也就是柱状图的高度。
width	柱状图的宽度，一般设置为1。
color	柱状图的填充颜色。

体验

绘制各校人数柱状图，关键程序代码如下：

```
gb=Df.groupby(by=['学校编号'],as_index=False)['学校编号'].agg({'人数':np.size})
font={'family':'Simhei'}
plt.rc('font',**font)
plt.bar(range(len(gb)),gb['人数'],color='rgb',tick_label=gb['学校编号'])
plt.show()
```

执行结果如图4-22所示。

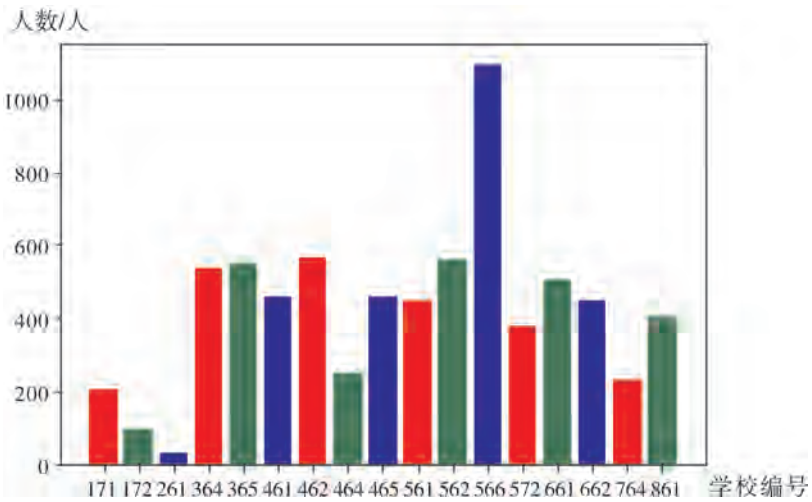


图4-22　各学校学生人数柱状图

3. 直方图

直方图用一系列等宽不等高的长方形来绘制，宽度表示数据范围的间隔，高度表示在给定间隔内数据出现的频数，变化的高度形态表示数据的分布情况。

直方图绘制方法：hist(x,color,bins,cumulative=False)，其参数说明如表4-8 所示。

表4-8　hist函数参数使用说明

参数	使用说明
x	需要进行绘制的向量。
color	直方图填充的颜色。
bins	设置直方图的分组个数。
cumulative	设置是否累积计数，默认是False。

观察

绘制各学校学生总分的直方图，关键程序代码如下：

```
plt.hist(Df.总分,bins=20)
plt.xlabel('分数/分')
plt.ylabel('人数/人')
plt.title('总分直方图')
plt.show()
```

执行结果如图4-23所示。

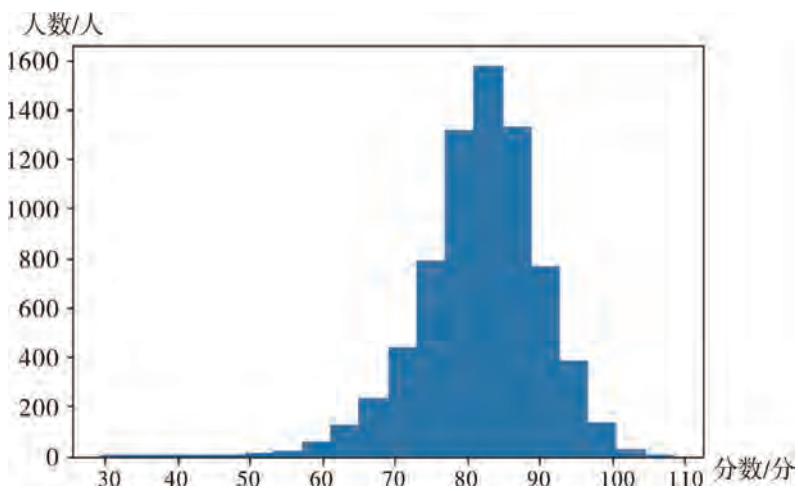


图4-23 各校学生总分直方图

4. 散点图

散点图是分别以自变量和因变量作为横、纵坐标，利用散点的分布形态反映变量关系的一种图形。当自变量与因变量线性相关时，在散点图中，点近似分布在一条直线上。

散点图绘制方法如下：

```
plt.plot（x,y,'.'，color=（r,g,b））
plt.xlabel（'x轴坐标'）
plt.ylabel（'y轴坐标'）
plt.grid（True）
```

参数说明：x和y表示X轴和Y轴的序列；'.'表示使用小点；color表示散点图的颜色，可以用RGB定义，也可以用英文字母定义，常用的RGB颜色如表4-9所示。

表4-9　常用RGB颜色对照

颜色	英文及简写	RGB	十六进制
白色	white（w）	（255，255，255）	#FFFFFF
黑色	black（k）	（0，0，0）	#000000
红色	red（r）	（255，0，0）	#FF0000
橙色	orange（o）	（255，165，0）	#FFA500
黄色	yellow（y）	（255，255，0）	#FFFF00
绿色	green（g）	（0，255，0）	#00FF00
蓝色	blue（b）	（0，0，255）	#0000FF
靛青色	indigo（i）	（75，0，130）	#4B0082
紫色	purple（p）	（128，0，128）	#800080
紫红色	magenta（m）	（255，0，255）	#FF00FF
蓝绿色	cyan（c）	（0，255，255）	#00FFFF

以身高作为自变量，体重作为因变量，讨论身高对体重的影响，绘制散点图，关键程序代码如下：

```
color=np.random.rand(1000)
plt.scatter(Df.身高,Df.体重,color)
plt.xlabel('身高/cm')
plt.ylabel('体重/kg')
plt.grid(True)
plt.show()
```

执行结果如图4-24所示。

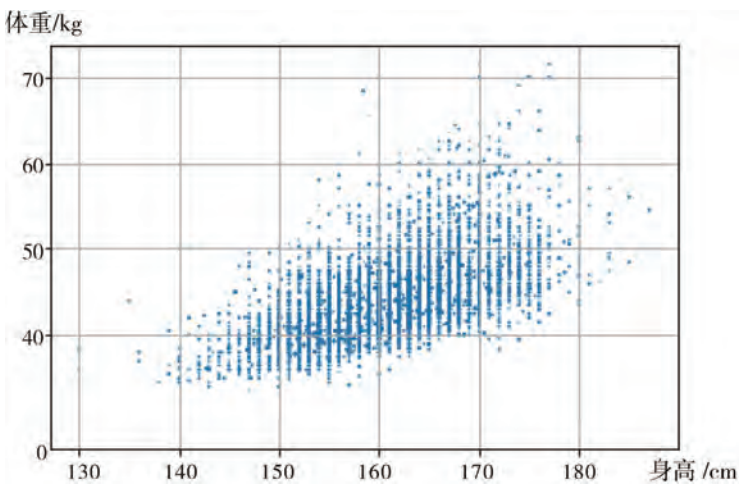


图4-24　身高体重散点图

5. 折线图

折线图也称趋势图，是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势。折线图绘制方法如下：

```
plt.plot（x,y,'-'，color）
plt.title（'图的标题'）
```

参数说明：x，y表示X轴和Y轴的序列；'-'表示画线的样式；color表示折线的颜色。画线的样式有多种，常用的样式说明如表4-10所示。

表4-10　plot函数画线样式使用说明

参数值	说明	参数值	说明
-	实线，连续曲线（默认样式）。	s	正方形标记散点图。
--	连续虚线（短划线）。	+	加号标记散点图。
:	由点连成的曲线。	p	五角星标记散点图。
-.	连线的带点网线。	h	六角形标记散点图。
.	小点，散点图。	x	十字标记散点图。
o	大点，散点图。	d	菱形标记散点图。
,	像素点的散点图。	*	星号的点，散点图。
>	右角标记散点图。	^	上指向三角形。
<	左角标记散点图。	v	下指向三角形。
1（2，3，4）	伞形上（下左右）标记散点图。		

绘制各校人数折线图，关键程序代码如下：

```
gb=Df.groupby(by=['学校编号'],as_index=False)['学校编号'].agg({'人数':np.size})
name=gb.学校编号
x= range(len(name))
plt.plot(x,gb['人数'],label='人数',linewidth=3,color='r',marker='o',
markerfacecolor='blue',markersize=8)
plt.xlabel('学校编号',fontsize=16)
plt.ylabel('人数/人',fontsize=16)
plt.title('各学校人数折线图',fontsize=16)
plt.legend()
plt.xticks(x,name,rotation=45)
plt.show()
```

执行结果如图4-25所示。

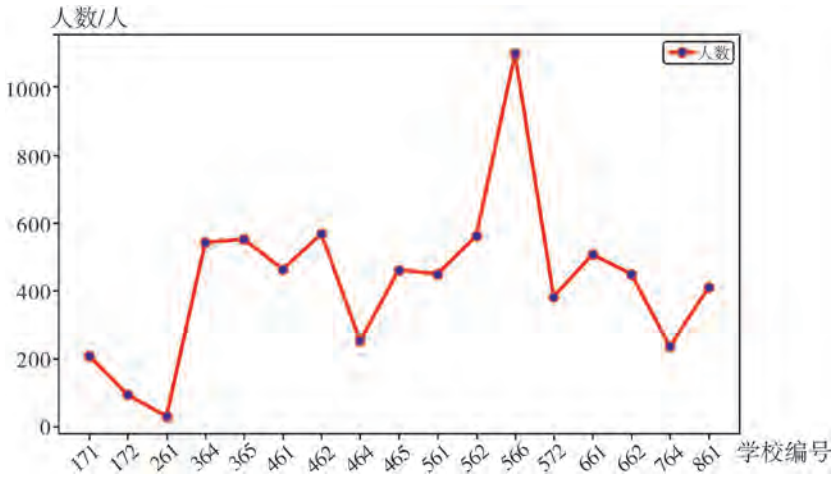


图4-25　各学校人数折线图

6. 箱形图

箱形图又称为盒须图、盒式图或箱线图。如图4-26所示，箱形图包含均值、分位数、中位数以及极值等统计量。箱形图不仅显示不同类别数据平均水平差异，还能揭示数据间离散程度、异常值、分布差异。

