



普通高中教科书

XINXI  
JISHU

# 信息技术

选择性必修 3

数据管理与分析



教育科学出版社

普通高中教科书

XINXI  
JISHU

# 信息技术

选择性必修 3

数据管理与分析

教育科学出版社  
· 北京 ·

总 主 编 李 艺 董玉琦  
本 册 主 编 解月光  
本册副主编 张晓卉 于 颖  
主要编者 解月光 张晓卉 于 颖 王 海 褚 丹 徐庆鑫

出 版 人 李 东  
责任编辑 贾立杰  
版式设计 国美嘉誉文化 王 辉  
责任校对 贾静芳  
责任印制 叶小峰

## 普通高中教科书

### 信息技术 选择性必修 3 数据管理与分析

教育科学出版社出版发行  
(北京·朝阳区安慧北里安园甲9号)  
邮编: 100101

总编室电话: 010-64981290 编辑部电话: 010-64989637  
出版部电话: 010-64989487 市场部电话: 010-64989009  
传真: 010-64891796

网址: <http://www.esph.com.cn>

各地新华书店经销

北京市大天乐投资管理有限公司印装

开本: 890 毫米 × 1240 毫米 1/16 印张: 8.5  
2020 年 1 月第 1 版 2021 年 12 月第 5 次印刷

---

ISBN 978-7-5191-1307-0

定价: 14.85 元(含光盘)

批准文号: 京发改规〔2016〕13号 价格举报电话: 12315

图书出现印装质量问题, 本社负责调换。

图片来源: 高品(北京)图像有限公司

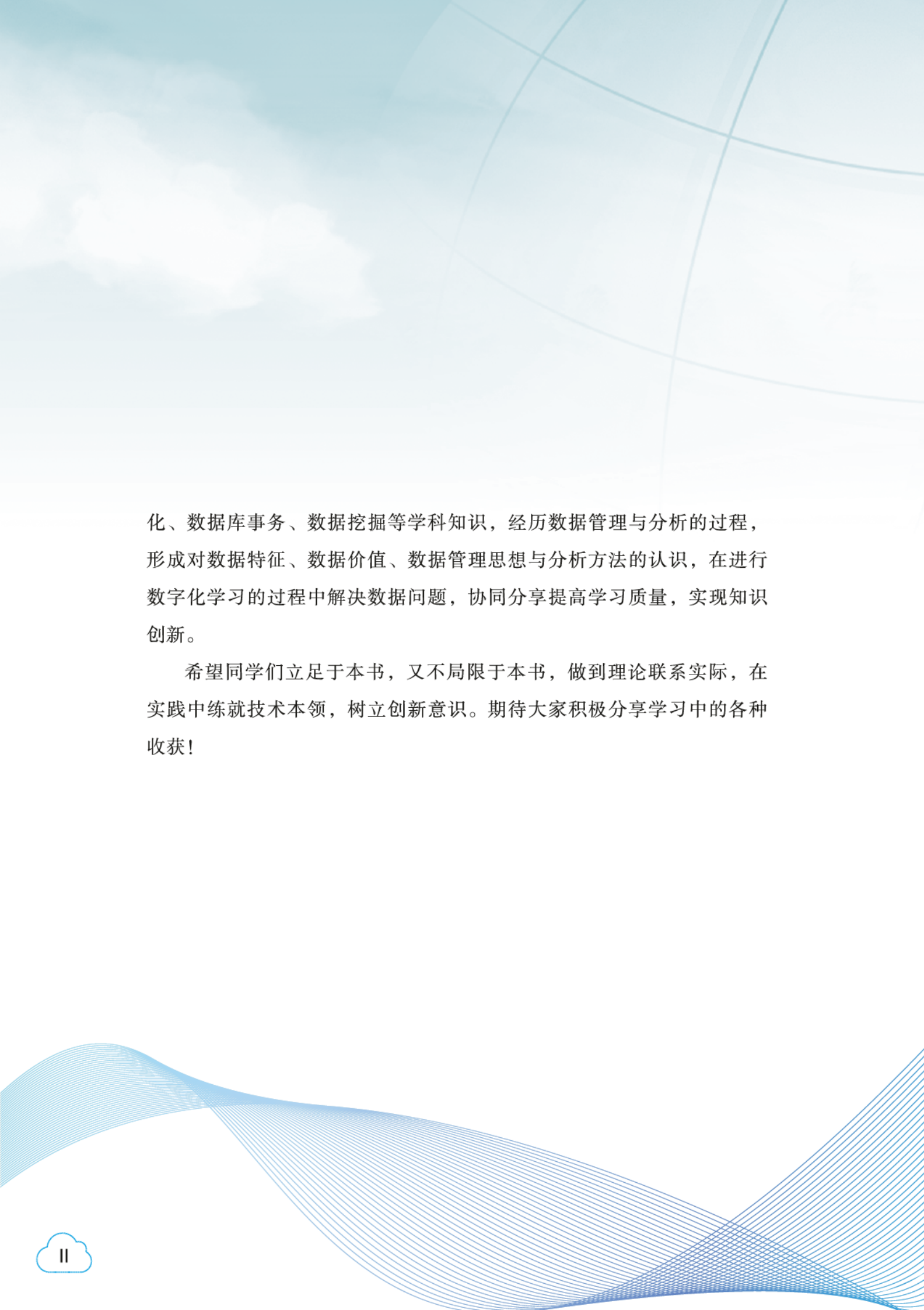
# 前 言

“数据创造价值！”

大数据时代，我们每天都会接触数据，我们自身也可能成为数据。“好奇心和思维方式是人类立足世界的根本”，好奇心让我们面对数据时，希望能从数据中发现它蕴涵的价值，思维方式能引领我们用正确的方法实现数据的价值。因此，在做出正确的决策之前，善于管理和分析数据是我们应具备的基本能力。

本书根据《普通高中信息技术课程标准（2017年版）》编写。全书以“网店的创办与运营”为主题展开，共分数据世界与问题解决、数据的收集与整理、数据的管理、数据的分析与可视化、数据安全、初识数据仓库与数据挖掘六个单元。全书以真实的项目任务和问题解决为主线组织教材结构，设计学习活动，选择学习内容，既有整体规划，也有单独设计，主旨明确，思路清晰，内容丰富而灵活。

本书旨在引导同学们自主学习，在项目实施、任务执行、活动开展、问题解决的过程中习得知识、提升能力、发展核心素养。同学们在学习过程中会扮演数据的管理者、设计者、使用者等角色，在技术多样、资源丰富的数字化环境下，掌握数据管理、数据分析、数据可视



化、数据库事务、数据挖掘等学科知识，经历数据管理与分析的过程，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，在进行数字化学习的过程中解决数据问题，协同分享提高学习质量，实现知识创新。

希望同学们立足于本书，又不局限于本书，做到理论联系实际，在实践中练就技术本领，树立创新意识。期待大家积极分享学习中的各种收获！



# 目 录

<b>第 1 单元</b>	<b>数据世界与问题解决</b>	<b>1</b>
1.1	认识数据管理与分析的重要性	2
1.2	需求分析与整体解决方案的制订	6
	单元学习评价	16
	单元学习总结	17
<b>第 2 单元</b>	<b>数据的收集与整理</b>	<b>19</b>
2.1	数据的收集	20
2.2	认识噪声数据	25
2.3	数据的结构化	30
	单元学习评价	34
	单元学习总结	35
<b>第 3 单元</b>	<b>数据的管理</b>	<b>36</b>
3.1	数据库的初识与剖析	37
3.2	数据库的设计与创建	46
3.3	数据库的操作与应用	54
	单元学习评价	65
	单元学习总结	66

<b>第 4 单元</b>	<b>数据的分析与可视化</b>	<b>68</b>
4.1	数据分析方法	69
4.2	数据分析工具	76
4.3	数据可视化	84
	单元学习评价	93
	单元学习总结	94
<b>第 5 单元</b>	<b>数据安全</b>	<b>96</b>
5.1	数据丢失与泄露的风险	97
5.2	数据备份与恢复	104
	单元学习评价	112
	单元学习总结	113
<b>第 6 单元</b>	<b>初识数据仓库与数据挖掘</b>	<b>114</b>
6.1	数据库和数据仓库	115
6.2	初识数据挖掘	118
6.3	数据挖掘的应用	121
	单元学习评价	125
	单元学习总结	126
<b>附 录</b>		<b>128</b>
	中英文术语对照表	128
<b>后 记</b>		<b>130</b>

# 第 1 单元 数据世界与问题解决

今天的我们正生活在一个“数据世界”之中，并受益于数据所带来的价值。例如，通过个性化的商品信息推送，我们能网购到称心如意的商品；通过在线订票系统，我们能便捷地预订车票、船票、飞机票等；通过实时交通流量监测，我们能成功避开道路拥堵；通过智能健康助手，我们能实时了解自己的运动状况……。以上数据的价值都是通过对数据进行科学管理与分析而实现的，如果缺乏科学有效的数据管理思想与分析方法，人们甚至会迷失在数据海洋中，为数据所累。

在本书中，我们将围绕“网店的创办与运营”，学会应用数据管理与分析技术解决问题，形成数据管理与分析的基本思想与方法。

制订问题整体解决方案，是网店创办与运营的第一步。在本单元中，我们将通过“制订网店创办与运营整体方案”项目展开学习，认识数据管理与分析的重要性，掌握问题整体解决方案制订的基本方法，初步形成分析、评价与优化方案的基本思想。

制订网店创办与运营整体方案时，主要考虑以下问题：网店创办与运营需要用到数据管理与分析技术吗？网店创办与运营有哪些业务需求？实现这些需求有哪些困难？数据管理与分析技术对解决困难有什么作用？如何设计网店创办与运营方案？制订出的方案有什么问题？如何改进？要解决这些问题，我们需要完成以下四个任务：

- ◆ 分析阿里云大数据体验馆中的案例；
- ◆ 分析网店业务需求；
- ◆ 设计网店整体方案；
- ◆ 评估网店整体方案。



## 1.1 认识数据管理与分析的重要性

我们生活在数据世界中，数据像土地、石油、水一样，已成为我们生产、生活中的重要资源。数据是会“说话”的，它可以帮助我们获取有用的信息，可以帮助我们做出判断，形成决策。但是，杂乱无章的数据是无意义的，只有对数据进行科学的管理和分析，才能实现其应有的价值。

### 学习目标

- ★ 理解数据是一种重要资源。
- ★ 体会数据管理与分析的重要性。
- ★ 认识有效管理与分析数据对实现数据价值的作用与意义。



数据体验馆有多种选择，可根据实际情况选取典型和有影响力的数据平台。例如，企业大数据平台、移动数据业务体验中心、大数据综合试验区展示中心、交通大数据监测中心等。

佳玮想创办一个主营中学生学习文化用品的网店。但是，由于缺乏创办网店的经验，她不清楚应该怎么做。她听说创办网店要利用数据管理与分析技术，是否有这个必要？让我们跟佳玮一起走进“数据体验馆”，一探究竟吧。



### 任务 分析阿里云大数据体验馆中的案例

#### ※ 活动1 走进阿里云大数据体验馆

使用淘宝账号登录，进入阿里云大数据体验馆。你发现了什么？请选择体验馆中与“场景体验”相关的栏目，浏览栏目内容，思考并讨论表1.1.1所列问题，将表1.1.1补充完整。

表 1.1.1 阿里云大数据体验馆体验列表

序号	思考并讨论	思考结果
1	在阿里云大数据体验馆里，你发现数据能做什么？	<ul style="list-style-type: none"> <li>◆ 生成报表</li> <li>◆ 个性化商品推荐</li> <li>◆ _____</li> </ul>
2	如果不进行数据分析，序号1中所说的结果能形成吗？数据分析给人们提供了什么帮助？	
3	数据管理对数据分析意味着什么？	

## ● 数据管理

数据管理是指利用计算机硬件和软件技术对数据进行有效的收集、分类、组织、编码、存储、查询和维护等活动。随着计算机技术的发展，数据管理经历了人工管理、文件系统管理、数据库系统管理三个发展阶段。

## ● 数据分析

数据分析是指用适当的分析方法对收集来的数据进行分析，提取有用信息和形成结论的过程。数据分析的目的是把隐藏在一大批看起来杂乱无章的数据中的信息进行集中、萃取和提炼，以找出所研究对象的内在规律，为人们做出判断、形成决策提供依据。例如，开普勒通过分析行星位置的观测数据，找出了行星运动规律；又如，企业领导人要通过市场调查，分析所得数据，以判定市场动向，从而制订合适的生产及销售计划。

数据分析是数学与计算机科学相结合的产物。数据分析的数学基础在20世纪早期就已确立，但直到计算机出现后才使得实际操作成为可能，并得以推广。



机器学习是研究如何使用机器来模拟人类学习活动的一门学科。它是人工智能的核心，其应用遍及人工智能的各个领域，如数据挖掘、计算机视觉、生物特征识别、搜索引擎、医学诊断、战略游戏和机器人运用等。如果你对机器学习感兴趣，可以深入探索机器学习的奥秘。

### ※ 活动2 分析阿里云大数据体验馆中的机器学习案例

进入“机器学习”栏目，分析栏目中的商品推荐类案例。按照表1.1.2中给出的要点对案例进行分析，并将表1.1.2补充完整，必要时可向老师请求帮助。

表1.1.2 机器学习案例分析要点与结果

案例类别	案例名称	分析要点	分析结果
商品推荐类		利用数据分析为顾客做个性化商品推荐，这对买卖双方有何影响？	买方：节约商品搜索与浏览时间，及时获取所需商品信息。 卖方：针对顾客需要，刺激顾客购买欲，提高顾客满意度，提高网店_____。
		个性化商品信息推荐是如何实现的？	个性化商品推荐信息是通过对大量顾客购物数据进行采集、利用特定的方法分析而得出的。
		佳玮的网店如果需要向顾客推荐商品信息，不用数据管理与分析能实现吗？为什么？	

## ● 数据管理与分析的作用

通过利用计算机硬件和软件技术，数据管理可以帮助我们有效地收集、组织、存储、处理和应用需要的数据，把原本看似杂乱无章的数据转换成可供利用的数据资源；数据分析又进一步帮助我们从小数据资源中提取出隐藏在数据背后的“有用”信息，“读”出数据的意义，为我们做出判断、形成行动决策提供依据。

通过上述体验，佳玮认识到网店的创办与运营离不开数据管理与分析技术。她初步形成了一份如表1.1.3所示的数据管理与分析体验报告，请帮她补充完整。

表1.1.3 数据管理与分析体验报告简表

主题	关注的问题	认知结果
数据管理的作用	杂乱无章的数据可以为我们所用吗？数据管理技术可以把这些杂乱无章的数据转换成可被利用的数据资源吗？	数据与资源间的关系：数据管理可以帮助我们有效地收集、组织、存储、处理和应用需要的数据，把原本杂乱无章的数据转换成可供利用的数据资源。
	数据管理对数据分析有什么作用？	数据管理与数据分析间的关系： _____
数据分析的作用	如何从数据资源中提取出有用的信息？	数据资源与数据分析间的关系：数据分析能够帮助我们从给定资源中“读出”有用信息。
	数据分析结果有何用处？是否有助于我们进一步做出决策？	数据分析与决策间的关系： _____

续表

主题	关注的问题	认知结果
总结	人与数据有什么关系?	人生活在数据世界中,离不开数据。因此,人们要掌握数据管理与分析技术,从数据中“读出”需要的信息。
	网店创办与运营需要数据管理与分析技术吗?为什么?	答案: _____ 理由: _____



## 拓展练习

生活中,我们经常看见、听到很多典型案例,感受着数据对人们生活的重要影响。比如,某大型电商通过分析所拥有的客户数据,预判客户的消费需求,以就近储备商品,保证客户下单后能以最快的速度将商品送达;搭乘公交车时,我们可以通过智能公交APP的实时数据监控,精准获取公交车进站时间,免去长时间候车或误车的困扰……,你有过这样的体验吗?

请选取生活中或网上的典型案例,说明数据管理与分析的作用,形成典型案例分析报告。

### 典型案例分析报告

#### 【选取案例介绍】

案例名称: \_\_\_\_\_ 案例主题: \_\_\_\_\_

案例目标: \_\_\_\_\_

案例简介: \_\_\_\_\_

#### 【案例分析要点】

1. 案例中形成的决策有依据吗? 决策依据是通过数据分析得出的吗?
2. 用于分析的数据是如何得来的?
3. 如果不利用数据管理与分析技术, 结果会怎样?

#### 【结论】

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

## 1.2 需求分析与整体解决方案的制订

制订整体解决方案是利用数据管理与分析技术解决问题的重要环节，它影响着问题解决的方向、进程和效果。需求分析是利用数据管理与分析技术解决问题、建立整体解决方案的第一步。



### 学习目标

- ★ 了解需求分析与制订问题整体解决方案的相关概念与基本流程。
- ★ 掌握制订问题整体解决方案的基本方法。
- ★ 初步形成分析、评价与优化方案的基本思想。



### 任务一 分析网店业务需求

经过对阿里云大数据体验馆的体验，佳玮认识到了数据管理与分析对于网店创办与运营的重要作用。她想请专业人士帮忙制订一个网店创办与运营的整体方案。让我们毛遂自荐，帮她来制订这个方案吧。我们首先需要进行网店业务需求分析，明确网店的业务需求。

#### ● 需求分析

在需求分析阶段，要全面、准确地分析用户的需求，明确问题的要求，确定利用数据管理与分析技术解决什么问题，并确认最终实现什么功能，达成什么效果。

#### ● 需求分析的基本步骤

在软件工程中，需求分析有明确的步骤，通常包括制订需求开发计划、开展需求调查与分析、需求验证、需求规则说明制作、需求确认。

软件工程需求分析的思想已渗透至各个领域的问题解决中。需求分

析通常可以按照以下步骤展开：提出业务需求—开展需求调研—需求协商与确认—需求变更管理（可选步骤）。

### ※ 活动1 调研网店的业务需求

创办网店需要考虑一系列需求问题，例如：现有资金有多少？创办的网店规模有多大？具体经营哪些商品？从哪里进货？促销活动如何组织？……要想弄清哪些是佳玮网店的业务需求，就需要进行需求调研。

#### ● 业务需求

业务需求通常来自项目投资人、购买产品的客户、实际用户的管理者、市场营销部门或产品策划部门。业务需求描述了为什么要开发一个系统和希望达到的目标。

#### ● 需求调研

需求调研是需求分析的重要环节，通常借助一定的需求引导方法，如与客户和领域专家面谈、使用调查表调研、观察和研究业务文档、头脑风暴法等，引导客户较为清晰地提出自己的业务需求。其中，头脑风暴法是较为常用的需求引导方法，在利益相关者对于业务需求比较模糊或者难以达成一致的情况下，头脑风暴法经常为需求分析者所使用。

佳玮的网店已经成立了筹建部、采购部、信息管理部、营销业务部、客服部、信息安全部等部门。但是，由于佳玮和伙伴们都是初次创办网店，他们对于网店的业务需求比较模糊，这给需求调研带来了一定难度。如果由你来负责需求调研工作，你将如何制订网店业务需求调研计划呢？请将下面的“网店业务需求调研计划”补充完整。

#### 网店业务需求调研计划

【调研时间】\_\_\_\_\_【调研地点】\_\_\_\_\_

【调研目标】提出网店的业务需求

【参与需求调研人员】\_\_\_\_\_

【主持人】（1名，熟悉头脑风暴法要点和操作要素）\_\_\_\_\_

【记录人】（1~2名）\_\_\_\_\_【调研对象】网店店主与各部门主管

【采用的需求引导方法及理由】

方法：头脑风暴法 理由：\_\_\_\_\_

【调研内容】网店各部门业务需求清单



**头脑风暴法**出自“头脑风暴”一词，主要指无限制地自由联想和讨论，其目的在于产生新观念或激发创新设想。使用头脑风暴法时，应遵循自由畅谈、延迟评判、追求数量、力求完善等基本原则。



在网店整体方案制订活动中，佳玮及其团队成为我们的“客户”。只有明确了佳玮团队的业务需求，才能有针对性地制订出网店整体方案。

事实上，与网店业务密切相关的还有另外一类“客户”——网店目标消费者。在进行需求调研时，也需要引导佳玮团队适当考虑目标消费者群体的需求。

## 【调研步骤】

会前准备：落实好主持人、记录人、调研对象和会议主题，主持人拟好引导提纲，必要时可预先对调研对象进行思维训练。

设想开发：在主持人的引导下，力争在有限的时间内，让各部门主管尽可能多地提出本部门的业务需求设想。

设想的分类与整理：整理、概括各部门初步提出的设想。

完善与确认设想：继续采用头脑风暴法对设想进行完善与确认。

请按照需求调研计划，进行分角色扮演。一组同学扮演需求调研人员，另一组同学扮演各部门主管，在需求调研人员的引导下，充分进行头脑风暴，设想网店的业务需求，并把设想的结果填入表1.2.1。

表1.2.1 经需求调研形成的网店业务需求清单

序号	部门名称	业务需求	序号	部门名称	业务需求
1	筹建部	网店规模	3	信息管理部	网店日常数据管理
		网店选址	4	营销业务部	营销策略
		网店资金筹集			
2	采购部	商品经营品类	5	客服部	客服设置
		货源	6	信息安全部	

## ※ 活动2 确认网店的业务需求

通过头脑风暴，大家设想了很多业务需求，其中哪些是佳玮的网店必需的？我们需要与佳玮及其团队进一步协商，设定需求的优先级，确认符合佳玮网店实际情况的业务需求。如果在需求协商与确认过程中遇到困难，还可通过访谈领域专家寻求帮助。请将佳玮团队协商确认后的结果填入表1.2.2。

## ● 需求的优先级

绝大多数项目没有足够的时间或资源来实现每个细节，必须决定哪些需求是必要的、哪些是重要的、哪些是可以不予重点考虑的。这就需要由客户负责设定需求的优先级，而需求分析人员应该协助客户确定需求的优先级。

## ● 需求确认

需求确认是指需求分析者与客户之间对需求进行协商与确认，对客户提出的不一致的需求或优先级较低的需求进行删改，或者增加客户暂时没有发现却非常有必要的需求。

表 1.2.2 网店业务需求协商确认清单

业务部门	业务需求	确认结果	增删理由
采购部	商品经营品类	无变动	无
	货源	无变动	无
客服部	客服设置	删除	网店规模较小, 客服工作暂由 客服部主管、副主管兼任
信息管理部	网店日常数据管理		
营销业务部	营销策略	无变动	无
	销售业绩		
信息安全部	网店数据安全	无变动	无
筹建部	网店规模	删除	已请专家咨询论证, 优先级较低
	网店选址	删除	
	网店资金筹集	无变动	无

在上述经过协商确认的网店业务需求中, 有没有计算机处理不了的或者不需要计算机处理就能解决的需求? 如果有, 需要剔除。请在表 1.2.3 中把要剔除的需求选出, 在  内标注 F, 并说明剔除理由。

表 1.2.3 网店业务需求确认清单


业务需求	剔除理由
<input checked="" type="checkbox"/> 网店资金筹集	不需进行数据管理与分析即可解决
<input type="checkbox"/> 商品经营品类	
<input type="checkbox"/> 货源	
<input type="checkbox"/> 网店日常数据管理	
<input type="checkbox"/> 营销策略	
<input type="checkbox"/> 销售业绩	
<input type="checkbox"/> 网店数据安全	

剔除超出范围的业务需求后, 业务需求得以最终确认。同时, 业务数据问题也得以进一步明确。接下来, 我们需要利用数据管理与分析技术, 为这些业务数据问题设计整体解决方案。



为了行文与工作方便, 暂且把这种围绕特定业务需求, 且需要通过数据管理与分析才能解决的问题简称为业务数据问题。



 迭代是重复反馈过程的活动，其目的通常是逼近所需目标或结果。每一次对过程的重复称为一次迭代，而每一次迭代得到的结果会作为下一次迭代的初始值。迭代是计算思维的思想与方法之一。

上述过程其实就是需求分析的过程。从提出业务需求到业务需求的最终确认，需要不断迭代，上述需求分析就经历了三次迭代的过程。



## 任务二 设计网店整体方案

### ※ 活动1 使用问题树理清问题解决思路

明确了网店业务数据问题之后，该如何设计网店整体方案呢？这时，可以借助问题分析工具来理清问题解决思路。问题树是一种常用的问题分析工具，请仿照“商品经营品类”子问题树的构造方法（如图1.2.1所示），分工协作，构造完成“网店创办与运营”问题树，并将图1.2.2补充完整。


#### ● 问题树


问题树又称逻辑树、演绎树或分解树，是将问题的所有子问题分层罗列，从最高层开始，逐步向下扩展。它把一个已知问题当成树干，然后考虑这个问题和哪些子问题有关。每想到一点，就给这个问题（树干）加一个“树枝”，并标明这个“树枝”代表什么问题。一个大的“树枝”上还可以有小的“树枝”，以此类推，直至找出问题的所有关联问题。

#### ● 问题树的构造步骤

问题树的构造步骤通常为：

- (1) 找出问题中存在的“核心问题”或“起始问题”；
- (2) 确定导致“核心问题”或“起始问题”的主要原因；
- (3) 确定“核心问题”或“起始问题”导致的主要后果；
- (4) 根据以上因果关系画出问题树；
- (5) 反复审查问题树，并根据实际情况加以补充和修改。

 从问题树的描述中可以看出，“网店创办与运营”是一个给定问题，其下又可分解为若干个子问题。围绕其中某个子问题，可形成相应的子问题树。“商品经营品类”就是“网店创办与运营”的一个子问题树。

 如果你对“商品经营品类”子问题树的构造有不同意见，可以对它进行修改。

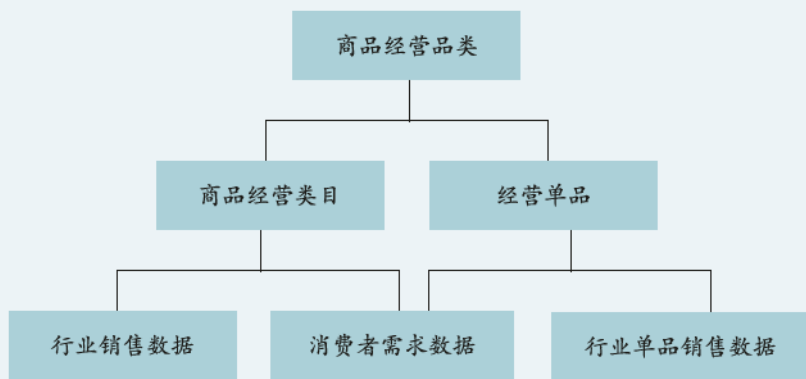


图1.2.1 “商品经营品类”子问题树

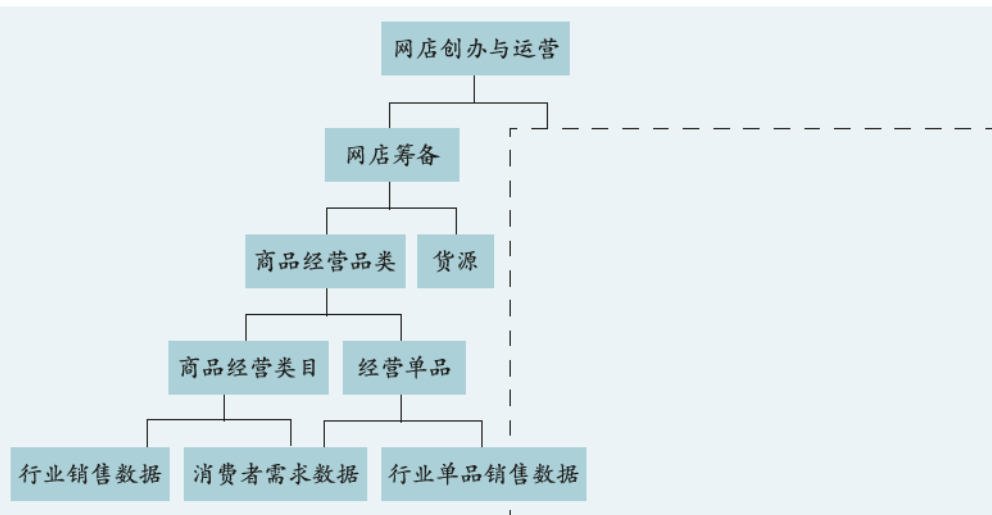


图 1.2.2 “网店创办与运营”问题树


在完成“网店创办与运营”问题树的构造之后，我们能清晰地找出与网店创办与运营相关的所有问题，并明晰解决各个问题所需要的关键数据。请结合问题树构造过程，把需要的关键数据填入表1.2.4。

表 1.2.4 网店整体方案需求数据清单

问题	问题描述	需要的关键数据
商品经营品类	经营哪些类别的哪些商品？	目标消费者需求数据与行业销售数据
货源	从哪里进货？货源如何确定？	
网店日常数据管理	网店运营产生的一系列数据如何管理？	
营销策略	使用什么样的营销/促销策略？	网店运营前期销售数据
销售业绩	卖得怎么样？销售业绩如何？	
网店数据安全	网店数据安全如何维护？	网店后台运营数据

### ※ 活动2 初步形成网店整体方案

问题树帮助我们理清了问题解决思路，明确了需要的数据，但还缺少解决问题的办法。因此，我们需要对照问题解决方案设计重点，依次考虑每个问题的解决办法，设计出网店整体方案，完善表1.2.5。

 使用问题树分析问题，要注意确认分支问题的关键问题和优先级，确认历史上该分支问题是否有解决方案可借鉴，必要时可安排分支问题的责任人和解决时限。

## ● 问题解决方案设计重点

1. 这个问题是谁的问题？（以谁为中心来设计）
2. 解决这个问题需要哪些数据？目前有哪些数据？已有的数据能否解决这个问题？
3. 需要从哪里获取哪些方面的数据？（考虑数据获得成本和风险）
4. 获取到数据后如何进行管理与分析？
5. 是否有多个可能的方案？
6. 问题解决的最佳方案是什么？（确保方案的完整性、合理性和可行性）

表1.2.5 “网店创办与运营”整体方案

问题	问题描述	解决方案	方案依据
商品经营品类	经营哪些类别的哪些商品？	选定被调查的学生群体和影响力较大的3个指数网站，利用调查法和互联网数据搜索法，收集目标消费者需求数据和行业销售数据，确定网店经营品类	目标消费者和行业整体销售数据与网店利益密切相关；指数网站数据获取容易，直接对目标消费者（学生群体）进行调查，数据信度高
货源	从哪里进货？ 货源如何确定？	通过网页采集器工具收集大型知名批发网站的单品批发数据，比对价格，确认质量，选定货源	大型知名批发网站有一定行业代表性和可靠性，且数据收集方便快捷，成本低，见效快
网店日常数据管理	网店运营产生的一系列数据如何管理？		选用可靠的数据管理应用软件，易实现，成本低，风险小
营销策略	使用什么样的营销/促销策略？	采用“0元购”促销策略，选用恰当的数据分析方法确定促销商品售价	仿照知名电商网站所采用的营销策略
销售业绩	卖得怎么样？ 销售业绩如何？		
网店数据安全	网店数据安全如何维护？	通过完全备份、增量备份、差异备份与日志备份对网店后台数据进行备份。如发生数据丢失的意外，选用合适的还原方法及时还原数据	



指数是一个经济学概念，可以用来反映经济现象的变化趋势与影响因素等。指数网站是指网站开发者利用程序自动收集数据并测算指数的网站。

### ※ 活动3 论证与改进网店整体方案

初步建立的方案能达成网店创办与运营的目标吗？有没有需要改进的地方？如何改进？请论证方案并发现方案存在的问题，提出改进建议，并说明改进理由，将方案改进记录填入表1.2.6。

## ● 方案论证与改进

初步建立的方案需要经过不断论证与改进，才能逼近问题的最佳方案。方案论证要重点考虑方案形成的依据、方案的可行性以及方案的优势与问题等。如果在方案论证的过程中不断发现新问题，可以通过问题解决方案改进记录表来记录方案的改进与优化过程。

表 1.2.6 问题解决方案改进记录表

问题方案	发现的问题	如何改进	改进理由
营销策略	采用“0元购”策略	根据促销规划与数据分析确定营销策略	“0元购”策略是假想策略，未经过数据分析与验证，无法预估实际收益

经历了建立网店整体方案这一过程之后，佳玮发现自己对数据管理与分析技术在问题解决过程中的作用有了更清晰的认识。于是，她归纳出了数据管理与分析的基本过程与可能涉及的技术工具，如图 1.2.3 所示。

她归纳得怎么样？你也是这样认为的吗？

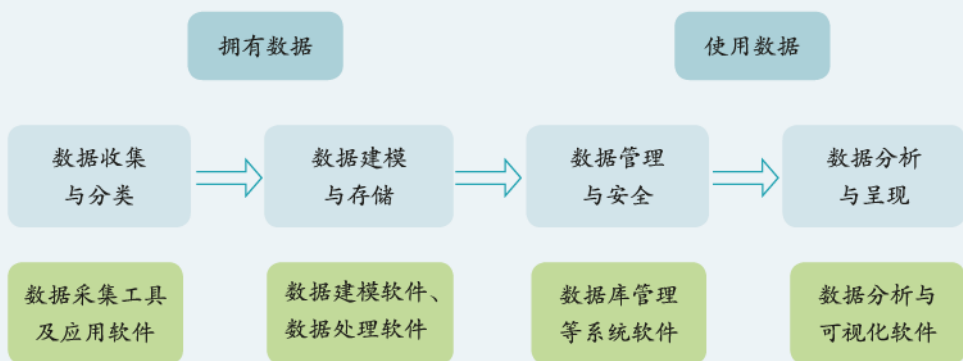


图 1.2.3 数据管理与分析的基本过程



## 任务三 评估网店整体方案

### ※ 活动1 评估拟订方案

初步拟订的网店整体方案怎么样？是否存在缺陷？请以小组为单位，各小组分别以采购部、营销业务部、信息管理部、信息安全部成员身份，自行评估这份方案的完整性、合理性和可行性，分析其中存在的问题，并提出改进意见，完善表 1.2.7。

## ● 方案评估



质性研究方法是以研究者本人为研究工具，在自然情境下采用多种资料收集方法，对研究现象进行整体性探究，从原始资料中形成结论，通过与研究对象互动，对其行为和意义建构获得解释性理解的活动。

在正式执行方案之前，要对方案进行评估。方案评估是运用系统性方法、质性研究方法、定量研究方法等来分析资料、搜集证据，客观地判断方案的成效与影响。

由于不同方案针对的领域及问题域不同，其评估标准也会有所不同。通常情况下，评估方案可以从方案的完整性、合理性和可行性等方面着手，主要是为了发现方案是否满足业务需求，提出的问题解决方案是否有合理性依据，方案是否能有效达成预期目标，是否具有良好的客户接受度，是否综合考虑了时间、成本、风险与质量等各项指标的约束与均衡，使用数据管理与分析技术是否能够实现方案，等等。

表1.2.7 网店整体方案评估表

指标	具体描述	存在的问题	改进意见
完整性	方案有较好的整体架构与规划意识，能够满足网店创办与运营的各项业务需求		
合理性	方案从需求分析出发，能为各项业务需求问题提供适切的方法和策略，能有效达成预期目标，具有良好的可操作性		
可行性	方案有较高的接受度，能被客户和利益相关者（如网店店主、营销策划部门等）接受；方案综合考虑了时间、成本、风险与质量等各项指标的均衡，使用数据管理与分析技术能够实现方案		