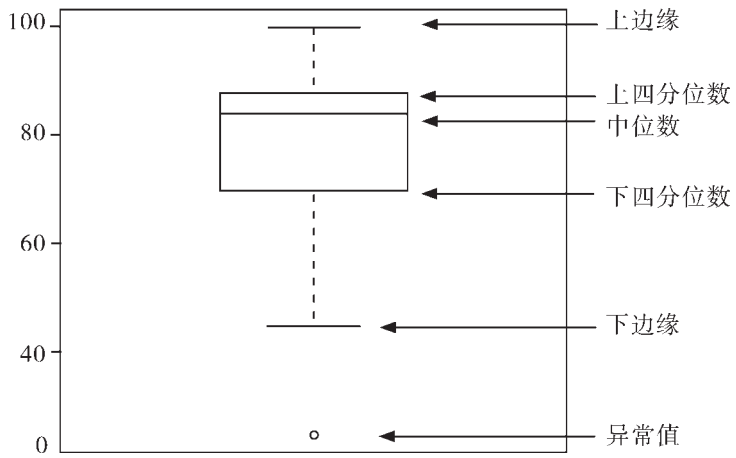


图4-26　箱形图统计量说明

箱形图的绘制方法是plt.boxplot( )。函数boxplot包含众多参数，涉及对箱盒的颜色及形状、线段线型、均值线、异常点的形状、大小设置，根据需要使用参数，能使图形的数据特征表现更直观。

绘制各测试项目得分的箱形图，关键程序代码如下：

```
Df.boxplot()
plt.show()
```

执行结果如图4–27所示。

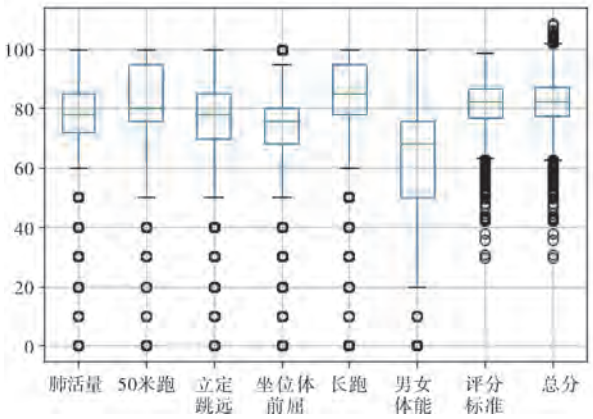


图4–27　测试项目得分箱形图

从图4–27的执行结果可以直观地得出以下信息：

（1）各项目测试中，长跑的平均成绩比较高，而男、女生体能（男生引体向上，女生1分钟仰卧起坐）的平均成绩比较低。

（2）肺活量、坐位体前屈、标准分、总分成绩分布比较集中，因为箱子比较短。而50米跑和男、女生体能成绩比较分散。

（3）从各个箱形图的中位数和上下四位数的间距也可以看出，肺活量的评分、测试各项得分总和（标准分和总分）成绩分布比较对称。50米跑和男、女生体能成绩分布不平衡，其中50米跑项目，大部分学生分布在80~95分；男、女生体能，大部分学生分布在50~70分。

（4）在各项目对应的箱形图中都出现了异常点，各项目按权重计算后得到的标准分和总分的异常点较多。总分中既有分数异常高的学生，也有分数异常低的学生。

4.4.2　数据可视化实例1——回归分析

“回归最初”是遗传学中的一个名词，是由英国生物学家兼统计学家高尔顿首先提出来的。他在研究人类的身高时，发现高个子回归于人口的平均身高，而矮个子则从另一个角度回归于人口的平均身高。回归分析是研究自变量与因变量之间关系形式的分析方法，它主要是通过建立因变量Y与影响它的自变量X_i（i=1,2,3,…）之间的回归模型，来预测因变量Y的发展趋势。

1. 回归分析的步骤

回归分析基于对数据的观测，建立变量间适当的依赖关系，以分析数据内在规律。回归分析是应用极其广泛的数据分析方法之一，多用于预报、控制问题。回归分析的基本步骤如图4–28所示。

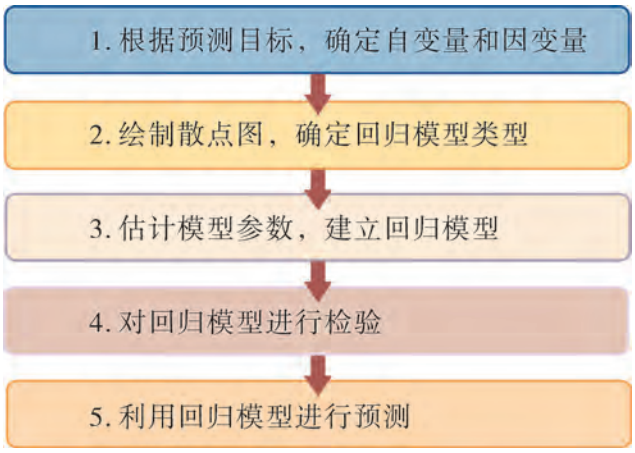


图4-28　回归分析的基本步骤

2. 回归分析的应用

回归分析被广泛地用于解释市场占有率、销售额、品牌偏好及市场营销效果，其作用主要表现在以下几个方面：

（1）判别自变量是否能解释因变量的显著变化——关系是否存在。

（2）判别自变量能够在多大程度上解释因变量——关系的强度。

（3）判别关系的结构或形式——反映因变量和自变量之间相关的数学表达式。

（4）预测自变量的值。

（5）当评价一个特殊变量或一组变量对因变量的贡献时，对其自变量进行控制。

通过学生身高和体重的相关分析，我们发现全体学生的身高和体重有显著的相关关系，即身高越高，体重越重。利用学生身高和体重的相关关系，通过学生体重预估学生的身高。

1. 根据预测目标，确定自变量和因变量。

因为是预测学生的身高，所以将“体重”作为自变量，将“身高”作为因变量。导入教科书配套学习资源包“第四章\课本素材\hgfx.xlsx”文件中的数据。

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import linear_model
import xlrd
fileNameStr=(':\第四章\课本素材\hgfx.xlsx')
xls = pd.ExcelFile(fileNameStr)
Df = xls.parse('Sheet1')        #读入dataframe
```

2. 绘制散点图，确定回归模型类型。

相关分析中已经绘制“身高”和“体重”两个变量的散点图（图4-24），两个变量之间存在明显的线性关系，因此采用简单的线性回归分析方法是合理的。

3. 估计模型参数，建立回归模型。

```
regr=linear_model.LinearRegression()  #建立线性回归模型
regr.fit(Df['体重'].reshape(-1,1),Df['身高'])  #拟合
a,b = regr.coef_,regr.intercept_  #求得直线的斜率、截距
print(a,b)  #打印直线的斜率、截距
```

4. 利用回归模型进行预测。

```
weight=110  #给出预测的体重值
print(a * weight + b)  #根据直线方程计算的身高
print(regr.predict(weight))  #根据predict方法预测的身高
plt.scatter(Df['体重'],Df['身高'], color='blue')  #真实的数据点
plt.plot(Df['体重'],a*Df.体重 + b, color='red', linewidth=2)  #拟合的直线
plt.title('体重预测身高')
plt.xlabel('体重/kg')
plt.ylabel('身高/cm')
plt.show()
```

程序代码运行后，得到如图4-29所示身高和体重的相关关系图形，根据蓝色的数据点作出红色的一次函数图象，通过程序代码中“weight”变量的学生体重值和线性关系预估学生的身高。

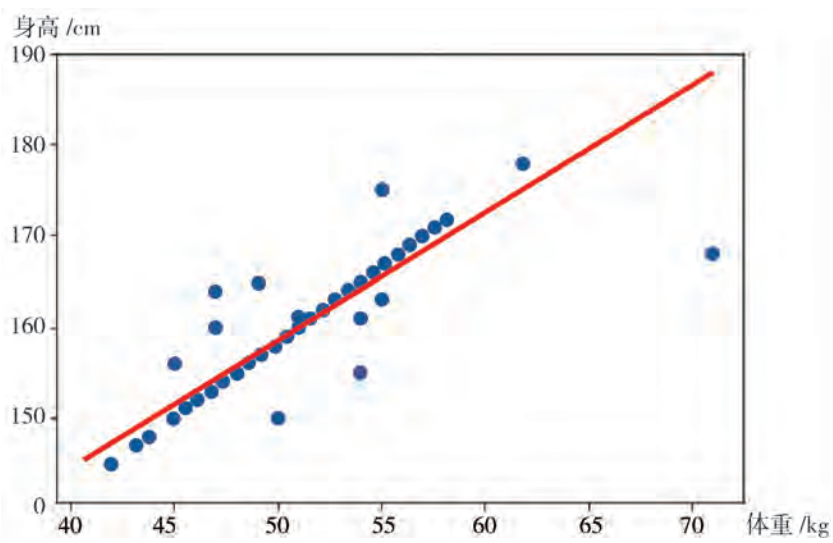


图4-29　体重、身高回归分析结果可视化图

4.4.3　数据可视化实例2——聚类分析

聚类分析对看似无序的对象进行分组、归类，以达到更好地理解研究对象的目的。聚类分析根据事物彼此不同的属性进行辨认，将具有相似属性的事物聚为一类，使得同一类的事物具有高度的相似性，而不同类别的对象相似性较低。

1. 聚类分析概述

聚类分析算法种类繁多，在众多的聚类算法中普及性最广、最实用、最具代表性的包括K-均值聚类、K-中心点聚类、密度聚类、系谱聚类、期望最大化聚类。

2. 聚类分析基本过程

聚类分析的基本过程如图4-30所示。

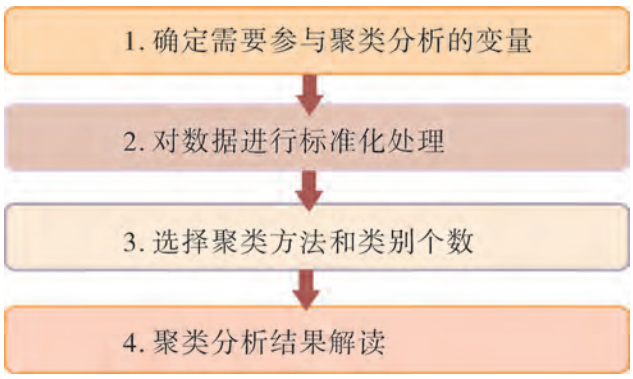


图4-30　聚类分析的基本过程

3. 聚类分析基本思想

聚类分析的基本思想认为研究的样本或变量之间存在着程度不同的相似性，根据样本的多个观测指标，具体找出一些能够度量样本或指标之间相似程度的统计量，以这些统计量为划分类型的依据，把关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到把所有的样本都聚合完毕，形成一个由小到大的分类系统。

K-means算法是典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此K-means聚类将得到紧凑且独立的簇作为最终目标。

学生体质健康通过肺活量、长跑、短跑、立定跳远等项目测评得到总分。现抽取短跑、立定跳远和坐位体前屈三个指标进行聚类分析，对学生进行分类，从而开展针对性的项目训练，提高学生体质健康测试得分。

由于使用短跑、立定跳远和坐位体前屈三个指标得分的单位及量级相当，所以采用原始得分进行聚类分析，无须进行数据标准化处理。如果变量间存在单位或量级的差异，就需要先对数据进行标准化处理。

K-means聚类分析，主要程序代码如下：

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans　#导入K均值聚类算法
import xlrd
fileNameStr=(':\第四章\课本素材\jlfx.xlsx ')　#待聚类的数据文件
xls = pd.ExcelFile(fileNameStr)
Df = xls.parse('Sheet1')
clf=KMeans(n_clusters=3)　#聚类类别数为3
y_pred=clf.fit_predict(Df)　#调用k-means算法，进行聚类分析
print(clf)
print(y_pred)
plt.title("50米跑与立定跳远聚类")
plt.xlabel("50米跑成绩/分")
plt.ylabel("立定跳远成绩/分")
plt.legend("分类")
plt.scatter(Df.短跑,Df.立定跳远,c=y_pred,marker='o')
plt.show()
```

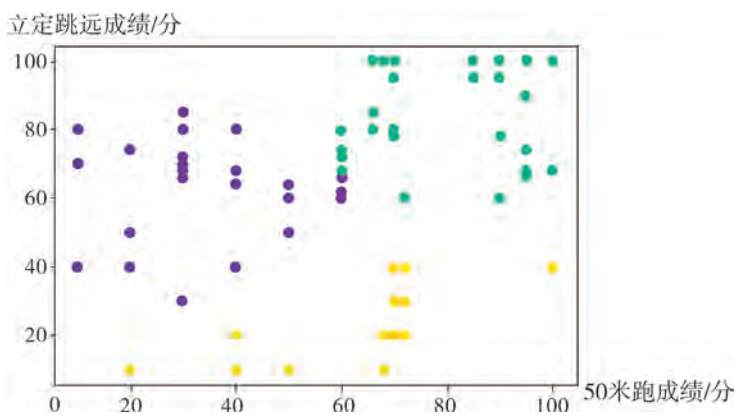


图4-31　聚类分析结果可视化图

从图4-31可以看出，学生被分成三个组别：

紫色类别的学生50米跑成绩偏低，立定跳远成绩一般。在加强这两个项目训练之余，应有针对性地加强50米跑的训练。

黄色类别的学生立定跳远成绩偏低，需要有针对性地加强立定跳远项目的训练。黄色类别学生中有个别学生50米跑成绩也较低。

蓝绿色类别的学生50米跑和立定跳远的成绩相对较高。

拓展

聚类分析将物理或抽象对象的集合分组为由类似的对象组成的多个类，互联网行业是聚类分析应用最为成熟的领域，同时聚类也是重要的人类行为之一。从上面的学习可见，采集准确的数据、选择合适的算法是数据分析成功的必要保障。有效的聚类分析可以揭示数据的内部结构，掌握数据结构是制订解决问题方案的先决条件。聚类分析已经成为数据分析和挖掘研究中的一个热点。聚类分析应用范例如表4-11所示。

表4-11　聚类分析应用范例

领域	应用场景
生物研究	对动植物及其基因进行分类，获取对种群固有结构的认识。
商业	研究消费者行为，发现不同的客户群，寻找新的潜在市场。
互联网	把用户浏览、消费行为进行聚类，研究总结用户特征。
金融研究	根据用户投资行为和资产状况对用户进行分类。
城市规划	根据区域特征对城市布局、功能进行分类。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，实践与体验各种数据可视化表达方法，分析并实践回归分析和聚类分析实例，参照项目范例的样式，撰写相应的项目成果报告。

成果交流

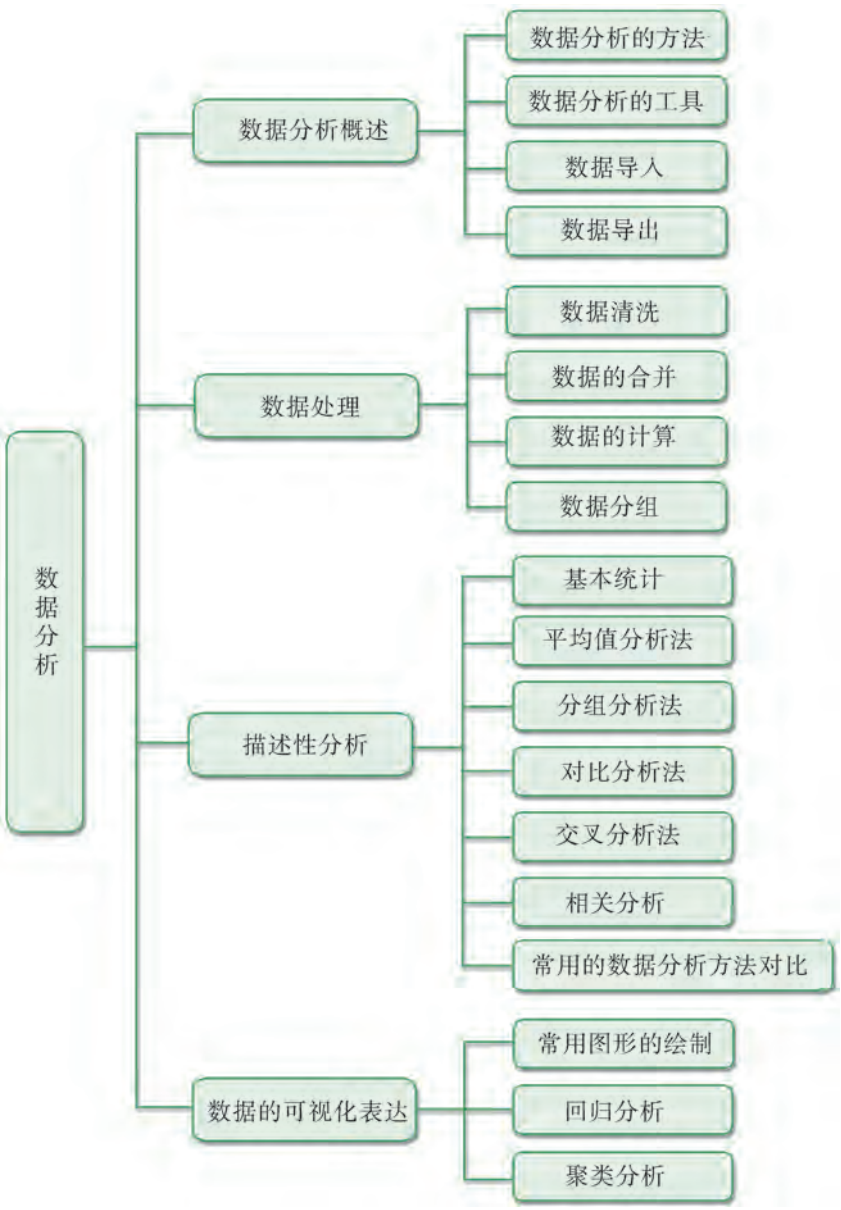
各小组运用数字化学习工具，将所完成的项目成果，在小组或班级上进行展示与交流，共享创造、分享快乐。

活动评价

各小组根据项目选题、拟订的项目方案、实施情况以及所形成的项目成果，利用教科书附录2的“项目活动评价表”，开展项目学习活动评价。

本章扼要回顾

同学们通过本章学习，根据“数据分析”知识结构图，扼要回顾、总结、归纳学过的内容，建立自己的知识结构体系。



回顾与总结

本章学业评价

同学们完成下列测试题（更多的测试题可以在教科书的配套学习资源包中查看），并通过“本章扼要回顾”以及本章的项目活动评价，综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。