各小组评估完成后，请汇总各小组的评估表，形成方案评估总表。

同时，请在班级范围内展示与研讨方案评估总表，根据研讨意见完善评估总表，形成完整的方案评估表。

### ※ 活动2 根据评估意见优化方案

方案评估表形成后，需要对照方案评估表中存在的问题和改进意见逐一修改方案，对方案进行优化。请将优化后的网店创办与运营整体方案填入表1.2.8。

表1.2.8 优化后的网店创办与运营整体方案

问题	问题描述	解决方案	依据
商品经营 品类	经营哪些类别的哪些商品？		

续表

问题	问题描述	解决方案	依据
货源	从哪里进货？货源如何确定？		
网店创建与数据管理	网店运营产生的一系列数据如何管理？		
营销策略	使用什么样的营销/促销策略？	根据促销规划，采用开业促销优惠券策略，选用平均分析法分析网店运营前期销售数据，确定优惠券金额	
销售业绩	卖得怎么样？销售业绩如何？		
网店数据安全	网店数据安全如何维护？		

需要注意的是，即便是问题整体解决方案确定之后，也仍有可能需要继续完善方案。比如，在方案执行过程中，由于业务需求变更，或者其他因素发生变化，都可能会需要对方案进行修正与完善。

## ● 制订问题整体解决方案的基本流程与主要思维活动

上述从需求分析直至优化方案的过程其实就是一个完整的问题整体解决方案确立的过程，蕴含了系统设计与结构化思想，体现了抽象、分解、算法、迭代和优化等主要思维过程，如图1.2.4所示。



算法是一个有穷规则的集合，规定了解决某一特定类型问题的一个运算序列。通俗地说，算法就是解决问题的方法和步骤。

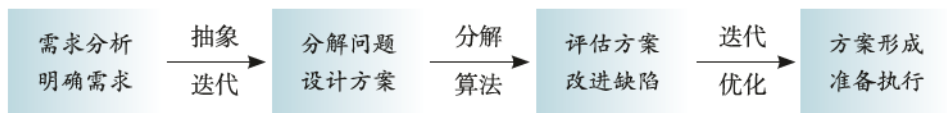


图 1.2.4 制订问题整体解决方案的基本流程与主要思维活动



## 拓展练习

学校团委打算近期举行一次学生创客大赛，团委王老师计划先制订一个创客大赛成绩排名整体方案。请你向王老师提供制订问题整体解决方案的思路，并协助王老师制订出创客大赛成绩排名整体方案。

## 单元学习评价

通过本单元的学习，我们感受了数据管理与分析的作用，体验了制订问题解决方案的基本过程。你是如何看待数据管理与分析的？你是否具备制订或评价问题解决方案的能力？请参与小组交流并反思，开展自评或小组评价。

1. 如果计划外出游玩，你通常会使用以下哪些移动APP？

- 在线订票APP      旅行住宿APP      自助点餐APP      智能计步APP  
智慧学习APP      智慧公交APP      在线支付APP      其他\_\_\_\_\_

除了快捷、方便之外，这些APP还为你带来了哪些便利？数据管理与分析在其中有何贡献？请将答案写在下面的横线上。

---



---



---

2. 请上网搜索一个业务数据问题整体解决方案成品，对方案进行评价，重点评价方案的完整性、合理性与可行性，指出方案存在的缺陷并加以改进。

### 业务数据问题整体解决方案评价

【问题整体解决方案简介】

方案名称：\_\_\_\_\_

方案要解决的主要问题：（简列2~3个）

---



---

【评价要点】（含优点与缺陷）

方案的完整性：\_\_\_\_\_

方案的合理性：

● 是否是从需求分析出发的？

---



---

● 方案是否有合理性依据？能否达成预期目标？

---



---

方案的可行性:

- 时间、成本、风险与客户接受度如何?

---

- 使用数据管理与分析技术能否实现方案?

---

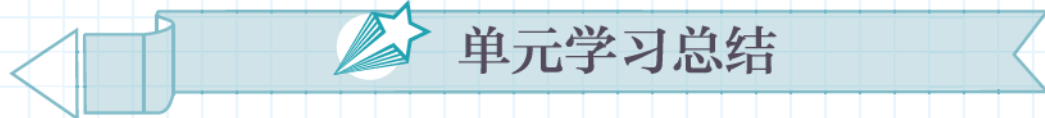
【方案的缺陷及改进】(至少1~2点)

缺陷1: \_\_\_\_\_

改进措施: \_\_\_\_\_

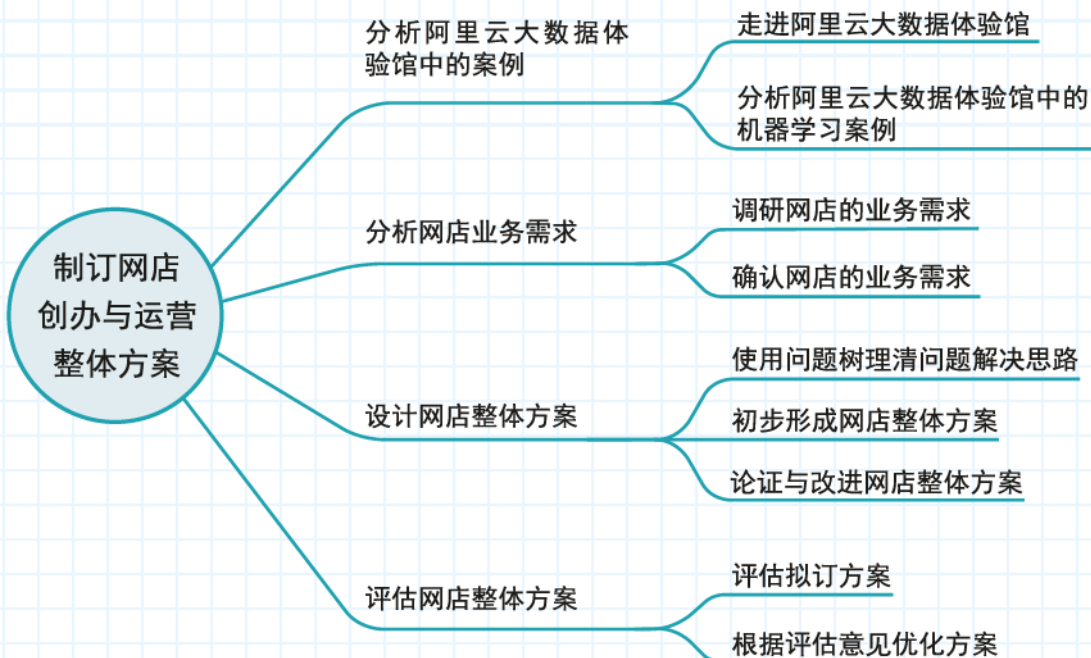
缺陷2: \_\_\_\_\_

改进措施: \_\_\_\_\_



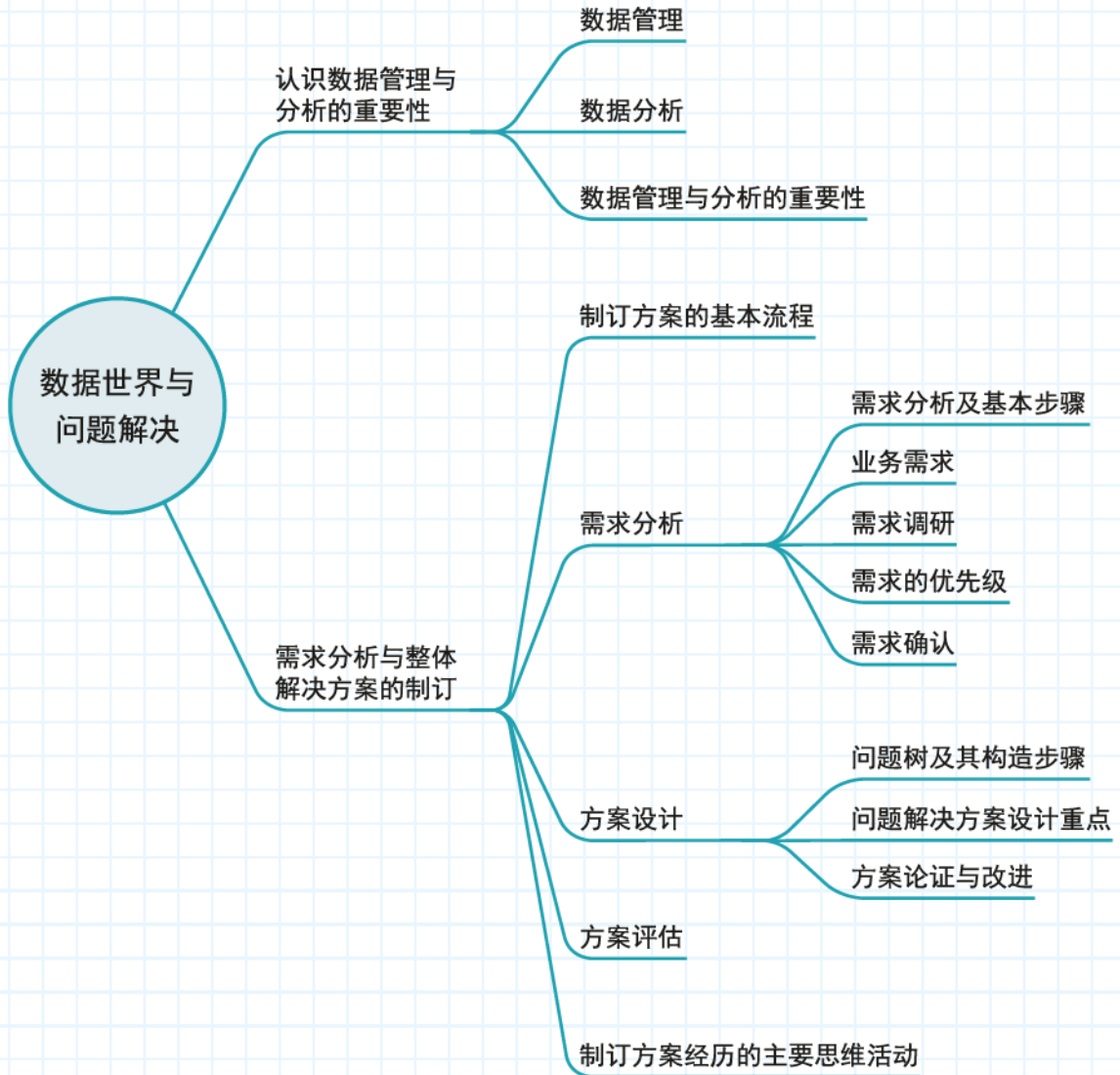
本单元围绕“制订网店创办与运营整体方案”项目，通过依次完成分析阿里云大数据体验馆中的案例、分析网店业务需求、设计网店整体方案和评估网店整体方案等任务，开展相应的主体活动，认识数据管理与分析的重要性，体验问题整体解决方案制订的过程，初步形成分析、设计、评价与优化方案的基本思想与方法。

### “项目—任务—活动”结构图





## 知识地图



## 第 2 单元 数据的收集与整理

数据的收集与整理是数据管理与分析的第一步，如果不对数据进行收集与整理，数据管理与分析就无从谈起。随着互联网的普及，全球数据呈指数级增长。截至2018年12月，中国网民规模达8.29亿，全年新增网民5653万人，互联网普及率为59.6%……。但是，这些数据与我们的需求也许并不相符，只有根据问题解决需要主动收集并整理的数据才是有价值的。

本单元，我们将通过“确定网店经营品类与货源”项目展开学习，经历数据的收集与整理过程，认识数据来源的多样性，根据解决问题的需要，选择适当的方法与工具，对数据进行必要的收集；探寻其中的噪声数据现象并对数据进行分类与存储；体会数据的收集与整理在解决业务数据问题中的重要作用。

确定网店的经营品类与货源，需要收集行业销售数据、批发数据等。到哪里收集这些数据？如何收集？对于收集来的数据，如何进行整理？要解决这些问题，我们需要完成以下三个任务：

- ◆ 收集文化用品行业销售数据；
- ◆ 确定货源；
- ◆ 整理商品信息。

## 2.1 数据的收集

在开展数据管理与分析之前，要进行数据的收集。数据可能来源于网络，可能来源于问卷调查，还可能来源于传感器。我们需要根据数据问题解决需求，选择可靠的数据来源，采用适当的方法与工具收集有价值的数



### 学习目标

- ★ 认识数据来源的多样性。
- ★ 能够选择适当的方法与工具收集数据。
- ★ 体会信息技术在数据收集过程中的重要作用。

根据网店整体方案，佳玮需要带领采购部成员一起收集行业销售数据和目标消费者需求数据来确定网店商品的经营品类。这些数据可能来源于哪里？可以用什么方法或工具收集这些数据？让我们和佳玮一起去完成数据的收集工作。



### 任务 收集文化用品行业销售数据

#### ※ 活动1 收集品类销售数据

佳玮决定先收集行业销售数据。让我们和她一起登录指数网站，观察淘宝文化用品行业的品类销售数据，以便确定哪些品类最受欢迎。

佳玮打开了各品类的销售数据页面，网站截图如图2.1.1所示。请确定各品类的销售情况排名。

2017年1月至4月，淘宝文化用品行业销量最高的前三个品类是：

\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。



行业整体销售数据可能来源于淘宝指数、阿里指数等指数网站发布的数据。

行业	文化用品			
销售趋势	类目趋势	类目分布	品牌分布	
类目	2017-01	2017-02	2017-03	2017-04
1 笔类/书写工具	22.81%	26.04%	25.78%	26.08%
2 绘画测量用品	1.19%	1.39%	1.37%	1.32%
3 教学演示/展示告示用品	2.68%	2.58%	3.16%	3.53%
4 点读笔	0.03%	0.04%	0.04%	0.03%
5 日常学习用品	12.68%	13.77%	12.61%	11.70%
6 点读机	0.04%	0.03%	0.03%	0.02%
7 纸张本册	21.83%	22.16%	20.85%	19.80%
8 印刷制品	6.18%	3.74%	3.99%	4.27%
9 计算器	0.75%	0.69%	0.85%	0.88%
10 学习机	0.08%	0.05%	0.04%	0.04%

图 2.1.1 某指数网站淘宝文化用品行业的品类销售数据页面截图

佳玮又对同学们进行了小范围调研，调研结果与指数网站数据收集结果比较一致。于是，她决定经营笔、纸张本册、日常学习用品这三种品类。

行业销售数据来源于指数网站，消费者需求数据来源于佳玮对消费者进行的小范围调研，这两种数据来源有什么区别？

## ● 数据的来源

指数网站中的数据并不是直接获得的，我们只是对数据加以利用，这样的数据获取渠道称为数据的间接来源。例如，带有统计数据的年鉴、数据交易平台等就属于数据的间接来源。

除了指数网站外，销售数据也可能来源于我们对消费者的调查。使用者可以通过调查、实验等方法，直接获得数据，这样的数据来源称为数据的直接来源。



在网络上，有许多公开的数据来源，如国家统计局网站、发布金融数据的银行网站等，这些都是非常可靠的数据来源。

通过数据交易平台，用户可以购买所需的数据。当用户发布数据收集需求后，开发者利用平台的收集引擎收集相应数据，完成用户的特定需求。

### ※ 活动2 收集单品销售数据

现在，佳玮需要进一步了解各品类下的单品销售情况，以确定经营哪些单品。请收集指数网站中的单品销售数据，讨论并思考这些数据是否全面、准确、科学，完成表2.1.1。

佳玮打开了文化用品行业中的书包销售数据页面，网站截图如图2.1.2所示。

对于书包来说，图案的样式、颜色的种类、规格的大小等，都会影响其销售情况。指数网站以这些不同的方面为依据，分别收集了销售数据。从图中可以看出，该指数网站将书包图案的样式分为“其



图 2.1.2 某指数网站书包销售数据页面截图

他”和“米宝”两类。现实中书包的图案有许多类别，这样的数据分类方式并不科学，不能帮我们确定具体要经营哪种图案的书包。

请打开书包的其他销售数据页面，观察并完成表2.1.1。

表2.1.1 书包销售数据分析表

相关销售数据	能否满足问题 解决需求	原因
不同图案的销售数据	不能	不同图案的书包销售数据只分成了“其他”和“米宝”两类，这种数据分类方式并不科学
不同价格区间的销售数据	能	价格数据区间设置合理，50~100元价格区间的书包销量最高
不同品牌的销售数据		

我们通过对指数网站中的数据进行观察发现，指数网站中的数据有的并不科学，或者与我们的需求不符。在使用直接来源获取数据时也会出现这样的情况吗？不同的数据来源有什么特点呢？

## ● 不同数据来源的特点



知识产权是指人们就其智力劳动成果所依法享有的专有权利，包括专利权、著作权、商标权等。

指数网站是数据的间接来源，通过间接来源所获取的数据是二手数据。二手数据获取较为容易，但使用时需要对其可靠性等进行判断。我们可以从发布单位是否权威、收集目的是否明确、收集方法是否科学、所收集的数据时效性如何、是否满足问题解决需求等方面进行思考。在直接使用二手数据时，要尊重知识产权。首先，判断数据是否被允许使用；其次，如果可以使用，在使用时要注明数据来源。

通过直接来源所获取的数据称为一手数据。通常来说，一手数据更全面，更符合我们的特定需求。过去，由于受到资金、技术条件等限制，一手数据的获取较为困难。今天，随着信息技术水平的不断进步和互联网的普及，获取一手数据变得越来越容易。在二手数据不能满足我们的需求时，可以充分利用信息技术手段，直接收集我们所需要的数据。

### ※ 活动3 收集地区店铺销售数据

佳玮要经营的是小件商品，价格比较便宜，消费者在购买时为了节省邮费，往往更倾向于选择附近地区网店中的商品。因此，佳玮想要看看她生活的东北地区店铺的销售数据。

在指数网站中并没有东北地区的销售数据，人工录入需要消耗大量的时间和精力。有没有自动收集这些数据的工具呢？请上网查找常用的数据收集方法与工具，比较其优缺点，从中选择合适的方法与工具来帮助佳玮收集某一单品的地区销售数据，存储为如图2.1.3所示的电子表格。