1. 单选题

（1）在高考志愿填报的调查中，想了解学生填报高校与专业的情况，最适合选用（　　）。

A. 频数分析　　B. 平均值分析　　C. 相关分析　　D. 聚类分析

（2）显示时间序列的对比关系最适合使用（　　）展示。

A. 雷达图　　B. 直方图　　C. 柱状图　　D. 折线图

（3）（　　）最能反映数据分布的离散程度。

A. 方差　　B. 中位数　　C. 标准差　　D. 频数

2. 思考题

通过本章的学习，你了解了多少种数据分析的方法？各种数据分析方法有哪些异同点？

3. 情境题

（1）中国居民消费价格指数变化。

居民消费价格指数是度量居民生活消费品和服务价格水平随着时间变动的相对数，综合反映居民购买的生活消费品和服务价格水平的变动情况。下表所示是国家统计局发布的2017年7月—2018年7月全国居民消费价格分类指数。

时间	2017/7	2017/8	2017/9	2017/10	2017/11	2017/12	2018/1	2018/2	2018/3	2018/4	2018/5	2018/6	2018/7
居民消费价格指数	101.4	101.8	101.6	101.9	101.7	101.8	101.5	102.9	102.1	101.8	101.8	101.9	102.1
食品烟酒类	99.9	100.4	99.6	100.3	99.8	100.3	100.2	103.6	102	101.1	100.7	100.8	101
衣着类	101.4	101.3	101.3	101.2	101.2	101.3	101.4	101.1	101.1	101.1	101.1	101.1	101.2
居住类	102.5	102.7	102.8	102.8	102.8	102.8	102.7	102.2	102.2	102.2	102.2	102.3	102.4
生活用品及服务类	101.1	101.3	101.4	101.5	101.5	101.6	101.5	101.8	101.6	101.5	101.5	101.5	101.6
交通和通信类	99.8	100.7	100.5	100.8	101.3	101.2	100.2	101.5	100.3	101.1	101.8	102.4	103
教育文化和娱乐类	102.5	102.5	102.3	102.3	102	102.1	100.9	103.7	102.2	102	101.9	101.8	102.3

（续表）

时间	2017/7	2017/8	2017/9	2017/10	2017/11	2017/12	2018/1	2018/2	2018/3	2018/4	2018/5	2018/6	2018/7
医疗保健类	105.5	105.9	107.6	107.2	107	106.6	106.2	106	105.7	105.2	105.1	105	104.6
其他用品和服务类	101.3	101.4	101.4	101.8	101.7	101.9	101.2	101.7	101.2	100.9	101	100.9	101.2

（资料来源：国家统计局）

①选用合适的图表类型，将以上统计数据直观地呈现出来。

②观察图表，请指出增幅位列前三名的指数指标项。

③居民消费价格指数变动率在一定程度上反映了通货膨胀或紧缩的程度。请根据统计图表，简述你的分析。

（2）中国人口数据的分析。

人口出生率是指某地在一个时期内（通常指一年）出生人数与平均人口之比，它反映了人口的出生水平，一般用千分数表示。下表所示是国家统计局发布的2011—2016年人口出生率、死亡率和自然增长率。

指标	2011年	2012年	2013年	2014年	2015年	2016年
人口出生率 / ‰	11.93	12.1	12.08	12.37	12.07	12.95
人口死亡率 / ‰	7.14	7.15	7.16	7.16	7.11	7.09
人口自然增长率 / ‰	4.79	4.95	4.92	5.21	4.96	5.86

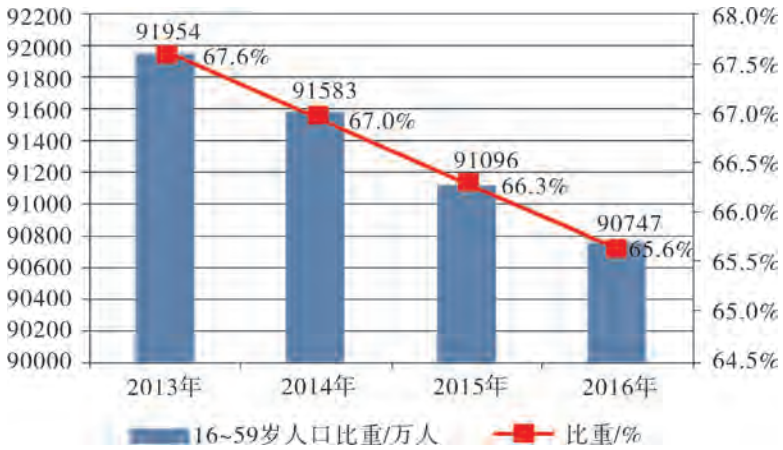
（资料来源：国家统计局）

人口出生率=（年内出生人数/年内平均人口数）× 1000‰

①选用合适的图表类型，将以上三种数据的变化趋势直观地呈现出来。

②请查阅相关资料，解释2016年人口自然增长率增幅较大的原因。

③人口红利是指一个国家的劳动年龄人口占总人口比重较大，抚养率比较低，为经济发展创造了有利的人口条件，整个国家的经济呈高储蓄、高投资和高增长的局面。2013年1月，国家统计局公布的数据显示，我国人口红利趋于消失，导致未来中国经济要过一个“减速关”。观察下图并结合“全面放开二孩”政策，对图中的数据进行解读。



2013—2016年我国劳动年龄人口统计

第五章
数据管理与分析的发展趋势

近年来，随着物联网、移动互联网、云计算以及信息通信技术的飞速发展，各行各业积累了越来越多的数据，并且每个人都是数据的产生源。如通过基于位置的服务可以采集个人在地球上的运动轨迹，通过在线支付可以采集个人的支付记录，通过社交网络服务可以采集个人的交往记录，通过邮件、文档、时间轴、视频监控等可以采集个人的言行记录……也就是说，我们的行为、位置，甚至身体生理数据等每一点变化都可成为被记录和分析的数据。这些大量的数据该如何观察，如何有效地管理与分析，进而如何更好地应用，都是当今探索的热点问题，同时也促进着数据管理与分析技术的不断发展。

本章将通过“体验数据的管理与分析新技术应用”项目，进行自主、协作、探究学习，让同学们运用数字化学习方式，了解数据管理与分析技术的新发展；结合恰当的案例分析，认识大数据和数据挖掘对信息社会问题解决和科学决策的重要意义，从而将知识建构技能培养与思维发展融入运用数字化工具解决问题和完成任务的过程中，促进信息技术学科核心素养达成，完成项目学习目标。

- 数据管理与分析的新发展
- 数据挖掘与大数据的意义

项目范例 体验电子商务数据的管理与分析新技术应用

情境

中国互联网企业的飞速发展，带动了数据量的剧增，继而产生了数据管理与分析技术的新需求，从而也促进了数据管理与分析技术的新发展与新应用。其中，当今中国较大的电子商务公司，得益于其丰富的生态和业务，积累了包括电商交易、搜索、物流、支付、广告、风控、电影、移动、视频、音乐、位置等种类繁多的高质量数据，为探索与发挥大数据价值提供了极大的方便，并且为很多电子商务公司提供了一系列数据产品服务。例如，基于全渠道数据融合、全链路数据产品集成，为商家提供数据披露、分析、诊断、建议、优化、预测等一站式数据产品服务，其中应用到的数据管理与分析技术非常先进，是我们学习数据管理与分析新应用的良好素材。

主题

体验电子商务数据的管理与分析新技术应用

规划

根据项目范例的主题，在小组中组织讨论，利用思维导图工具，制订项目范例的学习规划，如图5-1所示。

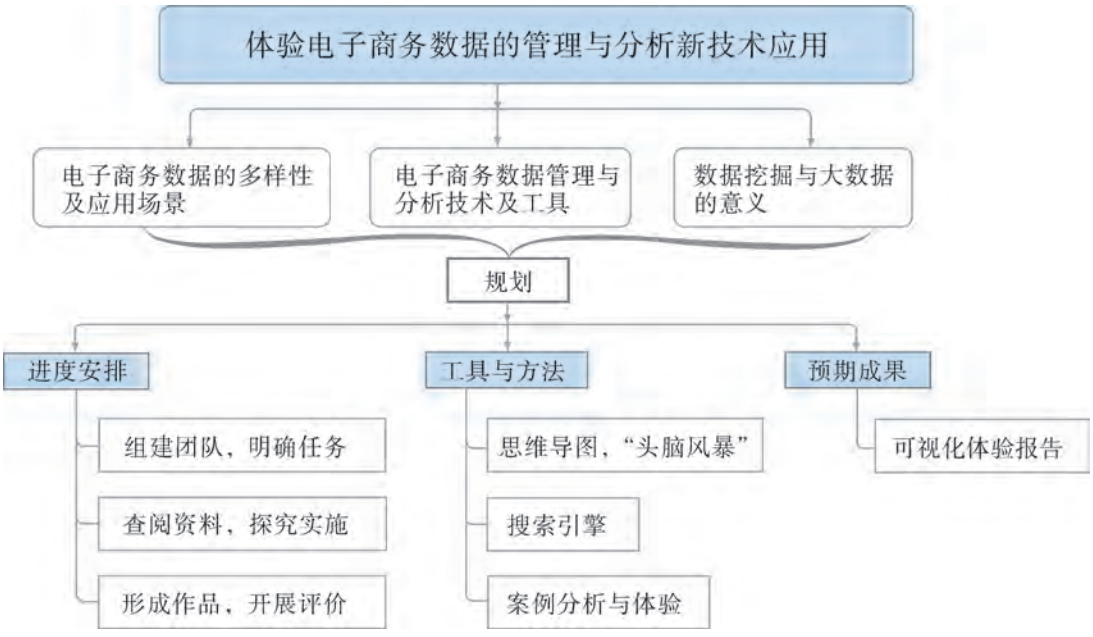


图5-1 “体验电子商务数据的管理与分析新技术应用”项目学习规划

探究

根据项目学习规划的安排，通过调查、案例分析、文献阅读和网上资料搜索，开展

“体验电子商务数据的管理与分析新技术应用”项目学习探究活动，如表5-1所示。

表5-1　“体验电子商务数据的管理与分析新技术应用”项目学习探究活动

探究活动	学习内容		知识技能
数据管理与分析技术的新发展	电子商务数据的多样性及应用场景。	了解相关数据及其应用场景。 了解数据库应用新需求。 了解数据库新技术。 了解数据分析新需求。 了解数据分析新技术及工具。	运用数字化学习方式，了解数据管理与分析技术的新发展。
	电子商务数据管理与分析技术及工具。		
数据挖掘与大数据的意义	数据挖掘的意义。	了解数据挖掘的发展历史。 数据挖掘技术的应用。	结合恰当的案例分析，认识大数据和数据挖掘对信息社会问题解决和科学决策的重要意义。
	大数据的意义。	大数据的发展历程。 大数据的应用。 大数据的影响。	

实施项目学习各项探究活动，进一步体验电子商务数据的管理与分析新技术应用。

在小组开展项目范例学习过程中，利用思维导图工具梳理小组成员在“头脑风暴”活动中的观点，建立观点结构图，运用多媒体创作工具（如演示文稿、在线编辑工具等），综合加工和表达，形成项目范例可视化学习成果，并通过各种分享平台发布，共享创造、分享快乐。例如，运用在线编辑工具制作的“体验电子商务数据的管理与分析新技术应用”可视化报告，可以在教科书的配套学习资源包中查看，其目录截图如图5-2所示。

图5-2　“体验电子商务数据的管理与分析新技术应用”可视化报告的目录截图

根据教科书附录2的“项目活动评价表”，对项目范例的学习过程和学习成果在小组或班级上进行交流，开展项目学习活动评价。

同学们以3～6人组成一个小组，选择下面一个参考主题，或者自拟一个感兴趣的主题，开展项目学习。

1. 体验票务系统大数据的管理与分析新技术应用
2. 体验图书大数据的管理与分析新技术应用
3. 体验医疗系统大数据的管理与分析新技术应用

各小组根据项目选题，参照项目范例的样式，利用思维导图工具，制订相应的项目方案。

各小组将完成的方案在全班进行展示交流，师生共同探讨、完善相应的项目方案。

5.1 数据管理与分析的新发展

5.1.1 数据的多样性与应用场景

以甲信息集成公司对从乙煤气电力公司和丙汽车公司采集而来的数据的应用为例。最初，煤气电力公司采集数据是为了保持其服务的稳定性，而汽车公司采集电动汽车的数据是为了提高经营效率，但是，信息集成公司将两者的数据建成数据集并整合成一个数据系

统，通过这个系统，车主能够从中掌握何时何地需要并可以为汽车充电，能源供应商则能够对电力负荷进行相应的调整。

利用大数据，以区域和时间两个维度，通过LBS（基于地理位置的服务）开放平台分析手机用户的定位信息，采用可视化呈现方式，动态、即时、直观地展现中国春节前后人口大迁徙，能够映射出手机用户的迁徙轨迹，可用于观察当前及过往时间段内，全国总体迁徙情况，以及各省、市、区的迁徙情况，直观地确定迁入人口的来源和迁出人口的去向。

一个大规模生产、分享和应用数据的时代正在开启，观察数据的方式在发生着巨大的变化。如今数据不再是一种静态的可支配资源，其意义不再像以往那样局限于一种单一的目的，而是已经成为延伸至多种功能用途的数据处理。正是由于数据作为商业的一种原材料也和其他生产的原材料一样，能够被应用于各种各样的领域而使得其价值超越了原始产品本身。

探究活动

实践

得益于电子商务系统丰富的生态和业务，电子商务公司积累了包括电商交易、搜索、物流、支付、广告、风控、电影、移动、视频、音乐、位置等种类多样的高质量数据，为探索与发挥大数据价值提供了极大的方便。同学们可上网了解电子商务数据的多样性及其应用场景。

5.1.2　数据管理技术新进展

以数据库为代表的数据管理技术已经历了近半个世纪的大发展。数据管理技术已经从第一代的层次与网状数据库系统、第二代的关系数据库系统，发展到新一代数据库，人们在不断努力开发能满足最新需求的数据库管理系统。

1. 数据库应用新需求

数据库技术是计算机科学技术中发展非常快的领域，也是应用最广的技术之一，随着用户要求的多样化和复杂化，数据库应用领域不断扩展，新的应用需求不断涌现。数据库应用新需求如表5-2所示。

表5-2 数据库应用新需求

新需求	特点	应用场景
数据类型多样化	不仅包括传统的数字、字符、文本等，还需要视频、音频、图形、图像、动画，以及HTML/XML、流数据等更复杂的数据类型。	大型图书馆管理系统就需要满足多种媒体的巨大数据量的统一管理。
数据结构新需求	结构化、半结构化、非结构化。	Web应用已成为许多企业提供服务的常用方式，而企业希望能对其大量的Web文件和数据进行统一管理。
数据存储新需求	海量、多维性等。	某大型互联网公司的业务分布在世界各地，面对的客户数量庞大，需要对分布在不同地方的数据进行统一管理，希望既能让各部门管理各自的数据，同时又可以共享其他部门的有关数据等。
数据操作新要求	不仅包含通常意义下的插入、删除、修改、查询等，还需要互操作（例如视频快进操作等）、主动性操作、领域搜索浏览、时态查询等，还要能够进行自定义操作。	一些企业希望自己的系统不仅仅是根据用户的要求被动地提供服务，还希望系统能根据库存不足或超量、生产进度偏离、财务费用超支、生产事故等异常事件，主动发出警告或进行相应应急处理等。
其他需求	领域需求。	为了适应数据库应用多元化的要求，在传统数据库的基础之上，结合各个应用领域的特点，研究和开发适合各个应用领域的数据库新技术，如工程数据库、统计数据库、科学数据库、空间数据库等。

2. 数据库新技术

数据库新技术是一个不断发展的范畴，在数据模型的改进、与相关技术融合以及面向应用领域等方面都在不断改进与发展。

（1）数据模型的改进。

相对于传统的数据库而言，在数据模型及其语言、事务处理与执行模型、数据库组织与物理存储等各个层面，都集成了新的技术、工具与机制的有：

①面向对象数据库模型。

面向对象数据库是数据库技术与面向对象程序设计方法相结合的产物，由于结合了面向对象方法学，所以具有了所有面向对象的优点。同时，由于数据库主要操作的是集合（而不是单个数据），所以又具有自身的特点和优点。应用面向对象数据库，可以提高数据库开发效率，完成复杂的逻辑运算，保证低冗余性与高效性，以及提高软件的可重用性。

②时态数据库系统。

时态数据的形式特征是其由不显含时间的数据和相应的时间标签组成，而其本质需要

将数据本身与特定的时间（例如数据的生命周期等）紧密结合，将时间的处理和数据的管理相融合。因此，常规数据库就不能有效进行时态数据的管理。一般认为，一个时态数据库管理系统需要支持以下功能：

a. 一种时态数据定义语言。

b. 一种时态数据操作语言。

c. 一种时态查询语言。

d. 时态约束，比如时态外键一致性约束。

③实时数据库系统。

实时数据库技术是实时系统和数据库技术相结合的产物。概括地讲，实时数据库就是其数据和事务都有显式定时限制的数据库，系统的正确性不仅依赖于事务的逻辑结果，而且依赖于该逻辑结果所产生的时间。实时数据库系统主要特性包括及时性、可预测性和可靠性等。

实时数据库并非实时系统和数据库技术在概念、结构和方法上的简单集成。它在概念、理论、技术、方法和机制方面具备自身特点。例如，数据和数据库的结构与组织；数据处理的优先级控制、调度和并发控制协议与算法；数据和事务特性的语义及其与一致性、正确性的关系；数据查询/事务处理算法与优化；I/O调度、恢复和通信的协议与算法等，这些问题之间彼此高度相关。

④主动数据库系统。

主动数据库系统是数据库技术与基于知识的系统（或广义地说是人工智能系统）技术相结合的产物，能够提供某种主动性的操作和服务。实现主动数据库系统需要解决许多关键的问题，这些问题包括实现有效的事件监视器，有效的规则表示和执行机制，数据库中的事件描述、运算和复合，以及在主动数据库中的有效事务处理机制等。

（2）数据库与相关技术结合。

传统的数据库技术和其他技术的有机结合、互相渗透，使得数据库技术新的内容层出不穷。数据库中的某些概念、技术内容、应用领域，甚至某些原理都有了重大的变化，不同数据库系统如图5–3所示。

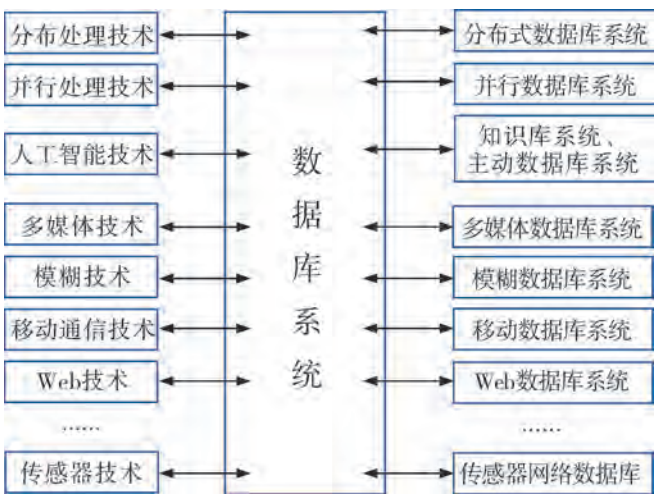


图5–3　数据库技术与其他计算机技术的相互渗透

下面简单介绍一下分布式数据库与Web数据库。

①分布式数据库基本特征。

a. 物理分布性——数据库中数据存储在不同计算机设备当中。

b. 逻辑整体性——数据在物理上分散存储，但在逻辑上相互关联，构成整体，数据被所有用户（全局用户）共享，由一个分布式数据库管理系统（Distributed Database Management System，DDBMS）统一管理。

c. 场地自治性——各个场地数据由本地数据库管理系统（Database Management System，DBMS）管理，具有自治处理能力，完成本场地的应用（局部应用）。

d. 场地间协作性——各个场地高度自治，但又相互协作构成一个整体。对用户来说，使用分布式数据库系统（Distributed Database System，DDBS）如同使用集中式数据库一样，可以在任何一个场地执行全局应用。