商品名称	发货地	价格	品牌
白字彩色直液式走珠笔针管签字商务中性笔	辽宁沈阳	1.45 - 1.75	品牌: 白字
星晨中性笔0.5mm子弹头黑色水笔办公用品学生写字笔批发包邮	辽宁沈阳	12.00	品牌: 星晨
风意摩易擦热可擦笔磨擦一擦中性笔子弹头晶蓝黑色一盒包邮	辽宁沈阳	9.99 - 62.00	品牌: 风意
铜笔工匠笔纯铜笔黄铜笔复古签字笔全金属六棱形手工怀旧中性笔	辽宁沈阳	25.00 - 35.00	品牌: 铜笔工匠
唯好可擦中性笔 摩易擦 摩擦笔 水笔 进口可擦油墨 12支包邮	辽宁营口	1.18 - 28.80	品牌: 唯好
特价纯风超大容量钻石头中性笔钻石笔水性笔 黑/蓝/红	辽宁沈阳	0.39 - 0.75	品牌: 纯风

图 2.1.3 东北地区网店销售数据截图

### ● 常用的数据收集方法与工具

根据不同的数据问题需求，我们需要选择适当的数据收集方法与工具来收集数据。常用的数据收集方法有调查、观察、测量、实验、查阅文献、基于互联网进行收集或搜索等。除此之外，随着技术的进步，利用传感器（如图2.1.4所示）、网页采集器等自动收集数据，也是目前常用的数据收集方法。

网页采集器是一种非常便捷的数据收集工具，它可以自动采集网络上分散的内容。利用网页采集器不仅可以收集文字、数字，也可以收集图片、音乐等。当利用网页采集器仍不能满足特定的数据收集需求时，可自行编写网络爬虫来收集网页中的数据。



图 2.1.4 利用手机中的传感器收集健康数据

💡 网络爬虫是遵循一定的规则，自动地抓取网络信息的程序或脚本。



## 拓展练习

1. 判断以下数据是否可靠并说明理由。
  - (1) 国家统计局数据显示, 2015年前三季度GDP同比增长6.9%。
  - (2) 某网站统计结果显示, 51%的人认为上大学没有用。
  - (3) 从北京到上海最早的火车是早上6:00发车(摘自某网页, 数据更新于2004年)。
2. 如果要收集以下数据, 你会采用什么样的方法或工具? 为什么?
  - (1) 2015年度最受欢迎的电影。
  - (2) 消费者对于网店风格、价格期待等的消费心理。
  - (3) 不同浓度盐水的沸点。
  - (4) 人民网在2016年两会期间的所有新闻。

## 2.2 认识噪声数据

一般情况下，收集到的数据中经常会存在不需要、不准确、重复等不能用的数据，这些数据会影响分析与决策，我们把这些不能用的数据称为噪声数据。为什么会出现噪声数据呢？这些噪声数据会产生哪些影响呢？



### 学习目标

- ★ 认识噪声数据现象。
- ★ 了解噪声数据的成因。
- ★ 体会噪声数据的影响。

为了确定货源，佳伟已经从某批发网站收集了部分单品的批发数据，包括商品名称、进货价格、批数等。现在需要弄清楚这些数据是否准确、有效，让我们和佳伟一起来探查其中是否存在噪声数据吧。



### 任务 确定货源

#### ※ 活动1 探查单品批发数据中的噪声数据现象

佳伟首先搜索的单品是某品牌7901型号的皮面本，图2.2.1所示是她按价格从低到高排序后的数据节选。请大家仔细观察，将噪声数据在图中直接标记出来。

编号	商品名称	价格	批数	厂家
1	包邮办公A4软抄本 B5笔记本 日记本记事本A5软面抄笔记本文具	6.7	≥1本	凯乐文具源头厂家
2	7901 皮面本 日记本 仿皮记事本 笔记本 80页 25K		1~39本	诗易文具厂
3	笔记本 7901高级皮面本 80页25K仿皮商务记事本 +现货		7 90本	新潮文化用品有限公司
4	正品7901仿皮面笔记本A5商务皮面本80页25K仿皮记事本记录本	8.5	1~15本	必升贸易有限公司
5	7901皮面记事本25K 80页 笔记本 仿皮记录本	8.6	1~9套	爱必思文具有限公司
6	包邮7901商务皮面本笔记本 记事本 仿皮日记本 25K 80页 黑色	9.8	1~15本	晨阳文具厂
7	包邮7901商务皮面本笔记本 记事本 仿皮日记本 25K 80页 黑色	9.8	1~15本	晨阳文具厂
8	7901皮面本日记本 80页25K仿皮商务记事本笔记本 正品	200	1~9本	中凯文具厂

图 2.2.1 单品批发数据节选



## ● 噪声数据现象

噪声数据现象主要有两种。一种是数据中存在错误，主要包括数据的值不正确、数据缺失、数据重复和数据不一致等。其中数据的值不正确是指数据存在错误或异常值，如正常人一天走路的步数为5，这一数值显然是不正确的。数据缺失主要是指数据的值空缺。数据重复是指数据的值全部相同。数据不一致则是指数据的编码不一致，或者格式及单位不统一等。另一种是数据不符合我们的需求，如要了解女性对于某一产品的需求情况，就需要将男性的购买数据刨除在外。

佳玮收集的单品批发数据中存在的噪声数据现象属于哪种呢？请分析其中的噪声数据现象并对其进行描述，将分析结果填入表2.2.1。

表 2.2.1 单品批发数据中的噪声数据现象

噪声数据现象	对噪声数据现象的描述
数据重复	第6条和第7条的数据完全重复

### ※ 活动2 分析噪声数据的成因

为什么佳玮收集到的数据中存在噪声数据呢？请展开讨论，从人为原因、设备原因、程序错误等方面逐一进行排查，推测使用不同方法或工具收集数据时可能出现的噪声数据现象及成因，根据讨论的结果完成表2.2.2。

表 2.2.2 噪声数据的成因分析表

收集方法或工具	可能产生的噪声数据现象	噪声数据的成因
调查	数据缺失	填答人员没有填写或录入时产生遗漏

续表

收集方法或工具	可能产生的噪声数据现象	噪声数据的成因
网页采集器	数据的值不正确	网站搜索程序异常、网站代码不一致等

### ● 噪声数据的成因

噪声数据的成因有许多。人为原因有填答失误、录入错误、重复录入、操作失误等；设备原因有收集设备故障、存储设备故障等；还有程序错误、传输过程中的数据丢失、多个数据合并时的编码不一致等。

此外，由于需求的变化，一些有用的数据也可能变成噪声数据。例如，对于日常运营来说，价格智能系统收集的节假日促销数据就是噪声数据，而对于开展促销活动来说，价格智能系统收集的日常数据就变成了噪声数据。



价格智能系统可以自动收集不同电子商城的商品售价并存储，以帮助商家根据不同的需求进行数据分析。

### ※ 活动3 分析噪声数据对货源决策的影响

佳玮收集批发数据的目的是确定进货渠道，批发数据中的噪声数据是否会影响佳玮的决策？如果其他数据中出现了噪声数据又会产生哪些影响？请围绕这些问题展开思考或讨论，完成表2.2.3。

表 2.2.3 噪声数据对货源决策的影响

讨论的问题	结论
1号商品并不是7901型号的皮面本，是否会影响佳玮的决策？	会影响，因为1号商品进价最低（6.7元），会导致采购部采购错误的商品
6、7号商品数据完全重复，是否会影响佳玮的决策？	
厂家的地址出现错误有哪些影响？	影响佳玮进货时的邮费和货品到达的时间
厂家将商品颜色等相关信息填写错误有哪些影响？	
恶意厂家利用非法手段刷取成交量、信誉等行为有哪些影响？	

## ● 噪声数据的影响

数据管理的目的在于充分发挥数据的作用，如果所管理的数据中存在噪声数据，会影响数据管理相关事务的正常运行。例如，信件的收件地址发生错误，可能会导致信件无法投递或投递至错误的地址；银行管理系统中的数据发生错误，可能会导致客户的存款信息发生变化。

数据分析能够帮助人们从杂乱无章的数据中提取有用的信息，进而为人们的决策提供依据。如果所要分析的数据中存在噪声数据，会影响人们的判断，使人们做出无效乃至错误的决策。

在进行数据管理与分析之前，需要对噪声数据进行处理，去除其不良影响。



### 拓展练习

某培训机构为了更好地了解学员的需求情况设计了调查问卷，同时在网络调查平台上和街头展开了调查。调查回收的问卷及收集的数据样例节选如图2.2.2至图2.2.4所示。

1. 您的性别是 ( ) 问卷编号 12  
 A.男            B.女
2. 您的年龄是 (A)  
 A. 10 岁以下    B. 10~15 岁    C. 15~20 岁    D. 20~25 岁  
 E. 25~30 岁    F. 30~40 岁    G. 40~50 岁    H. 50 岁以上
3. 您现在是否在上学，学历是什么? (D)  
 A. 小学          B. 初中          C. 高中          D. 大学  
 E. 工作          F. 退休

图 2.2.2 街头问卷填答截图

问卷编号	1. 您的性别是	2. 您的年龄是	3. 您现在是否在上学，学历是什么?
12		B	D
13	A	B	B
13	A	B	B

图 2.2.3 人工统计的数据节选

序号	1. 您的性别是	2. 您的年龄是	3. 您现在是否在上学，学历是什么?
20	3	4	

图 2.2.4 网络调查平台自动统计的数据节选

现在该机构需要将数据合并到一起进行统计。请找到噪声数据并分析其成因，完成表2.2.4。

表 2.2.4 调查数据中的噪声数据现象分析

编号	噪声数据现象	噪声数据的成因
图2.2.3	不完整数据	编号为12的问卷调查对象没有填答第1题
图2.2.4		
数据 合并后		

## 2.3 数据的结构化

为了便于管理与分析数据，需要将数据存储到数据库中。除了文本、数字之外，数据库也需要对声音等进行管理。那么这些数据有什么特点？如何存储这些数据？



### 学习目标

- ★ 理解不同结构化程度数据及其区别。
- ★ 体会不同结构化程度数据的存储过程。

确定好经营品类和货源之后，佳玮要整理商品信息准备开店。在网店中，与商品有关的信息包含哪些？信息背后的数据又该如何存储？



### 任务 整理商品信息

#### ※ 活动1 考察网店商品信息

为了了解网店中的商品信息都包括哪些，佳玮决定亲自对网店进行一番考察。请同学们打开网店中任一单品页面进行考察，同时思考以下问题：网店商品信息包括哪些内容？这些信息都有哪些表现形式（如文本、数字、声音等）？不同的单品所包含的商品信息是否相同？考察结束之后请完成表2.3.1。

表2.3.1 商品信息考察表

所考察的单品	所包含的信息	信息的表现形式
7901型号皮面本	商品名称	文本
	售价	数字

续表

所考察的单品	所包含的信息	信息的表现形式

在考察商品信息的过程中可以发现，商品信息包括商品名称、价格、货号、商品广告等，其表现形式有文本、数字、视频等。规整的文本、数字等可以直接存储进数据库中，而图片、视频等数据该如何存储呢？

### ● 结构化数据

结构化数据即行数据，能够用统一的结构加以表示，如人们的出生日期、性别等。常见的二维表中的文本就属于结构化数据。

### ● 非结构化数据

非结构化数据是指数据的结构不规则或不完整，不方便用二维表来表现的数据，如Word文档、PDF文件、PowerPoint课件等。

### ● 半结构化数据

半结构化数据是介于结构化数据和非结构化数据之间的一种数据。它的结构变化很大，如员工的简历。有的员工的简历很简单，可能只包括教育情况；有的却很复杂，可能会包括工作情况、政治面貌等。

请同学们思考常见的数据中哪些属于结构化数据、哪些属于半结构化数据、哪些属于非结构化数据。将思考结果与大家讨论，并阐述思考过程。

#### ※ 活动2 存储商品数据至电子表格

为了管理网店中的商品，佳玮要用电子表格存储商品数据。这些数据应如何存储？请同学们参考表2.3.2，任选两种不同的单品分别制作商品数据表。表2.3.2中仅列了部分内容，在制作时，可以根据实际情况进行补充。

表 2.3.2 商品数据表 样例

商品名称	商品照片	商品详情
A5日记本	C:\Jiawei\pic.jpg	C:\Jiawei\Desc001.doc

### ● 结构化数据的存储

在存储结构化数据时，我们需要先建立相应的电子表格，再将文字、数值等直接进行存储。

### ● 非结构化数据的存储

我们通常将非结构化数据以文件的方式存储在文件系统中，同时将文件路径存储在数据表中。

### ● 半结构化数据的存储

半结构化数据在存储时需要化解为结构化数据。以简历为例，通常是对现有简历中的信息进行分类整理，总结出所有的类别，对每一类信息单独建立数据表进行存储。比如可以将员工简历中的基本信息、教育信息、工作信息、政治面貌等单独建立数据表，分别进行存储。

在制作数据表时，对于非结构化数据，请建立文件夹统一保存文件，并在电子表格中存储其路径。对于商品详情等图文混排信息，请粘贴至Word文档中，在电子表格中存储文档的路径。



## 拓展练习

如图2.3.1所示是某专业学生的档案，请你帮助老师建立相应的电子表格，并对结构化数据、非结构化数据和半结构化数据进行注明。

市场营销1班学生档案	
照片	姓名: _____ 性别: _____
	出生日期: _____
	籍贯: _____
	QQ号码: _____ 电子邮箱: _____
	手机: _____
	家庭电话: _____
	家庭住址: _____
个人爱好: _____	
特长: _____	
曾获奖项: _____	
个性化介绍: _____	

图 2.3.1 学生档案



## 单元学习评价

通过本单元的学习，我们认识了数据来源的多样性，了解了噪声数据的成因及影响，体会了不同结构化程度数据的存储过程。现在，你能根据不同需求完成数据的收集与整理工作吗？请结合本单元的学习内容，完成以下评价。

1. 某气象局利用温度传感器收集了全天24小时的温度数据并将其发布至气象局网站。浩然想利用这些数据开展一个小型的科学研究项目，于是他利用网络爬虫对气象局网站中的数据进行了收集。请问这两种数据中，哪些是一手数据？哪些是二手数据？这两种数据各有什么优缺点？

(1) 气象局利用温度传感器收集的数据属于\_\_\_\_\_，浩然利用网络爬虫收集的数据属于\_\_\_\_\_。

(2) 一手数据的优点：\_\_\_\_\_；

一手数据的缺点：\_\_\_\_\_。

(3) 二手数据的优点：\_\_\_\_\_；

二手数据的缺点：\_\_\_\_\_。

2. 嘉琦通过学习得知，计步器能够记录步数是由于里面内置了相应的传感器。他查看妈妈的步数数据，发现某些天的步数过少甚至为0。请判断这属于以下哪种噪声数据现象，并分析其成因。

数据的值不正确     数据缺失     数据重复     数据不一致

噪声数据的成因可能是：\_\_\_\_\_。

3. 为了方便对非物质文化遗产进行管理，需要建立相应的数据表。请以“民间文学”这一类别为例，收集相应数据并存储至电子表格中。在完成的过程中，请思考以下问题：

(1) 需要收集哪些数据？

(2) 从哪里收集数据？数据来源是否可靠？

(3) 如何收集这些数据？用什么方法或工具进行收集比较好？

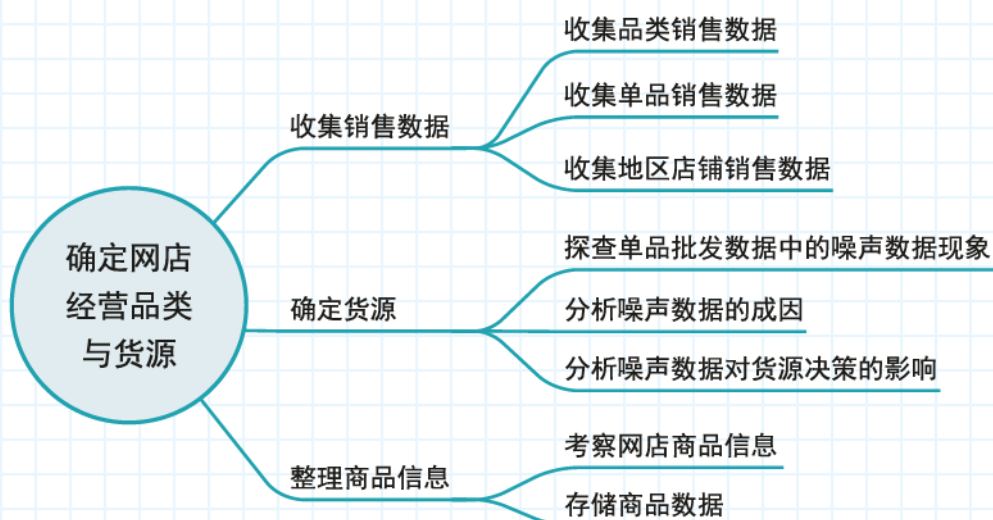
(4) 收集到的数据中是否存在噪声数据？这些噪声数据会有哪些影响？

(5) 数据中是否存在非结构化或半结构化数据？如何将其进行存储？

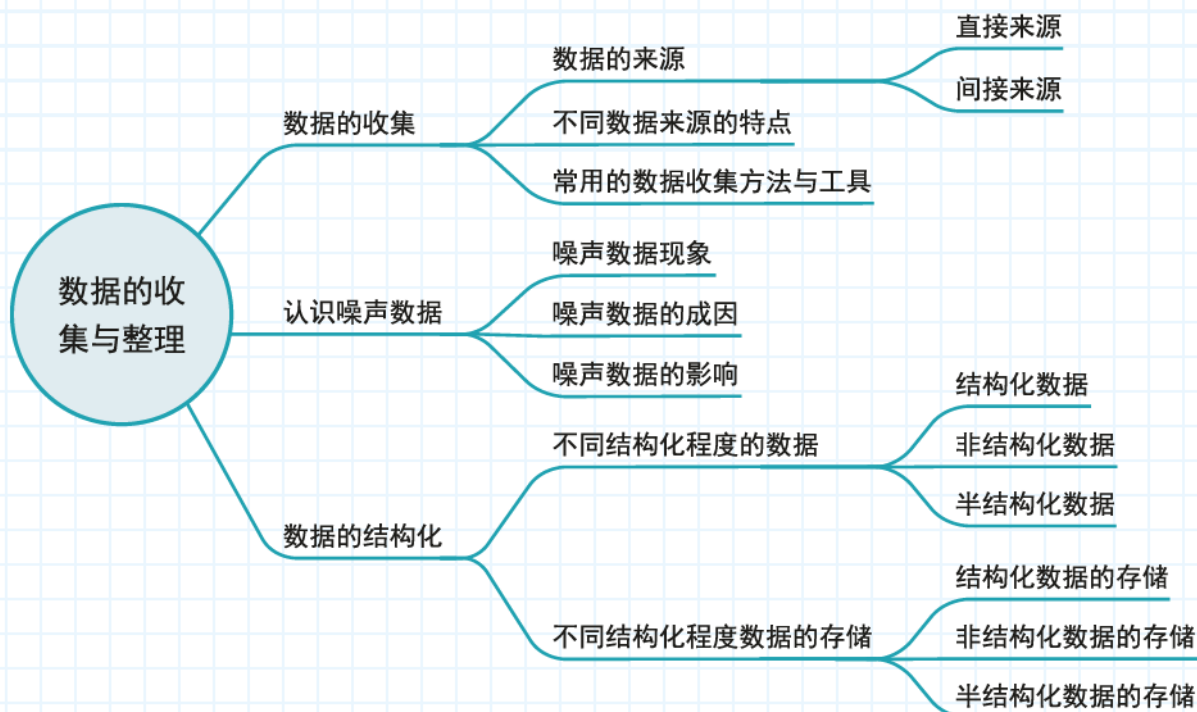
## 单元学习总结

本单元围绕“确定网店经营品类与货源”项目展开学习。通过依次完成收集文化用品行业销售数据、确定货源和整理商品信息等任务，认识数据来源的多样性；根据解决问题的需要，选择适当的方法与工具对数据进行收集；认识噪声数据的现象、成因及影响；理解不同结构化数据的区别及存储方式；深入体会数据的收集与整理在解决业务数据问题中的重要作用。