②Web数据库基本特征。

a. 能够在多平台、多操作系统上应用。

b. 能够通过网络实现数据库的远程存取和动态交互。

c. 提供通用的图形用户接口界面。

d. 实现基于WWW标准接口的网络数据库的开发，提高二次开发的简捷性，使操作简单、维护方便。

（3）面向应用领域。

为了适应数据库应用多元化的要求，在传统数据库的基础上，结合各个应用领域的特点，研究和开发适合各个应用领域的数据库新技术，如工程数据库、科学数据库、统计数据库、空间数据库、数据仓库等，如图5-4所示。

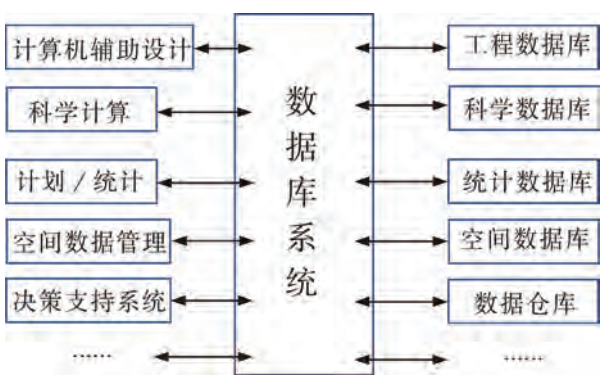


图5-4　面向应用领域的数据库

（4）非结构化数据库。

随着网络技术的发展，特别是Internet和Intranet技术的飞快发展，大数据的应用，使得非结构化数据的数量日趋增大。目前流行的有ibase，Hbase，TRIP等非结构化数据库。例如，ibase具有以下特点：

①Internet应用中，存在大量的复杂数据类型，ibase通过其外部文件数据类型，可以管理各种文档信息、多媒体信息，并对各种具有检索意义的文档信息资源，如HTML，DOC，RTF，TXT等提供强大的全文检索能力。

②ibase采用子字段、多值字段及变长字段的机制，允许创建许多不同类型的非结构化的或任意格式的字段。

③ibase将非结构化和结构化数据都定义为资源，使得非结构化数据库的基本元素就是资源本身，而数据库中的资源可以同时包含结构化和非结构化的信息。所以，非结构化数据库能够存储和管理各种各样的非结构化数据，实现了数据库系统数据管理到内容管理的转化。

④ibase采用了面向对象的基石，特别适合表达复杂的数据对象和多媒体对象。

⑤ibase提供一个网上资源管理系统ibase Web，将Web Server和Database Server直接集成一个整体，实现数据库和Web的有机无缝组合，从而为网上进行信息管理和开展电子商务应用开辟了更为广阔的领域。

⑥ibase全面兼容各种大、中、小型的DB，对传统关系数据库，如Oracle，SQL Server，DB2，Informix等提供导入和链接的支持能力。

讨论

分小组了解电子商务数据的管理技术与工具。云数据库RDS中的多结构数据存储，如图5-5所示。在数据类型多样的应用中，可将高热存取数据存储于缓存产品，如云数据库Memcached版 、云数据库Redis版，将图片等非结构化资源存储于对象存储OSS，而将链接等结构化数据存储于RDS，实现对业务数据高效存取，并相应降低成本投入。试讨论一下，使用什么数据管理技术与工具才能更方便地处理上述的存储数据。

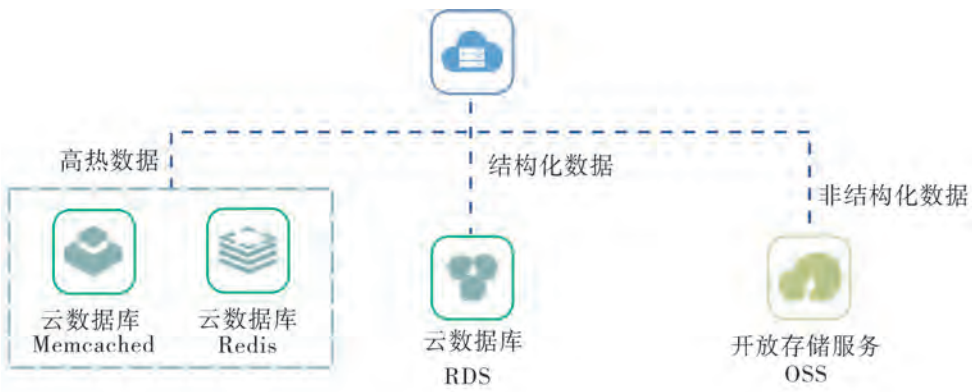


图5-5　多结构数据存储

5.1.3　数据分析技术新进展

1. 数据分析新需求

信息技术沿着以个人计算机为核心、到以互联网为核心、再到以数据为核心的发展脉络，逐步改变着社会的经济结构和生产方式，加快了全球范围内的知识更新和技术创新。

数据分析的对象是数据，但数据的价值，随着数据量的指数级数增长，已经不能够通过传统的图表得以显现，价值隐藏在庞杂的数据中，需要使用更加合理的数据分析方案才可释放这些价值。

人类思维方式在发生着变革。互联网已经渗透到社会的各行各业，就像电力和道路一样，互联网正在成为现代社会真正的基础设施之一。互联网不仅仅是用来提高效率的工具，还是构建未来生产方式和生活方式的基础设施。互联网思维已成为一切商业思维的起点，在这种思维方式下，人们也在不断地提出新的数据分析需求。

（1）精细化运营的需求。

大数据时代的各行各业，尤其是互联网行业都面临着日新月异的竞争格局，为提升企业的效益和效率，必须寻找比传统的粗放型运营更加有效的精细化运营思路，那就需要更细分、更准确、更个性化的数据分析策略。

（2）实时应用的需求。

以购物网“双十一”“双十二”为例。商家会在购物网上或者在店铺内投放相应的广告来吸引客户，同时，商家也可能会准备多个广告样式、文案，根据广告效果来做出调整，这就需要对广告的点击情况、用户的访问情况进行分析，但以往这类分析采用分布式离线分析，分析结果都有几小时甚至一天的延时，而“双十一”“双十二”的促销活动通常持续时间只有一两天，延时得到的分析结果满足不了用户的实时分析响应，便失去了价值。

（3）大规模图数据的需求。

许多大数据都是以大规模图或网络的形式呈现，如社交网络、传染病传播途径、交通事故对路网的影响等，已有的图计算框架和图算法库不能很好地满足大规模图的计算分析需求。

2. 数据分析新技术

随着数据量的日渐剧增，新需求与新应用也不断产生，在数据分析方面，人们已不满足于简单的数据查询、统计等传统数据事务处理，提出了如知识发现、决策支持等更加高级的应用需求，这些需求推动了机器学习、数据挖掘、分布式并行编程模型、计算框架等技术的产生和发展。

分布式并行编程与传统的程序开发方式有很大的区别。传统的程序都是以单指令、单数据流的方式顺序执行，这种程序的性能受到单台机器性能的限制，可扩展性差。分布式并行程序可以在大量计算机构成的集群上运行，从而充分利用集群的并行处理能力。

数据处理的问题复杂多样，单一的计算模式是无法满足不同类型的计算需求的，目前大数据计算模式有批处理计算、流计算、图计算及查询分析计算等。批处理计算的代表产品有MapReduce，Spark等；流计算是针对流数据的实时计算分析，代表产品有Storm，S4，Flume，Stream，Puma，DStream，Super Mario，银河流数据处理平台等；图计算是针对大规模图结构数据的处理，代表产品有Pregel，GraphX，Giraph，PowerGraph，Hama，GoldenOrb等；查询分析计算代表产品有Dremel，Hive，Cassandra，Impala等。

项目实施

各小组根据项目选题及拟订的项目方案，结合本节所学知识，了解数据管理与分析技术的新发展。

1. 通过互联网等各种方式了解数据的多样性及应用场景。
2. 选择合适的平台进行数据产品的实测体验。

5.2 数据挖掘与大数据的意义

5.2.1 数据挖掘的意义

1. 数据挖掘的发展历史

随着数据库与互联网技术的发展应用，数据的积累不断膨胀，导致简单的查询和统计已经无法满足一些企业的商业需求，当数据量极度增长时，如果没有有效的方法来提取有用的信息和知识，人们面对信息海洋就会感到像大海捞针一样束手无策，收集在大型数据库中的数据就变成了“数据坟墓”，因而急需一些革命性的技术去挖掘数据背后的信息。由于人工智能取得了巨大进展，进入了机器学习的阶段，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，尝试挖掘数据背后的信息，从而催生了一门新的学科，即数据库中的知识发现（Knowledge Discovery in Databases，KDD）。在1989年8月召开的第11届国际人工智能联合会议的专题讨论会上首次出现了“知识发现”这个术语。而数据挖掘（Data Mining）则是知识发现的核心部分，它指的是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程。

总体来说，数据挖掘是一门广义的交叉学科，它融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的理论和技术。

2. 数据挖掘技术的应用

数据挖掘工具和软件已在很多领域得到了很好的应用，并收到明显的效益，表5-3是数据挖掘在部分领域的应用。

表5-3 数据挖掘在部分领域的应用

领域	数据挖掘的应用
金融行业	预测存/贷款趋势，优化存/贷款策略，科学的股票投资等，用DM将市场分成有意义的群组和部分，从而协助市场经理和业务执行人员更好地开展有促进作用的活动或设计新的市场模式。
交通航空行业	智能交通系统是近年来迅速发展的控制管理城市道路、高速公路与铁路的新技术。数据挖掘技术已广泛应用于城市交通流量预测、高铁票价的制定及高铁轨道安全性检测等方面。
电信行业	市场分群、精确营销、新业务响应及客户流失等方面的数据，用DM中聚类、决策树等技术可以进行有效的管理。
过程控制/质量监督保证方面	DM协助管理变量之间的相互作用；自动发现某些不正常的数据分布；暴露制造和装配操作过程中的异常情况和消极因素，从而协助质量工程师很快地注意到问题发生的范围并及时采取改正措施。
体育行业	数据挖掘在竞技体育中的应用由来已久，例如：NBA的教练利用数据挖掘工具来优化他们的技术战术安排；中国女排2016年奥运会夺冠背后也有强有力的数据分析。
互联网行业	相对于传统行业而言，互联网行业具有新应用源源不断、数据海量性、数据挖掘的周期短、数据挖掘成果的时效性短等特点，相应地对数据分析挖掘的应用需求也更为苛刻。

总之，数据挖掘可广泛用于金融、保险、电力、交通、远程通信、零售与批发、制造业、公共设施、行政管理、教育、体育、国防等多个领域，数据挖掘对信息社会问题解决与科学决策具有重要的意义。

探究活动

阅读

电子商城商品推荐

利用关联规则发现共现关系（挖掘频繁项目集），这个应用场景放在购物上是非常合适的，如大量的用户购买了A产品之后还会接着购买B和C产品，于是一旦发现用户购买了A产品，系统就会给用户推荐B和C产品。比如购买了数码相机，系统会推荐SD卡、数码相机电池等。

交流

近年来，人工智能、机器学习、神经网络是数据挖掘技术研究人员关注的热点。图5-6是机器学习兴起的因素，有兴趣的同学可以去了解一下：机器学习是如何实现的？

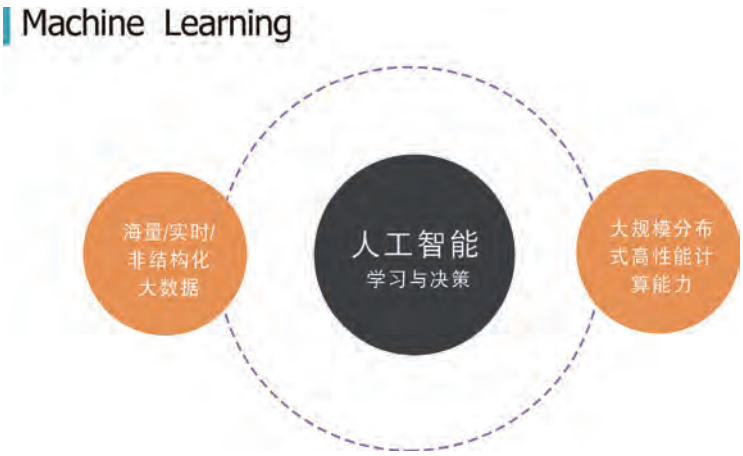


图5-6　机器学习兴起的因素

5.2.2　大数据的意义

1. 大数据的发展历程

从大数据的发展历程看，总体上可以划分为三个重要阶段：萌芽期、成熟期和大规模应用期，如图5-7所示。

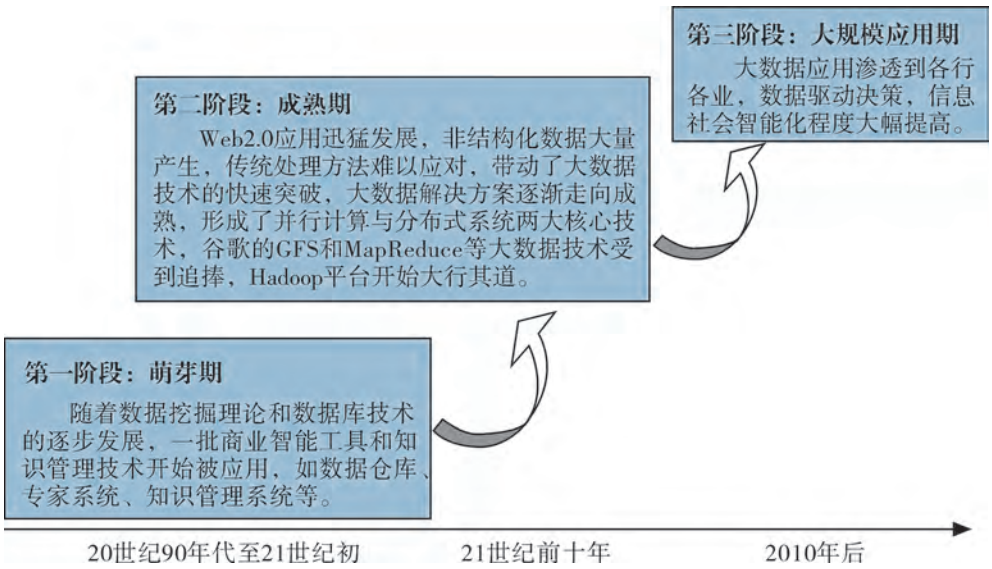


图5-7　大数据的发展历程

2. 大数据的应用

大数据无处不在，包括金融、保险、汽车、零售、餐饮、电信、能源、政务、教育、医疗、体育、娱乐等在内的社会各行各业都已经留下大数据的印迹。表5-4是大数据在部分领域的应用情况。

表5-4 大数据在部分领域的应用

领域	大数据的应用
金融行业	大数据在高频交易、社交情绪分析和信贷风险分析三大金融创新领域发挥重要作用。
汽车行业	利用大数据和物联网技术的无人驾驶汽车，在不远的未来将走入我们的日常生活。
互联网行业	借助大数据技术，可以分析客户行为，进行商品推荐和有针对性的广告投放。
餐饮行业	利用大数据实现餐饮O2O模式，彻底改变传统餐饮经营方式。
电信行业	利用大数据实现客户离网分析，及时掌握客户离网倾向，出台客户挽留措施。
能源行业	利用大数据技术分析用户用电模式，可以改进电网运行，合理地设计电力需求响应系统，确保电网运行安全。
物流行业	利用大数据可以优化物流网络，提高物流效率，降低物流成本。
城市管理	利用大数据实现城市改造规划、智能交通、环保监测、智能安防等。
生物医学	利用大数据可以实现流行病预测、智慧医疗、健康管理等。
体育行业	利用大数据可以实现训练球队、赛事排兵布阵、预测比赛结果等。
娱乐行业	利用大数据可以决定投拍哪种题材的影视作品，预测收视率等。
安全领域	利用大数据，政府可以构建强大的国家安全保障体系，公安可以借助大数据来预防犯罪，企业可以利用大数据抵御网络攻击等。
个人生活	利用与每个人相关联的“个人大数据”，分析个人生活行为习惯，为其提供更加周到的个性化服务。

拓展

电子商务平台知道谁需要贷款

每天，海量的交易和数据在电子商务平台上运行着，平台通过分析商户最近100天的数据，就能知道哪些商户可能存在资金问题，此时贷款平台就有可能出马，同潜在的贷款对象进行沟通。那么，该电子商务公司通过海量数据分析就能得出企业的经营情况，这也是大数据的应用之一。

3. 大数据的影响

大数据为信息技术的发展带来了巨大变革，其影响力和作用力正迅速触及社会的各个角落。全球范围内，世界各国政府均高度重视大数据技术的研究和产业发展，并把大数据上升到国家战略加以重点推进。2015年8月31日，国务院发布的《促进大数据发展行动纲要》认为，坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

大数据对科学研究、思维方式和社会发展都具有重要而深远的影响。在科学研究方面，大数据使得人类科学研究在经历了实验、理论、计算三种范式之后，迎来了第四种范式——数据。在思维方式方面，大数据具有“全样而非抽样、效率而非精确、相关而非因果”三大显著特征。这完全颠覆了传统的思维方式。在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现。总之，大数据对信息社会问题解决与科学决策具有深远的影响与重要的意义。

全国首创广东省制造业大数据指数（MBI），是贯彻落实“实施国家大数据战略加快建设数字中国”要求的具体举措，是深入贯彻“四个走在全国前列”重要指示精神的具体行动。它的创新点和应用价值主要体现在：一是利用大数据手段有效突破数据孤岛，汇聚了与制造业密切相关的企业的用电、商品进出口、货运、贷款、用地、通信、用工等第三方一手海量数据，确保数据客观、真实；二是运用机器学习等大数据方法，构建一个有别于传统经济分析框架的指数模型，形成一套具有科学性和时效性的全新制造业发展评价体系；三是利用红绿灯图等可视化手段初步实现了对制造业宏观、中观、微观的运行监测预警，尤其在精准发现区域，行业、企业异动方面效果明显。下一步广东将把其打造成为广东大数据工作品牌，形成运用大数据推进国家治理体系和治理能力现代化的广东经验。

近年来，互联网和电子商务急剧发展。自2009年创建“双十一”以来，每逢“双十一”，电子商务平台销售额都有显著增长。请同学们将近几年某电子商务平台的“双十一”数据做一个汇总。

有些数据分析师做了不少的分析工作，也完成了不少的分析项目，自己觉得已经很努力，却收不到预期的效果。这主要是因为从业务方看来，这些内容并不是业务方所需要的，自然也就得不到他们的理解和支持了，更不会在实际应用中体现出明显的商业价值。所以，数据分析师在这种跨专业、跨团队、跨部门的开放合作式运营环境下，不能局限于

自己眼前的技术，而应该先明确业务的痛点，而后按图索骥找到良好的方案，满足业务需求，实现商业价值和自身价值。这就要求数据分析师要经常换位思考，从业务方的角度思考数据分析，也就是说要学会换位思考。

2012年，国内一些大的电子商务企业先后宣布设置首席数据官的岗位，并将其作为企业的核心管理岗位之一。数据分析师作为数据管理团队中的一员，应具备什么样的素质与技能呢?