### “项目—任务—活动”结构图



### 知识地图



## 第 3 单元 数据的管理

在对数据进行收集与整理的基础上，我们需要利用计算机硬件和软件技术对数据进行有效的存储、处理和应用，以充分发挥数据的作用，而管理数据最有效的工具之一是数据库。使用数据库工具对数据进行科学管理，是对数据进行科学有效分析、充分实现数据价值的重要方法。

在本单元中，我们将通过“管理网店数据”项目展开学习，通过设计、操作与应用数据库，体验数据管理的全过程。初步掌握数据库设计与操作的基本方法，形成数据管理的基本思想。数据是记录客观事物属性的符号，信息是对数据进行加工处理后得到的结果。数据是信息的载体，信息是对数据的解释。网店运营过程中会产生大量的商品信息、交易信息及评价信息。这些信息用哪些数据表示？这些数据存储在哪里？如何实现对数据的处理与应用？要解决这些问题，我们需要完成以下七个任务：

- ◆ 创建网店并制订运营管理规划；
- ◆ 分析网店运营数据；
- ◆ 剖析网店后台数据的结构；
- ◆ 设计离线网店数据库；
- ◆ 创建离线网店数据库；
- ◆ 操作离线网店数据库；
- ◆ 处理与网店相关的数据库事务。

## 3.1 数据库的初识与剖析

在信息社会，充分有效地管理和利用各类信息资源，是进行科学研究和决策管理的前提。数据库是隐藏于管理信息系统、办公自动化系统、决策支持系统等各类信息系统背后的核心部分。那么，数据库是由什么构成的？数据库中的数据又是如何组织的？



### 学习目标

- ★ 了解概念模型与数据模型。
- ★ 掌握关系数据模型的表征方法。
- ★ 初步体会分类和抽象的思想。

经过前期的准备，佳玮决定在网络商城创建一个文具店，命名为“时光文具优选店”。



### 任务一 创建网店并制订运营管理规划

#### ※ 活动1 创建网店

本活动中的创建网店是指在“网络商城”里注册账号并申请开店。具体创建过程如下：

第一步，开通一张银行卡并存储一定数额的保证金；

第二步，通过对各大网络商城的查询、评价、比较，最终选定商城并注册；

第三步，登录网络商城，申请开店；

第四步，商城客服对店主提交的身份信息及银行账户信息进行严格的审核；

第五步，审核通过表示创建网店成功，否则根据提示重新提交材料，继续进行审核直至通过。

请同学们体验创建网店的过程，然后反思：网店创建过程中遇到了哪些问题？是如何解决的？对网店创建有什么认识与体会？将反思结果进行总结与梳理。

### 网店创建过程及结果反思

#### 一、网店创建过程中遇到的问题及解决方法

#### 二、网店创建的认识与体会

1.

2.

3.

4.

### ※ 活动2 制订网店运营管理规划

现在时光文具优选店仅仅是一个空店铺，我们需要仔细规划网店如何运营的问题。网店初期的运营管理主要涉及以下问题：网店如何“装修”？网店的特色如何打造？网店的商品如何发布？网店的服務如何定位？网店运营管理规划的制订需围绕这些问题进行。请查阅各网络商城培训中心提供的资源，在此基础上认真思考，分小组讨论，积极构想，帮助佳玮制订一份详细的网店运营管理规划书。

### 网店运营管理规划书

#### 一、网店装修风格

#### 二、网店特色打造

#### 三、网店商品发布

#### 四、网店服务定位

通过创建网店并制订运营管理规划任务，我们可以掌握网店创建的流程及网店运营管理的内容。那么，是否可以尝试在网络商城中为父母或亲朋好友的实体店创建一个网店呢？



## 任务二 分析网店运营数据

### ※ 活动1 网店运营数据分组归类

假设时光文具优选店已经试运营，且在网店的日常运营过程中，每天都会产生大量信息，如商品信息、交易信息、支付信息、物流信息、评价信息等，而且这些信息中包含了商品名称、价格、品牌、型号、规格、颜色、货号、运费、库存量等大量数据。佳玮很想知道的是，网店运营的数据是如何被有规律、有秩序地管理的？

请查看网店运营信息及数据，认真思考以下问题并与同学们一起讨论：网店信息和数据之间是否有关联？网店数据能否按信息归属分类？

分类和抽象是信息技术学科重要的思想方法。把客观存在的事物以数据的形式存储到计算机中，其间经历了现实世界→概念世界→数据世界的逐级抽象过程。

### ● 概念模型

从现实世界到概念世界的抽象结果称为概念模型，概念模型的要素主要包括实体、属性和联系。

实体指客观存在并可相互区分的事物，分为客观实体和信息实体。如商品是客观实体，订单信息、评价信息是信息实体。

属性指事物所具有的特征，如商品的属性包括编号、名称、品牌、货号、规格等。评价信息的属性包括评价编号、评价人、评价内容等。实体的概念模型表征如图3.1.1所示。

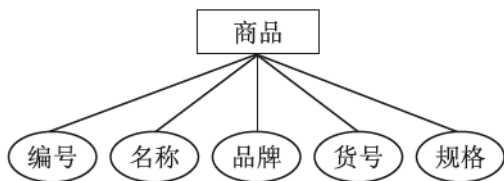


图 3.1.1 实体的概念模型表征图

联系指事物之间的关联关系，如商品和消费者之间的消费关系。



#### 图示说明

矩形框：表示实体，在框中记入实体名。

椭圆形框：表示实体的属性，在框中记入属性名。

直线：表示属性与实体之间的所属关系。

请依据概念模型的有关知识，将网店的库存数据、订单数据等以所属实体为依据进行分组归类，尝试建立网店各实体的概念模型表征图。

### ※ 活动2 分析各组数据之间的关联关系

我们将网店的全部数据以所属实体为依据进行分组归类后，尝试分析各组数据之间的关联关系，发现存在三种联系类型：一对一联系、一对多联系和多对多联系。

#### ● 联系类型

实体之间存在三种联系类型：一对一联系、一对多联系和多对多联系。

一对一联系表征：要在两个实体连线方向各写1。

一对多联系表征：要在一的一方写1，多的一方写N。

多对多联系表征：要在两个实体连线方向分别写N和M。



#### 图示说明

菱形框：表示联系，在框中记入联系名。

直线：实体与联系之间用直线相连，并在直线上标注联系的类型。

#### ● E-R图

E-R图也称实体—联系图，它提供了表示实体、属性和联系的方法，用来描述现实世界的概念模型。

制作E-R图的步骤如下：

第一步，确定所有的实体；

第二步，确定各个实体应包含的属性；

第三步，确定实体集之间的联系。

商品和供应商的E-R图如图3.1.2所示。

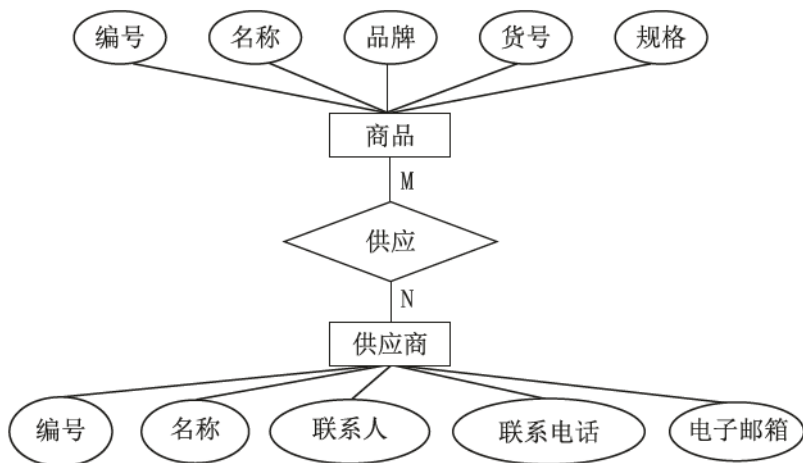


图 3.1.2 商品和供应商的E-R图

依据E-R图的有关知识，我们可以尝试建立起网店全部实体间的联系，形成网店管理信息系统的E-R图。



### 任务三 剖析网店后台数据的结构

#### ※ 活动1 观察网店操作界面中的数据

在时光文具优选店运营管理过程中，佳玮发现网店订单中的数据和客户交易数据非常有条理，而且当她把网店的历史订单和客户交易数据导出后，这些数据可以很方便地用Excel进行查询和使用。佳玮猜想后台数据库中的数据和Excel表格中的数据一样，是结构化的数据。于是她便产生了疑问：既然用Excel管理数据很方便，为何还要使用数据库？数据库与Excel有何不同？请思考佳玮的猜想和疑问，通过查找资料、分组讨论、头脑风暴，做出判断分析，进行解答与求证，梳理并总结数据库与Excel的异同。



**数据库：**是以一定方式储存在一起、能为多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。它具有整体性、共享性、独立性和安全性等特点。

#### 数据库与Excel的比较

##### 一、相同点

##### 二、不同点

##### 三、结论



## ● 为何使用数据库

毫无疑问，使用数据库的目的是方便、快速、精准、完备、安全地存储和管理大量的业务数据。借助数据库管理系统，可以对海量数据进行方便快速的检索、精准完备的存储、高效安全的管理。

### ※ 活动2 解析网店后台数据模型

随着对数据库的深入了解，佳玮更想知道网店后台应用了什么样的数据结构来提高网店的运营效率。请帮助佳玮解析后台数据的结构，创建网店后台数据模型。

## ● 数据模型

从概念世界到数据世界的抽象结果称为数据模型，数据模型的要素主要包括数据结构、数据操作和数据约束。

数据结构：相互之间存在一定联系的数据元素的集合，是对数据世界静态特征的描述。层次结构、网状结构和关系结构是三种典型的数据结构，三种结构的比较如表3.1.1所示。

表 3.1.1 三种典型数据结构的比较

类别	层次结构	网状结构	关系结构
定义	用树状结构表示实体内部及实体间的联系	用网状结构表示实体内部及实体间的联系	用二维表表示实体内部及实体间的联系
特点	<ul style="list-style-type: none"> <li>◆ 只有一个最高节点，即根节点</li> <li>◆ 其余节点有且只有一个父节点</li> <li>◆ 上下层节点之间表示一对多的联系</li> </ul>	<ul style="list-style-type: none"> <li>◆ 允许节点有多于一个的父节点</li> <li>◆ 可以有一个以上的节点没有父节点</li> <li>◆ 允许两个节点之间有多种联系</li> </ul>	<ul style="list-style-type: none"> <li>◆ 一个二维表代表了一个关系</li> <li>◆ 结构简单，容易实现</li> </ul>
优点	<ul style="list-style-type: none"> <li>◆ 结构简单，层次分明</li> <li>◆ 查询效率高</li> <li>◆ 提供良好的数据完整性支持</li> </ul>	<ul style="list-style-type: none"> <li>◆ 可表示实体间的多种复杂联系</li> <li>◆ 具有良好的性能和存储效率</li> </ul>	<ul style="list-style-type: none"> <li>◆ 具有更高的数据独立性</li> <li>◆ 具有更好的安全保密性</li> </ul>
缺点	<ul style="list-style-type: none"> <li>◆ 不能直接表示多对多联系</li> <li>◆ 插入和删除数据限制太多</li> <li>◆ 查询子节点必须通过父节点</li> </ul>	<ul style="list-style-type: none"> <li>◆ 数据结构复杂</li> <li>◆ 数据定义语言复杂</li> <li>◆ 数据操作语言复杂</li> </ul>	<ul style="list-style-type: none"> <li>◆ 查询效率不如层次结构和网状结构</li> </ul>

数据操作：指的是对数据的一些操作，是对数据世界动态特征的描述，如查询、插入、修改、删除等。

数据约束：数据及数据关系的制约和存储规则，如取值唯一、不能为空等。

## ● 关系数据模型

关系数据模型是以关系数据结构为基础的一种数据模型，它把描述实体的属性及联系的属性看成是集合，而将实体和联系认为是建立在这些集合之上的关系。

为了形象地描述关系数据模型中的关系，通常是用一个二维表来表征。表中的行称为元组，列称为属性，表头称为属性名，单元格称为属性值，如图3.1.3所示。在软件开发术语中，关系叫作数据表或文件，元组叫作记录，属性叫作字段，属性名叫作字段名称，属性值叫作数据项。

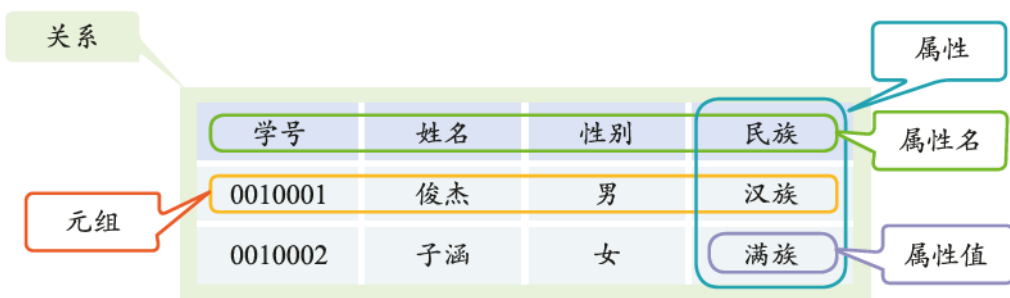


图 3.1.3 关系数据模型表征图

在关系数据模型中，二维表之间存在关联，主要通过主码和外码关联表征，比如商品表中的“供应商编号”和供应商表中的“供应商编号”之间，如图3.1.4所示。

**主码：**在关系数据模型中，指能够唯一标识一个元组的属性，也称主键。

**外码：**在关系数据模型中，一个关系中包含另一个关系的主码，该主码在此关系中称为外码，也称外键。

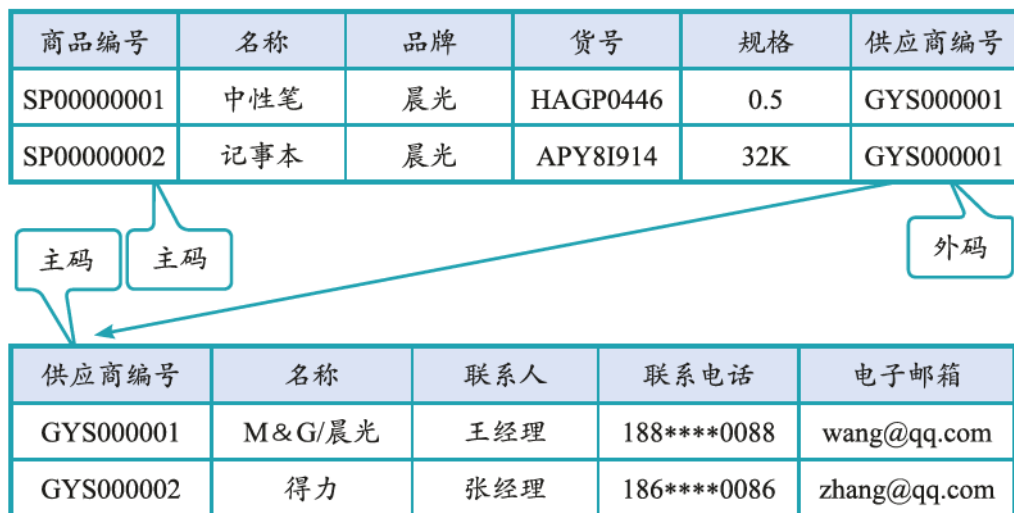


图 3.1.4 关系数据模型图例

依据数据模型及关系数据模型的有关知识，请帮助佳玮解析后台数据的结构，形成网店后台数据的关系数据模型。



## 拓展练习

请自行查找网上或生活中有关数据管理的案例，如学生学籍管理系统、学校图书管理系统、餐厅管理系统等，分析数据之间的关系，形成E-R图；剖析后台数据的结构，形成关系数据模型。



## 拓展知识

### 面向对象

面向对象的基本概念是在20世纪70年代形成的，它的基本做法是把系统工程中的某个模块和构件视为问题空间的一个或一类对象。到了20世纪80年代，面向对象的方法得到快速发展，在系统工程、计算机、人工智能等领域获得了广泛应用。

面向对象的基本思想是通过对问题领域进行自然的分割，用更接近人类通常思维的方式建立问题领域的模型，并进行结构模拟和行为模拟，从而使设计出的软件能尽可能地直接表现出问题的求解过程。

面向对象的方法就是以接近人类通常思维方式的思想，将客观世界的一切实体模型化为对象。每一种对象都有各自的内部状态和运动规律，不同对象之间的相互联系和相互作用就构成了各种不同的系统。

面向对象数据模型是将面向对象方法与数据库相结合所构成的数据模型，其示意图如图3.1.5所示。面向对象数据模型的优点包括：适合处理各种类型的数据，如文本、数字、图片、声音、视频等；有效地提高数据库应用程序开发人员的开发效率；比基于关系值的联系更能提高数据访问性能。面向对象数据模型的缺点包括：没有准确的定义；当改变对象的定义和移植数据库时，维护起来比较

困难；当用于普通应用时，其性能会降低且对处理能力的要求很高。

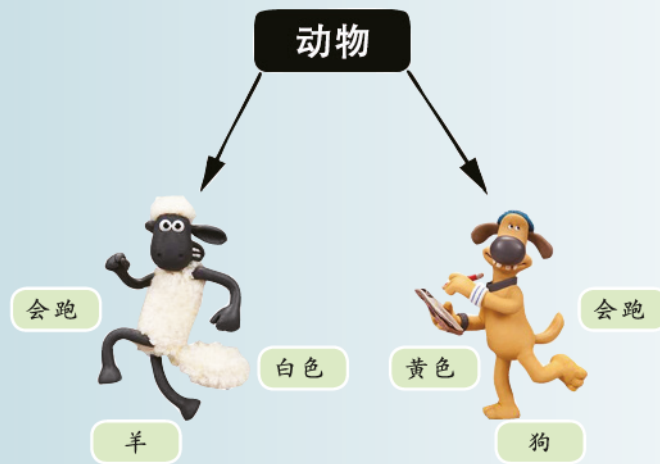


图 3.1.5 面向对象数据模型示意图

## 3.2 数据库的设计与创建

数据库是用来组织、存储数据的仓库，是最有效的数据管理工具之一。为了能够对数据进行安全、高效的管理，数据库的设计与创建必须科学、规范。那么，如何设计并创建数据库呢？



### 学习目标

- ★ 理解数据库范式与数据库管理系统。
- ★ 体验数据库设计、创建的过程与方法。



### 任务一 设计离线网店数据库

#### ※ 活动1 初步设计离线网店数据库

离线网店数据库是相对时光文具优选店的在线数据库而言的，由于数据库存储在本地计算机上，且并不实时地与网店相连，故称为离线网店数据库。佳玮想自己管理时光文具优选店的后台数据库，而同学们想通过管理时光文具优选店的后台数据库来学习数据库技术。由于入驻商城没有给客户提供数据库管理权限，所以大家决定模拟网店后台数据库，设计一个离线网店数据库，为以后自己真正管理网店数据库做好准备。

#### ● 数据库设计

完整的数据库设计内容包括需求分析、概念结构设计、逻辑结构设计、物理结构设计、数据库的实现和数据库的运行与维护。在此，我们更关注的是设计内容的核心部分，即需求分析、概念结构设计、逻辑结构设计和物理结构设计。

需求分析指在用户调查的基础上,通过分析,逐步明确用户对系统的需求,包括数据需求和围绕这些数据的业务处理需求。

概念结构设计指将用户需求中描述的现实世界抽象为概念世界,建立概念模型,形成E-R图。

逻辑结构设计指将E-R图转换成关系数据模型,形成数据库逻辑模式,在基本数据表的基础上建立必要的视图,形成数据库的外模式。

物理结构设计指根据数据库管理系统的处理和需要的需要,进行物理存储安排,建立索引,形成数据库的内模式。



视图是介于应用程序和数据表之间的虚表。

外模式是数据库用户能看见和使用的局部数据的逻辑结构和特征描述,对应于用户级。

内模式是数据库中全体数据的内部表示或底层描述,对应于物理级。

在离线网店数据库设计过程中,需求分析的调查对象主要是网店店主,需求的内容主要以业务流程和数据流程为主;概念结构设计主要是形成离线网店数据库的E-R图;逻辑结构设计主要是将离线网店数据库的E-R图转换成关系数据模型;物理结构设计主要是确定离线网店数据库中数据的存储结构、存取路径及存放位置。请采用分组分块的方式开展活动,即将数据库设计任务按照业务划分为客户管理、货源管理、商品管理、库存管理、订单管理、支付管理、物流管理及评价管理等若干块,每组同学分别选择一块,按照数据库设计流程进行离线网店数据库局部设计。最后,汇总各小组的成果,完成离线网店数据库的设计,形成离线网店数据库设计方案。

### 离线网店数据库设计方案

#### 【需求分析】

##### 1. 业务流程

##### 2. 数据流程

#### 【E-R图】

#### 【关系数据模型】

## 【数据库的物理结构】



设计关系数据库时，遵从不同的规范要求，设计出合理的关系数据模型，这些规范要求被称为数据库范式。

## ※ 活动2 优化离线网店数据库

数据库设计是一个“反复优化、逐步求精”的过程。我们已经设计好了离线网店数据库，接下来需要对其进行优化，即按照数据库范式进行规范。

## ● 数据库规范

数据库规范包括数据表规范、索引规范、视图规范等。在此，我们更为关注的是数据表规范。关于数据表的规范主要有六种范式，各种范式呈递次规范，越高的范式数据库冗余越小。六种范式中，常规使用的是前三种范式。

第一范式：要求数据表的每一列都是不可再分的原子数据项，而不能是集合、数组、记录等非原子数据项。

第二范式：在满足第一范式的基础上，要求数据表中的每一列完全依赖于主码。所谓完全依赖，是指不能存在仅依赖主码一部分的列。

第三范式：在满足第二范式的基础上，要求任何非主码不依赖于其他非主码，即任何非主码不得传递依赖于主码。

请按照前三种范式对设计好的离线网店数据库进行优化，并分享优化后的离线网店数据库设计方案。

## 离线网店数据库设计方案（优化后）

## 【需求分析】

1. 业务流程

2. 数据流程

【E-R图】

【关系数据模型】

【数据库的物理结构】



## 任务二 创建离线网店数据库

### ※ 活动1 创建离线网店数据库中的数据表

依据设计好的离线网店数据库设计方案，选择一个数据库管理系统，如Access、MySQL、SQL Server、Oracle等，创建离线网店数据库中的数据表。

### ● 数据库管理系统

数据库管理系统（Database Management System, DBMS）是一种创建和管理数据库的大型软件，用于建立、使用和维护数据库，其作用在于对数据库进行统一的管理和控制，以保证数据库的完整性、共享性、安全性。

一般而言，大部分数据库管理系统提供数据定义语言（Data Definition Language, DDL）和数据操作语言（Data Manipulation Language, DML），供用户定义数据库的模式结构与权限约束，实现对数据的添加、修改、删除、查询等操作。

数据定义语言包括数据库模式定义语言和数据库存储结构与存取方法定义语言。

数据操作语言用来表示用户对数据库的操作请求，是用户与数据库管理系统之间的接口。一般而言，对数据库的主要操作包括查询数据库中的信息、向数据库中插入新的信息、从数据库中删除信息以及修改数据库中的某些信息等。



**数据库模式定义语言：**用于定义数据库外模式、模式、内模式及它们之间的映射，作为数据库管理系统存取和管理数据的基本依据。

**数据库存储结构与存取方法定义语言：**用于定义数据库存储结构与存取方法，在存储设备上创建相关的数据库文件，建立起相应的物理数据库。



常见的数据库管理系统有Access、MySQL、SQL Server、Oracle等，各系统的比较如表3.2.1所示。

表 3.2.1 常见的数据库管理系统

类别	Access	MySQL	SQL Server	Oracle
出品公司	微软公司	MySQL AB公司	微软公司	甲骨文公司
系统特点	<ul style="list-style-type: none"> <li>◆ 界面友好</li> <li>◆ 易学易用</li> <li>◆ 开发简单</li> <li>◆ 接口灵活</li> </ul>	<ul style="list-style-type: none"> <li>◆ 体积小</li> <li>◆ 速度快</li> <li>◆ 成本低</li> <li>◆ 开放源码</li> </ul>	<ul style="list-style-type: none"> <li>◆ 使用方便</li> <li>◆ 伸缩性好</li> <li>◆ 与相关软件集成度高</li> </ul>	<ul style="list-style-type: none"> <li>◆ 高效可靠</li> <li>◆ 移植性好</li> <li>◆ 使用方便</li> <li>◆ 功能强大</li> </ul>
应用领域	适用于 小型企业	适用于 中小型企业	适用于 中小型企业	适用于 大、中、小型企业
应用平台	Windows平台	Linux、Unix、 Windows等多 平台	Windows平台	Linux、Unix、 Windows等多 平台

### ● 数据表

数据表也称文件，是数据库中的一个非常重要的对象，由若干字段构成，可存放若干记录。多个表之间存在关联，构成数据库。

数据表的创建过程可分为两步：首先，利用数据定义语言在存储设备上创建相关的数据库文件，建立起相应的物理数据库；之后，利用数据定义语言创建数据表，定义数据表各字段的字段名称、数据类型等，如定义离线网店数据库中的商品数据表（如表3.2.2所示）、供应商数据表（如表3.2.3所示）。

表 3.2.2 定义商品数据表

字段名称	数据类型	字段说明
GoodsID	长整型	商品编号
GoodsName	字符串	名称
Brand	字符串	品牌
GoodsNumber	字符串	货号
Specifications	字符串	规格
SupplierID	长整型	供应商编号

表 3.2.3 定义供应商数据表

字段名称	数据类型	字段说明
SupplierID	长整型	供应商编号
SupplierName	字符串	名称
Contacts	字符串	联系人
Telephone	字符串	联系电话
E-mail	字符串	电子邮箱

请依据数据库管理系统及数据表的相关知识，从Access、MySQL、SQL Server、Oracle等数据库管理系统中选择一个来创建离线网店数据库中的数据表。

#### ※ 活动2 为离线网店数据库中的数据表建立关系

请依据优化后的离线网店数据库设计方案，为离线网店数据库中已创建的数据表建立关系。

## ● 数据表间关系

### 主表

拥有主码的表称为主表，也称主键表。

### 相关表

拥有外码的表称为相关表，也称外键表。

### 关系完整性约束

关系完整性约束指为保证数据库中数据的正确性和相容性，对关系数据模型提出的约束条件或规则。它通常包括域完整性、实体完整性、参照完整性和用户定义完整性，其中前三个完整性是关系数据模型必须满足的完整性约束条件。

**域完整性：**指保证数据库字段取值的合理性。

**实体完整性：**指关系的主码不能重复也不能为空。

**参照完整性：**指保证关系的主码与外码引用的约束条件。例如，外码字段中的值必须在主码字段中存在，删除或修改主码字段中的值必须保证其没有被相关表引用。

我们可以利用数据库管理系统编辑关联表之间的关系，确定主表及主码、相关表及外码。例如，可以利用SQL Server编辑商品信息表和供应商信息表的关系，如图3.2.1所示。默认情况下，利用SQL Server编辑关系的过程中强制实施参照完整性约束。编辑关系的结果如图3.2.2所示。

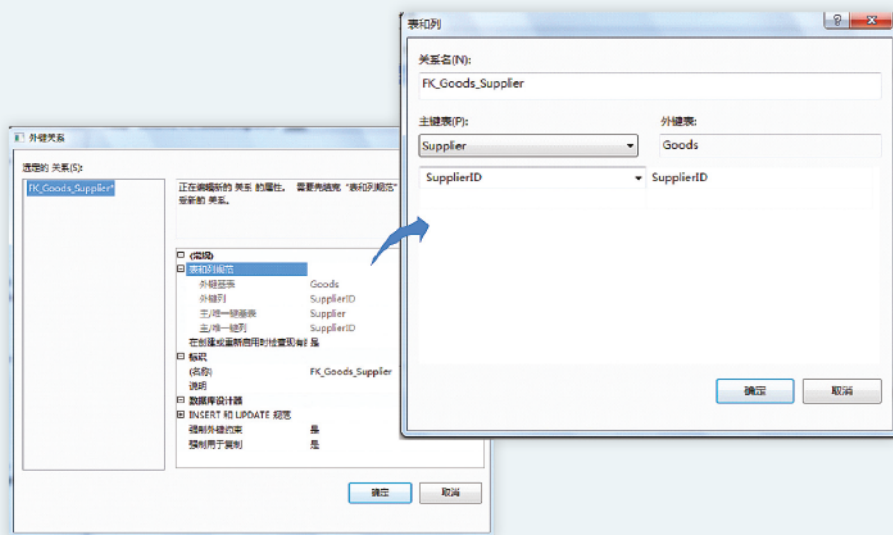


图 3.2.1 数据表间的关系编辑图

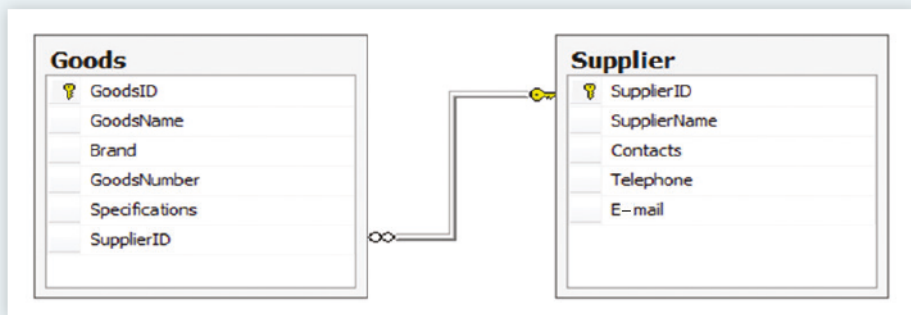


图 3.2.2 数据表间建立关系结果图

接下来，我们需要继续完善其他表之间的关系，完整地创建离线网店数据库OfflineDatabase。



### 拓展练习

请自行查找网上或生活中有关数据管理的案例，如学生学籍管理系统、学校图书管理系统、餐厅管理系统等，分析并规范后台数据库，形成后台数据库设计方案；依据设计方案，完整创建、分享、评价后台数据库。



## 拓展知识

### 数据库索引

在数据库管理系统中，利用数据库索引可快速查询、更新数据表中的数据，提高数据库访问速度。

创建数据库索引有以下优点：通过创建唯一性索引，可以保证数据表中每一行数据的唯一性；可以大大加快数据的检索速度，这也是创建索引的最主要的原因；可以加速表和表之间的连接，特别是在实现数据的参考完整性方面更有意义；在使用分组和排序子句进行数据检索时，同样可以显著减少查询中分组和排序的时间；借助索引，可以在查询的过程中使用优化隐藏器，提高系统的性能。

既然创建数据库索引有这么多优点，为什么不对数据表中的每一列创建一个索引呢？因为，创建数据库索引也有许多不利的方面：创建索引和维护索引要耗费时间，这种时间随着数据量的增加而增加；除了数据表占数据空间之外，每一个索引还要占一定的物理空间，如果要建立聚簇索引，那么需要的空间就会更大；当对数据表中的数据进行添加、删除和修改操作的时候，也要动态地维护索引，这样会降低数据的维护速度。

那么针对一个数据表，创建多少个索引比较合适呢？这需要数据库管理员根据实际的用途以及数据库中记录的情况来进行判断，具体策略就是在更新速度与查询速度之间取得一个均衡点。对于一些数据仓库或者决策型数据库系统来说，其主要用来进行查询，可以多设置一些索引；对于一些事务型的应用系统，需要经常更新数据表中的记录，此时就需要限制索引的数量，只在一些必要的字段上建立索引。

总之，基本的原则是在适当的数据表、适当的列上建立适当的索引。

## 3.3 数据库的操作与应用

利用数据库管理系统，不但能够存储数据，更重要的是能够处理和  
应用数据。它通过操作数据库，处理数据库事务，发挥数据应用的价值。  
那么，如何操作数据库？如何应用数据库管理系统处理数据库事务？



### 学习目标

- ★ 学会使用简单的 SQL 语句。
- ★ 体验数据库简单操作与事务处理的过程。
- ★ 感受数据管理的价值与意义。

在时光文具优选店的运营管理过程中，佳玮可以很方便地查询到  
订单、将货物上架。这些功能的实现都离不开数据库的操作与应用。  
为了能够真实体验数据库的操作与应用，佳玮打算和同学们一起把收  
集并整理好的时光文具优选店部分数据手动添加到自己创建的离线网  
店数据库OfflineDatabase中，从而了解前台应用程序是如何和后台数  
据库打交道的。



### 任务一 操作离线网店数据库

#### ※ 活动1 连接数据库

请根据离线网店数据库创建所使用的数据库管理系统，选择熟悉  
的编程语言来编写代码，完成数据库的连接。

#### ● 数据库连接

连接数据库是操作数据库的前提。创建数据库连接是前台应用程序  
和后台数据库进行交流的基础和桥梁，前台应用程序通过创建数据库连  
接方可访问后台数据库，如图3.3.1所示。