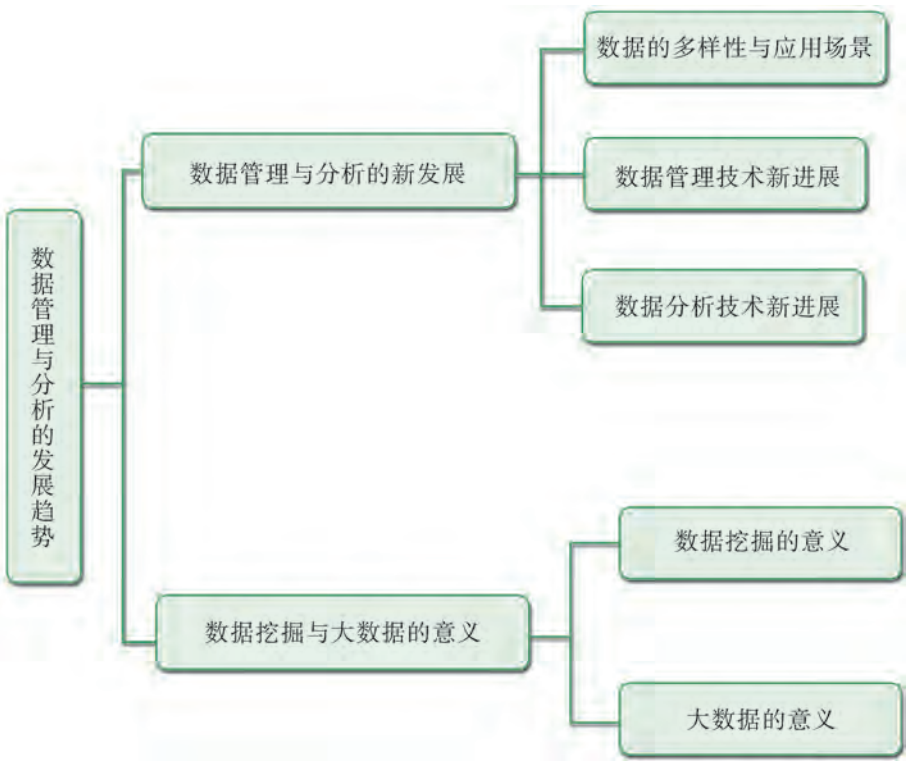
各小组根据项目选题及拟订的项目方案，结合本节所学知识，阅读相关资料，交流数据挖掘与大数据的意义，选择合适的平台进行体验，并参照项目范例的样式，撰写相应的项目成果报告。

各小组运用数字化学习工具，将所完成的项目成果，在小组或班级上进行展示与交流，共享创造、分享快乐。

各小组根据项目选题、拟订的项目方案、实施情况以及所形成的项目成果，利用教科书附录2的“项目活动评价表”，开展项目学习活动评价。

本章扼要回顾

同学们通过本章学习，根据“数据管理与分析的发展趋势”知识结构图，扼要回顾、总结、归纳学过的内容，建立自己的知识结构体系。



回顾与总结

本章学业评价

同学们完成下列测试题（更多的测试题可以在教科书的配套学习资源包中查看），并通过“本章扼要回顾”以及本章的项目活动评价，综合评价自己在信息技术知识与技能、解决实际问题的过程与方法，以及相关情感态度与价值观的形成等方面，是否达到了本章的学习目标。

1. 单选题

（1）（　　）提供的支撑技术，有效解决了大数据分析、研发的问题，比如虚拟化技术、并行计算、海量存储和海量管理等。

A. 线计算　　B. 云计算　　C. 点计算　　D. 面计算

（2）某超市研究销售记录数据后发现，买啤酒的人很大概率也会购买尿片，这种属于数据挖掘的（　　）问题。

A. 关联规则发现　　B. 聚类　　C. 分类　　D. 自然语言处理

（3）大数据的本质是（　　）。

A. 挖掘　　B. 联系　　C. 收集　　D. 洞察

2. 思考题

随着用户要求的多样化和复杂化，数据库应用领域不断扩展，新的应用需求也在不断扩展，数据库应用新需求主要体现在哪些方面？

3. 情境题

随着大数据、云计算、物联网等新技术向商业领域的渗透，越来越多的企业要素被数字化和数据化，数据已经成为企业的重要资源和生产要素。大数据正在成为商业的基础，成为一切管理和决策的依据。国内一些著名的互联网企业在拥有海量的用户数据之后，开始着手开展各类数据分析工作，用以支撑自身的电子商务、定向广告和影视娱乐等业务。

例如，影视动画、电影特效、建筑设计表现、城市规划、游戏片头动画、商业广告等多领域对海量渲染计算与数据传输有着巨大的需求，以往通常借助于高性能计算实现，但是存在部署复杂、缺乏灵活调度、总成本较高等问题。批量计算可支持海量作业并发规模，自动完成资源管理、作业调度和数据加载，并按实际使用量计费，以满足3D渲染的应用需求。

请同学们谈谈大数据分析的魅力，并展示一个“批量计算”的实例。

附录1　部分术语、缩略语中英文对照表

ADBS（Active DataBase System）	主动数据库系统（1）
Data Analysis	数据分析（4）
DBMS（Database Management System）	数据库管理系统（1）
DCL（Data Control Language）	数据控制语言（3）
DDL（Data Definition Language）	数据定义语言（3）
DM（Data Mining）	数据挖掘（5）
DML（Data Manipulation Language）	数据操纵语言（3）
DW（Data Warehouse）	数据仓库（1）
Entity-Relationship Model	实体—联系模型（2）
Hierarchical Model	层次模型（2）
KDD（Knowledge Discovery in Databases）	知识发现（5）
Machine Learning	机器学习（5）
Network Model	网状模型（2）
OODBS（Object-Oriented Database System）	面向对象数据系统（1）
RTDBS（Real-Time Database System）	实时数据库系统（1）
QL（Query Language）	查询语言（3）
semi-structured data	半结构化数据（1）
SQL（Structured Query Language）	结构化查询语言（3）
structured data	结构化数据（1）
TDBS（Temporal Database System）	时态数据库系统（1）
unstructured data	非结构化数据（1）

附录2　项目活动评价表

以培养信息素养为目标，以知识体系为载体，以项目学习活动过程与评价为途径，促进同学们的信息技术学科核心素养达成。

项目学习主题：_____

项目学习过程	学科核心素养达成	一级指标	二级指标	评价结果	支撑材料
选定项目	从现实世界中选择明确的项目主题，形成对信息的敏感度和信息价值的判断力。 分析项目目标与可行性。	项目选题	从现实世界选择项目主题的能力。 化抽象概念为现实问题的能力。 对信息的敏感度和价值的判断力。	□优秀 □良好 □中等 □仍需努力	
		项目分析	分析项目目标的能力。 分析项目可行性的能力。 从现实世界发现项目素材的能力。	□优秀 □良好 □中等 □仍需努力	
规划设计	组建团队与明确项目任务，体现正确的信息社会责任意识。 规划项目与交流方案。	项目规划	组建团队与明确项目任务的能力。 规划项目学习工具与方法的能力。 预期项目成果的能力。	□优秀 □良好 □中等 □仍需努力	
		方案交流	交流项目方案的能力。 完善项目方案的能力。 体现正确的信息社会责任意识。	□优秀 □良好 □中等 □仍需努力	
活动探究	通过团队合作，围绕项目进行自主、协作学习。 开展探究活动，提升信息获取、处理与应用、创新能力。	团队合作	自主学习能力。 分工与协作能力。 交流与沟通能力。	□优秀 □良好 □中等 □仍需努力	
		探究活动	信息获取与处理能力。 探究与联想能力。 实践与创新能力。	□优秀 □良好 □中等 □仍需努力	

（续表）

项目学习过程	学科核心素养达成	一级指标	二级指标	评价结果	支撑材料
项目实施	针对给定的任务进行分解，明确需要解决的关键问题，并采用计算机科学领域的思想方法，在形成问题解决方案的过程中产生一系列思维活动。 完成方案中预设的目标。	工具方法	采用计算机领域的思想方法能力。 使用数字化工具与资源能力。 数字化学习能力。	□优秀 □良好 □中等 □仍需努力	
		实施方案	针对给定的任务进行分解。 明确需要解决的关键问题。 完成方案中预设的目标。	□优秀 □良好 □中等 □仍需努力	
项目成果交流与评价	与团队成员共享创造与分享快乐，提升批判性思维能力与信息社会责任感。 评价项目目标与成果质量效果。	成果交流	清晰表达项目主题与过程。 与团队成员共享创造与分享快乐。 提升批判性思维能力与信息社会责任感。	□优秀 □良好 □中等 □仍需努力	
		项目评价	运用新知识与技能实现项目目标。 项目成果的可视化表达质量。 项目成果解决现实问题效果。	□优秀 □良好 □中等 □仍需努力	
综合评价	□优秀　□良好　□中等　□仍需努力				

注：1. 评价得分90～100分为优秀（A）；75～89分为良好（B）；60～74分为中等（C）；60分以下为仍需努力（D）。

2. 综合得分=互评×30%+自评×30%+教师评×40%。

绿色印刷产品

批准文号：粤发改价格［2017］434号　　举报电话:12358

ISBN 978-7-5548-3030-7

定价：11.13元