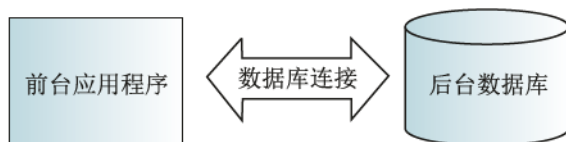


图 3.3.1 数据库连接的原理

创建数据库连接的一般步骤：

- 第一步，定义数据库连接字符串；
- 第二步，创建数据库连接；
- 第三步，打开数据库连接；
- 第四步，操作数据库；
- 第五步，关闭数据库连接。

选用C#语言，连接自己创建的OfflineDatabase数据库的代码示例如下。

```

using System.Data;
using System.Data.SqlClient;

// 定义数据库连接字符串,真正的程序中不要使用超级管理员 sa 做用户名,
// 密码也不能为空
string strConnection="Data Source=localhost;Initial Catalog =
OfflineDatabase;User ID=sa;Password='';
//创建数据库连接
SqlConnection objConnection=new SqlConnection(strConnection);
objConnection.Open();           //打开数据库连接
.....
objConnection.Close();         //关闭数据库连接
  
```

### ※ 活动2 操作数据库

目前，离线网店数据库已连接成功。接下来，我们可以尝试对离线网店数据库进行读写操作，具体包括添加、删除、修改和查询操作。

通常，前台应用程序借助数据库访问接口和结构化查询语言完成对后台数据库的操作。

### ● 数据库访问接口

数据库访问接口是前台应用程序和后台数据库交流的方法和工具，常见的有ODBC、JDBC、OLEDB、DAO、RDO、ADO等。其中，

ODBC、JDBC、OLEDB是底层的数据库接口，它们通过驱动程序访问后台数据库，而DAO、RDO、ADO、ADO.NET是上层数据库接口，它们向上与前台应用程序交互，向下与ODBC、JDBC或OLEDB对话。

### ● 结构化查询语言

结构化查询语言（Structured Query Language，SQL）是一种数据库查询和程序设计语言，用于存取数据以及查询、修改和管理关系数据库系统，具有极大的灵活性和强大的功能。在此，我们主要关注的是添加数据记录、删除数据记录、修改数据记录和查询数据记录四类SQL语句，如表3.3.1所示。

表 3.3.1 常用的SQL语句

类别	语法
添加数据记录	INSERT INTO <表名> (<字段列表>) VALUES (<表达式列表>)
删除数据记录	DELETE FROM <表名> [WHERE <删除条件>]
修改数据记录	UPDATE <表名> SET <表达式列表> [WHERE <修改条件>]
查询数据记录	SELECT [ALL DISTINCT] <字段列表> [INTO <新表名>] FROM <表名列表> [WHERE <查询条件>] [GROUP BY <字段名> [HAVING <条件表达式>]] [ORDER BY <字段名> [ASC DESC]]

### ● 添加数据记录操作

添加数据记录的一般步骤：

- 第一步，定义数据库连接字符串；
- 第二步，创建数据库连接；
- 第三步，定义添加数据记录的SQL语句；
- 第四步，创建数据库操作命令；
- 第五步，打开数据库连接；
- 第六步，执行数据库操作命令；
- 第七步，关闭数据库连接。

首先,进行添加数据记录的操作。根据离线网店数据库创建所使用的数据库管理系统,选择熟悉的编程语言来编写代码,完成添加数据记录操作。选用C#语言,对自己创建的OfflineDatabase数据库中的Goods表添加数据记录的代码示例如下。

```
using System.Data;
using System.Data.SqlClient;

//定义数据库连接字符串,真正的程序中不要使用超级管理员 sa 做用户名,
//密码也不能为空
string strConnection = "Data Source=localhost;Initial
Catalog=OfflineDatabase; User ID=sa; Password=''";
//创建数据库连接
SqlConnection objConnection = new SqlConnection
(strConnection);
//定义添加数据记录的SQL语句
string strInsertSql = "INSERT INTO Goods (GoodsName, Brand,
GoodsNumber) Values ('订书器', '得力', 'Deli0399')";
SqlCommand objCommand = new SqlCommand(strInsertSql,
strConnection);           //创建数据库操作命令
objConnection.Open();     //打开数据库连接
objCommand.ExecuteNonQuery(); //执行数据库操作命令
.....
objConnection.Close();    //关闭数据库连接
```

## ● 删除数据记录操作

删除数据记录的一般步骤:

- 第一步,定义数据库连接字符串;
- 第二步,创建数据库连接;
- 第三步,定义删除数据记录的SQL语句;
- 第四步,创建数据库操作命令;
- 第五步,打开数据库连接;
- 第六步,执行数据库操作命令;
- 第七步,关闭数据库连接。



接下来进行删除数据记录的操作。选用C#语言，对自己创建的OfflineDatabase数据库中的Goods表删除数据记录的代码示例如下。

```
using System.Data;
using System.Data.SqlClient;

//定义数据库连接字符串，真正的程序中不要使用超级管理员sa做用户名，密码也不能为空
string strConnection="Data Source=localhost; Initial Catalog=OfflineDatabase;User ID=sa;Password='';"
//创建数据库连接
SqlConnection objConnection = new SqlConnection (strConnection);
//定义删除数据记录的SQL语句
string strDeleteSql = "DELETE FROM Goods WHERE Brand = '得力'";
SqlCommand objCommand = new SqlCommand(strDeleteSql, strConnection);           //创建数据库操作命令
objConnection.Open();                       //打开数据库连接
objCommand.ExecuteNonQuery();               //执行数据库操作命令
.....
objConnection.Close();                     //关闭数据库连接
```

## ● 修改数据记录操作

修改数据记录的一般步骤：

- 第一步，定义数据库连接字符串；
- 第二步，创建数据库连接；
- 第三步，定义修改数据记录的SQL语句；
- 第四步，创建数据库操作命令；
- 第五步，打开数据库连接；
- 第六步，执行数据库操作命令；
- 第七步，关闭数据库连接。

接下来进行修改数据记录的操作。选用C#语言，对自己创建的OfflineDatabase数据库中的Goods表修改数据记录的代码示例如下。

```
using System.Data;
using System.Data.SqlClient;
```

```

//定义数据库连接字符串，真正的程序中不要使用超级管理员sa做用户名，密码也不能为空
string strConnection="Data Source=localhost; Initial Catalog=OfflineDatabase;User ID=sa;Password='';
//创建数据库连接
SqlConnection objConnection = new SqlConnection (strConnection);
string strUpdateSql="UPDATE Goods SET Brand='晨光'WHERE Brand='M&G/晨光';           //定义修改数据记录的SQL语句
SqlCommand objCommand=new SqlCommand(strUpdateSql, strConnection);           //创建数据库操作命令
objConnection.Open();           //打开数据库连接
objCommand.ExecuteNonQuery();           //执行数据库操作命令
.....
objConnection.Close();           //关闭数据库连接

```

## ● 查询数据记录操作

查询数据记录的一般步骤：

第一步，定义数据库连接字符串；

第二步，创建数据库连接；

第三步，定义查询数据记录的SQL语句；

第四步，创建数据库操作命令；

第五步，打开数据库连接；

第六步，执行数据库操作命令；

第七步，关闭数据库连接。

最后进行查询数据记录的操作。选用C#语言，对自己创建的OfflineDatabase数据库中的Goods表查询数据记录的代码示例如下。

```

using System.Data;
using System.Data.SqlClient;

//定义数据库连接字符串，真正的程序中不要使用超级管理员sa做用户名，密码也不能为空
string strConnection="Data Source=localhost; Initial Catalog=OfflineDatabase; User ID=sa;Password='';

```

```

//创建数据库连接
SqlConnection objConnection=new SqlConnection
(strConnection);
//定义查询数据记录的SQL语句
string strSelectSql="SELECT GoodsID,GoodsName,
Brand,GoodsNumber,SupplierID FROM Goods WHERE GoodsName=
'记事本' GROUP BY SupplierID ORDER BY GoodsID ASC";
SqlCommand objCommand=new SqlCommand(strSelectSql,
strConnection);           //创建数据库操作命令
objConnection.Open();     //打开数据库连接
//执行数据库操作命令
SqlDataReader objdataReader=objCommand.ExecuteReader();
.....
objConnection.Close();   //关闭数据库连接

```

我们还可以继续尝试对其他数据表中的数据进行添加、删除、修改和查询操作。



## 任务二 处理与网店相关的数据库事务

### ※ 活动1 分析网店交易过程

佳玮发现，网店的订单生成的同时，用户的付款信息也一并生成了，并且库存信息和用户购物信息会自动更新。请帮助佳玮分析网店交易过程，判断网店交易过程能否成为数据库事务，解释“一并生成”与“自动更新”现象。

### ● 数据库事务

数据库事务（Database Transaction）是作为单个逻辑工作单元执行的一系列操作，要么完全地执行，要么完全不执行。具体而言，如果某一事务成功，则在该事务中进行的所有数据库操作都会提交，成为数据库中的永久组成部分；如果事务遇到错误且必须取消或回滚，则所有数据库操作均被清除。其目的是增加应用程序的可靠性。

一个逻辑工作单元要想成为一个事务，必须具有四个属性：原子性、一致性、隔离性和持久性。

原子性：事务必须是原子工作单元，对于其数据修改，要么全都执行，要么全都不执行。

一致性：事务开始前，数据库处于一致性的状态；事务结束后，数据库必须仍处于一致性状态。例如，银行转账前后总金额一致，商品交易前后商品总数一致。

隔离性：由并发事务所做的修改必须与任何其他并发事务所做的修改隔离，即一个事务的执行不能被其他事务干扰，系统必须保证事务不受其他并发执行事务的影响。

持久性：事务完成之后，它对于系统的影响是永久性的。即使出现系统故障，该影响也将一直保持。

请依据数据库事务的定义及属性特征，分析网店交易过程，根据数据库事务的四个属性来判断网店交易过程能否成为数据库事务。

### 网店交易过程分析

#### 【分析目的】

判断网店交易过程能否成为数据库事务。

#### 【分析过程】

属性特征	具有	不具有
原子性		
一致性		
隔离性		
持久性		

通过分析可以确定，网店交易过程是数据库事务。网店交易过程需要对数据库进行一系列操作，具体包括生成订单信息、生成付款信息、更新库存信息、更新用户的购物信息，而且这几个操作如同捆绑在一起一样，要么全部成功，要么全部撤销，否则就会出现商品交易前后总数不一致、用户付款前后总金额不一致等现象。

#### ※ 活动2 处理网店交易事务

请依据显式事务处理步骤，对自己创建的OfflineDatabase数据库中的订单表（Order）、支付表（Payment）、库存表（Stock）、购物表（Shopping）进行事务处理，模拟处理网店交易事务。

## ● 事务处理原理

如前面所述，数据库事务是将一系列操作作为一个单元执行，要么成功，要么失败而回滚到最初状态。在事务处理术语中，事务要么提交，要么中止。若要提交事务，所有参与者都必须保证对数据的任何更改是永久的。不论系统崩溃或是发生其他无法预料的事件，更改都必须是持久的。只要有一个参与者无法做出此保证，整个事务就会失败。事务范围内的所有数据更改将回滚到特定设置点，事务处理过程如图3.3.2所示。

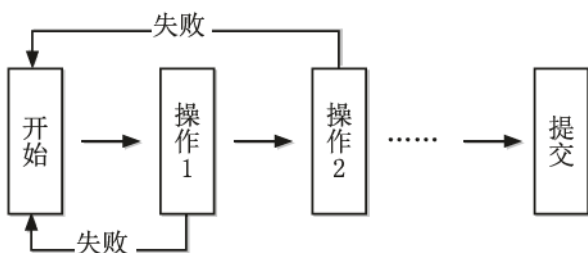


图 3.3.2 事务处理原理

## ● 事务处理过程

数据库事务处理可分为显式和隐式两种。显式事务通过命令实现，隐式事务由系统自动完成提交或撤销工作，无须用户的干预。在这里，我们主要介绍显式事务处理过程。显式事务均以“事务开始”语句显式开始，以“事务提交”或者“事务回滚”语句显式结束。其中，事务提交是指成功完成事务，数据库操作结果永久化；事务回滚是指中止当前事务，撤销对数据库的所有更改。显式事务的具体处理步骤如下：

- 第一步，事务开始；
- 第二步，执行一系列数据库操作；
- 第三步，事务提交或者事务回滚。

选用C#语言，对自己创建的OfflineDatabase数据库中的订单表（Order）、支付表（Payment）、库存表（Stock）、购物表（Shopping）进行事务处理的代码示例如下。

```

using System.Data;
using System.Data.SqlClient;

string strConnection = "Data Source=localhost;Initial
Catalog=OfflineDatabase;User ID=sa;Password='';
SqlConnection objConnection = new SqlConnection
  
```

```
(strConnection);
SqlTransaction myTransaction = null;
try
{
objConnection.Open();
//事务开始
myTransaction = objConnection.BeginTransaction();
SqlCommand objCommand = new SqlCommand("",objConnection,
myTransaction);
objCommand.CommandTimeout = 120;
objCommand.CommandType = System.Data.CommandType.Text;

string strInsertOrderSql = "INSERT INTO Order……";
string strInsertPaymentSql = "INSERT INTO Payment ……";
string strUpdateStockSql = "UPDATE Stock……";
string strUpdateShoppingSql = "UPDATE Shopping……";

objCommand.CommandText = strInsertOrderSql;
objCommand.ExecuteNonQuery();
objCommand.CommandText = strInsertPaymentSql;
objCommand.ExecuteNonQuery();
objCommand.CommandText = strUpdateStockSql;
objCommand.ExecuteNonQuery();
objCommand.CommandText = strUpdateShoppingSql;
objCommand.ExecuteNonQuery();

throw new Exception("test exception");
myTransaction.Commit();           //事务提交
}
catch (Exception ex)
{
myTransaction.Rollback();       //事务回滚
}
finally
{
if (objConnection.State != System.Data.ConnectionState.
Closed)
objConnection.Close();
}
……
```

我们还可以继续尝试处理网店管理中的其他数据库事务。



## 拓展练习

请自行查找网上或生活中的数据管理案例，如学生学籍管理系统、学校图书管理系统、餐厅管理系统等，完整创建后台数据库，对数据库进行添加、删除、修改、查询操作，分析并处理数据库事务。



## 拓展知识

### 数据库中间件

中间件在操作系统、网络和数据库之上，在应用软件之下，总的作用是为处于自己上层的应用软件提供运行与开发的环境，帮助用户灵活、高效地开发和集成复杂的应用软件。

中间件是一类软件，而非一种软件；中间件不仅仅实现互联，还要实现应用之间的互操作；中间件是基于分布式处理的软件，最突出的特点是其网络通信功能。

数据库中间件主要分为分布式数据库分库分表、数据增量订阅与消费、数据库同步、跨数据库迁移四类。对于分布式数据库分库分表类的中间件，主要以两种形式向应用提供服务。一种是以JDBC的jar包形式为Java应用提供直接依赖，Java应用通过提供的JDBC包实现透明访问分布式数据库集群中的各个分库分表，典型代表是网易的分布式数据库（Distributed Database, DDB）和阿里的淘宝分布式数据层（Taobao Distributed Data Layer, TDDL）。另一种是为应用部署独立的服务来满足应用分库分表的需求，在这种方式下通过标准JDBC访问Proxy，而Proxy则根据MySQL标准通信协议对客户端请求解析，还原应用SQL请求，然后通过本地访问数据库集群，最后再将得到的结果根据MySQL标准通信协议编码返回给客户端。典型代表是阿里的Cobar、Cobar变种MyCat、阿里的DRDS、网易的DDB Proxy模式以及DDB的私有云模式。

MyCat主要有以下优势：可以通过MyCat统一管理所有的数据源，后端数据库集群对前端应用程序透明；采用独创的E-R关系的数据分片策略，解决分片难处理问题；采用全局分片技术，每个节点同时并发插入和更新数据，每个节点都可以读取数据；通过人工智能的catlet支持跨分片复杂SQL实现以及存储过程支持等。

MyCat的下一步规划是要强化分布式数据库中间件的面的功能，使之具备丰富的插件、强大的数据库智能优化功能、全面的系统监控能力以及方便的数据运维工具，实现在线数据扩容、迁移等高级功能。

## 单元学习评价

通过本单元的学习，我们完成了数据库的设计、操作与应用任务，体验了数据管理的全过程。你能否胜任数据管理的相关工作？请参与小组交流、讨论并反思，开展小组评价或自我评价。

1. 概念模型的要素主要包括\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_，数据模型的要素主要包括\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_，数据库事务的属性特征有\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。

2. 一位教师可以给多名学生上课，一名学生可以选择多位教师的课，教师和学生之间是典型的多对多联系。请依据E-R图的有关知识，绘制教师和学生的E-R图。

3. 请依据数据库范式的相关知识，完成如下表所示的任务。

学号	姓名	年龄	生日	父母姓名	课程	成绩	等级
001	佳玮	17	2001-5-1	张志，李梅	语文	80	良

观察上表，进行调整，使其符合第一范式，在此基础上可继续调整，使其符合第二范式、第三范式。

4. 请对自己或小组负责的数据管理项目的实施过程及结果进行总结，分析项目实施的重点、难点，评价项目各个阶段的成果有无缺陷，给出相应的改进策略。

### 数据管理项目总结

【分析项目实施的重点】

【分析项目实施的难点】

【各个阶段的成果评价】

1. 数据库设计方案



2. 数据库创建熟练程度

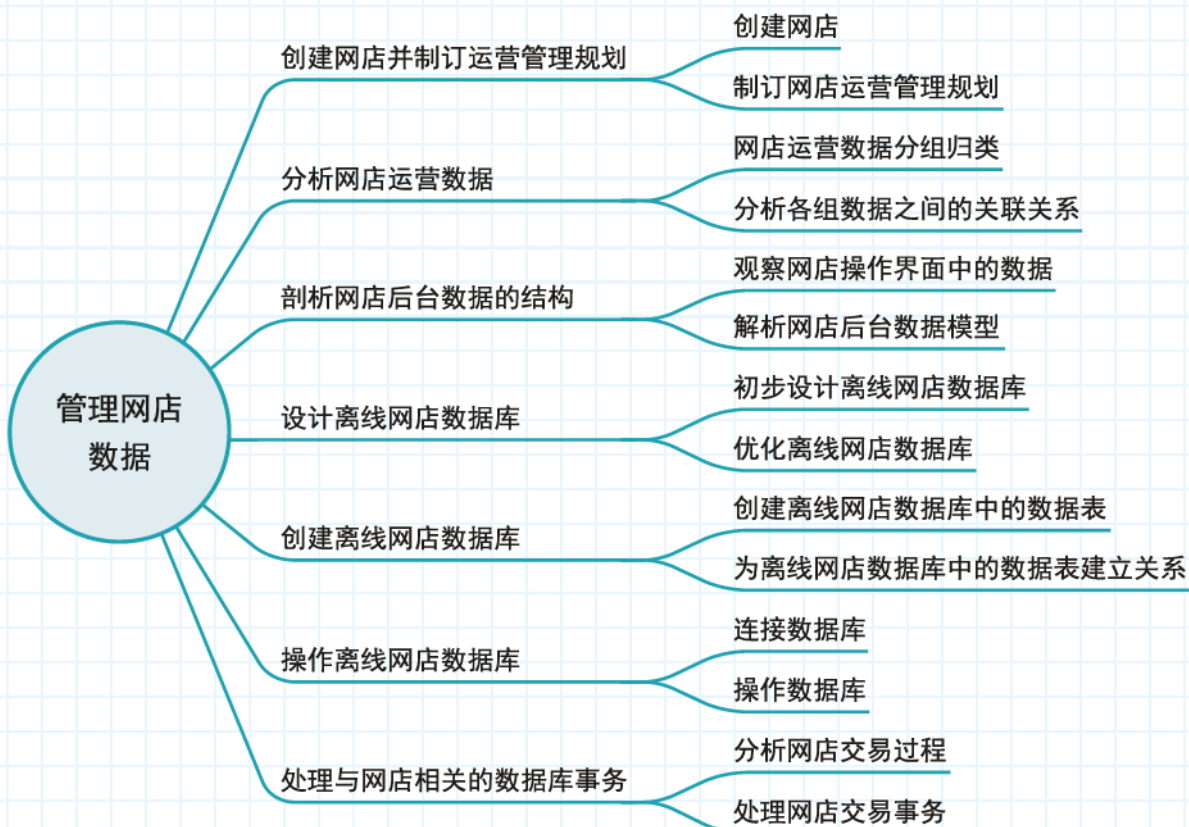
3. 数据库操作熟练程度

【项目实施与阶段成果改进策略】

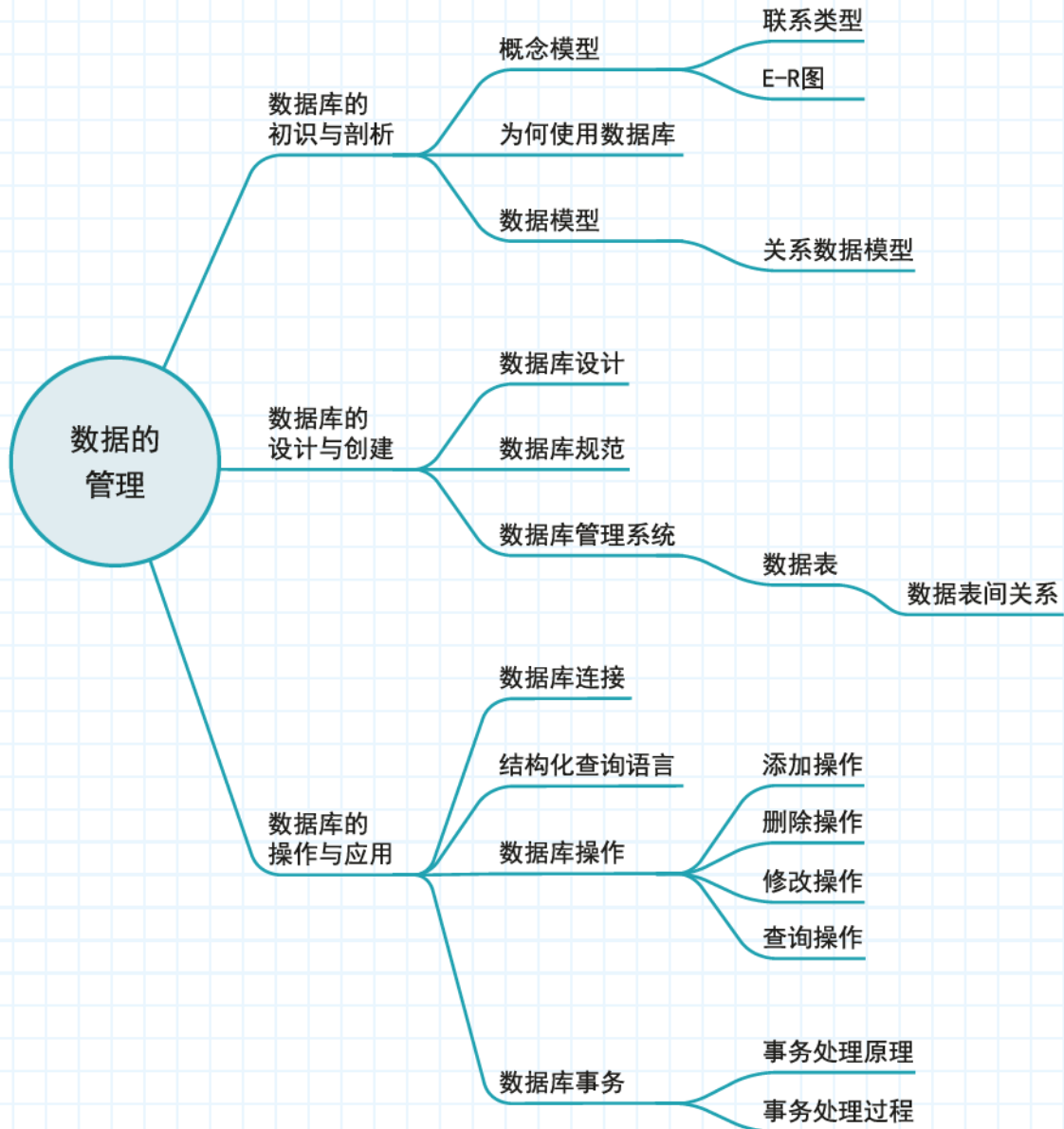
## 单元学习总结

本单元围绕“管理网店数据”项目展开学习，通过项目的规划与设计，明确项目实施的具体任务，开展相应的主体活动，理解概念模型、数据模型，掌握关系数据库结构，了解数据库设计、数据库规范等概念，学习如何创建并简单操作数据库，以及如何处理常见的数据库事务，解决现实中的数据管理问题，从而最终形成对数据存储、处理和应用的完整认识。

### “项目—任务—活动”结构图



### 知识地图



## 第 4 单元 数据的分析与可视化

信息时代是一个用数据说话、用数据竞争的时代。目前，数据分析部门和数据分析团队已经成为大多数企业的常设机构，越来越多的人意识到数据已经成为一种新的资产和资源。

在本单元中，我们将通过“网店开学季促销与活动总结”项目展开学习，掌握数据分析与可视化的方法，学会使用数据分析与可视化工具，经历数据分析与可视化的过程，形成数据分析与可视化的思想。

为了做好网店促销活动及活动后的总结，需要考虑一系列的问题：网店将采用哪种形式的促销活动？促销期间网店的销售业绩如何？如何展示销售业绩？要解决这些问题，我们需要完成以下四个任务：

- ◆ 规划促销活动；
- ◆ 分析网店的销售业绩；
- ◆ 生成网店销售业绩报表；
- ◆ 撰写网店销售业绩分析报告。

## 4.1 数据分析方法

掌握数据分析方法是对数据分析人员的基本要求。本节将通过完成“规划促销活动”任务，让大家参与数据分析，体验数据分析的过程，掌握数据分析的一般方法。

### 学习目标

- ★ 了解数据分析方法论，学会使用平均分析法。
- ★ 体验数据分析框架构建的过程。
- ★ 初步形成利用数据分析方法解决数据问题的意识。

佳伟的网店开业不久，赶上电商平台开展开学季的促销活动。为了提高店铺的销售业绩，佳伟准备针对本店铺的某些学习用品开展一次促销活动。因此，需要在活动开始前先做出促销规划。

### 任务 规划促销活动

#### ※ 活动1 设计促销方式

任何有目的的活动都需要做好前期规划。一个有效的规划需要有一定的理论方法做指导，促销活动也是如此。只有在促销活动开始之前理清思路，才能使活动顺利开展。

### ● 数据分析方法论

数据分析方法论主要是从宏观角度指导如何进行数据分析，它是数据分析的前期规划，指导着后期数据分析工作的开展，它更大程度上代表的是数据分析思路。



**数据分析方法**还有很多种,如管理方面的有5W2H、PEST、金字塔理论,营销方面的有用户使用行为、STP理论、SWOT等。可以根据不同的需求,运用不同的方法开展相关活动的规划。

## ● 4P理论

4P是指产品(Product)、价格(Price)、渠道(Place)、促销(Promotion)。4P理论是一种能对数据分析进行专门指导的数据分析方法,为企业的营销策划提供了一个可参照的框架。

产品指满足人们使用、消费的需求,让经营者产生利润的东西。

价格指顾客购买产品时的价格,包括折扣、支付期限等。价格关系到利润以及是否有利于产品销售、促销等问题。

渠道指商品从生产者到消费者手上的全过程中所经历的各个环节和推动力量。例如,在线销售、线下销售等。

促销指通过销售行为的改变来刺激用户消费,促成消费的增长。例如,买二赠一、赠送优惠券等。

在网店促销框架的设计中,要考虑参与促销的产品、促销价格、促销渠道与具体的促销方式等基本要素。请根据4P理论,构建网店促销框架,完成表4.1.1。

表 4.1.1 网店促销计划表

活动框架	具体内容
促销产品	
促销价格	
促销渠道	
促销方式	

### ※ 活动2 确定优惠券面额

根据网店的促销计划,佳玮想在本次促销活动中选择向消费者赠送优惠券的促销方式,但是优惠券赠送的门槛需要结合店铺的客单价,一般建议优惠券赠送的门槛比店铺客单价高10%。例如,客单价是90元,则建议优惠券赠送的门槛为99元。



**客单价**是指商场(超市)每一个顾客平均购买商品的金額,即平均交易金額。

## ● 数据分析方法

数据分析方法主要从微观角度指导如何进行数据分析,具体包括平均分析法、比较分析法、相关分析法等。

## ● 平均分析法

平均分析法是指利用计算平均数的方法，反映总体在一定时间、地点条件下某一数量特征的一般水平。一般而言，数据分析中的平均分析法分为位置平均数和数值平均数两种类型，且两类方法都有相应的平均指标作为参考，如图4.1.1所示。

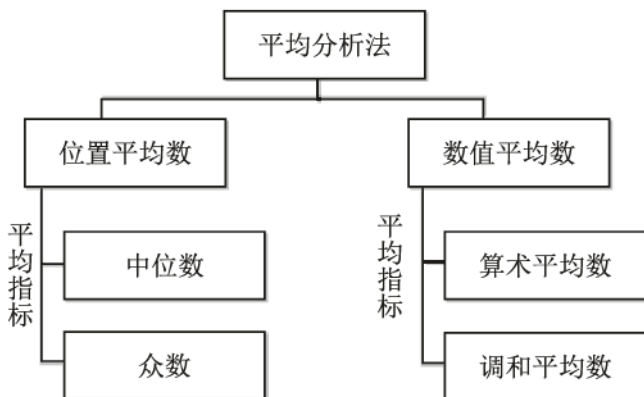


图 4.1.1 平均分析法

位置平均数的平均指标是中位数和众数。中位数是指将数据按大小顺序排列起来，形成一个数列，居于数列中间位置的那个数据。众数是指一组数据中出现次数最多的那个数据，一组数据可以有多个众数，也可以没有众数。从分布角度看，众数是具有明显集中趋势的数值。

数值平均数的平均指标是算术平均数和调和平均数。算术平均数是指在—组数据中所有数据之和再除以数据的个数；调和平均数又称倒数平均数，是各数据倒数的算术平均数的倒数。

在数据分析中，运用得最多的就是算术平均数。它是非常重要的基础性指标，其计算公式如下：

$$\text{算术平均数} = \frac{\text{总体单位数值的总和}}{\text{总体单位个数}}$$

客单价平均值的计算主要运用的是平均分析法。我们可将客单价的运算过程用文字的方式记录在下面。

---



---



---



客单价的计算公式是“客单价=销售总额÷顾客总数”，或者是“客单价=销售总额÷成交总笔数”。



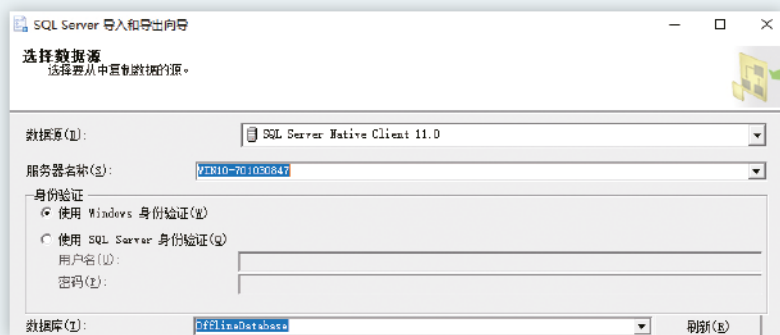


图 4.1.3 数据导出操作2

第三步，在“选择目标”对话框中，设置目标为“Microsoft Excel”，并设置目标文件所在的位置，如图 4.1.4 所示。

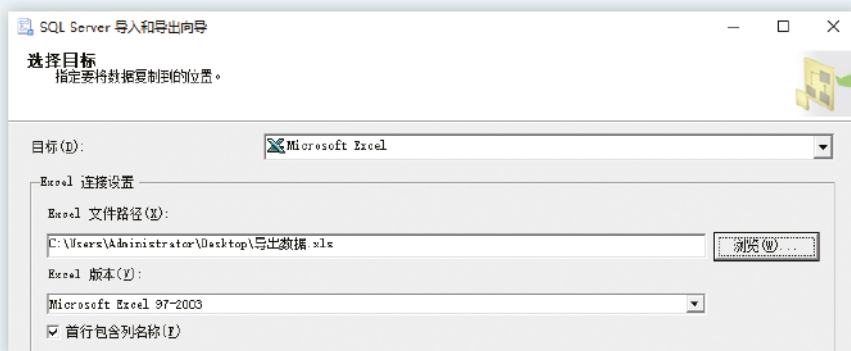


图 4.1.4 数据导出操作3

第四步，在“指定表复制或查询”对话框中，选中“复制一个或多个表或视图的数据”单选按钮，如图 4.1.5 所示。

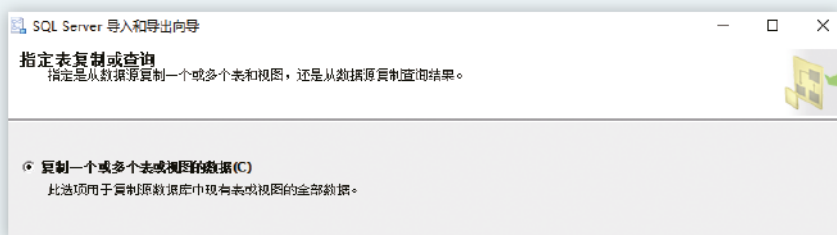


图 4.1.5 数据导出操作4

第五步，在“选择源表和源视图”对话框中选择一个或多个源表或源视图，例如“[dbo].[Goods]”，如图 4.1.6 所示。

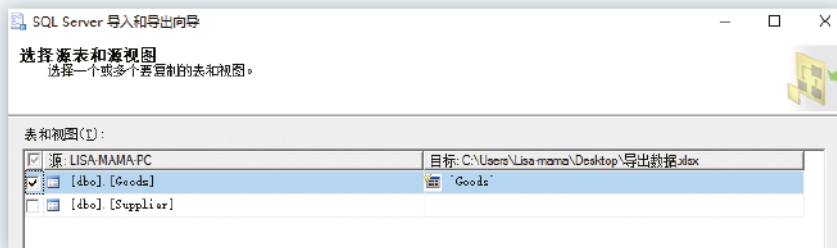


图 4.1.6 数据导出操作5



第六步，在“保存并运行包”对话框中选中“立即运行”复选框，如图4.1.7所示。

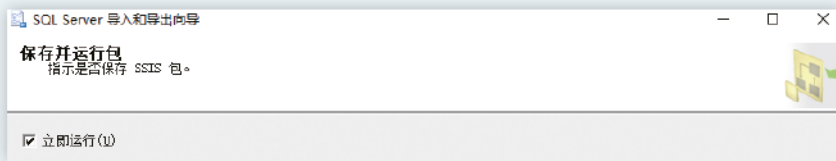


图4.1.7 数据导出操作6

第七步，单击“完成”按钮，出现“完成向导”对话框，如图4.1.8所示。



图4.1.8 数据导出操作7

第八步，打开目标文件，验证导出的结果，如图4.1.9所示。

编号	商品名称	品牌	销售量	价格	供货商
1	AT0044 奥品水晶学生办公桌24号桌	晨光	53	24.9 01	
2	A00099 永佳牌(0)管理桌(0.7m*1.6.7m)	晨光	92	5.9 01	
4	8100 炫彩圆珠笔(0)笔尖0.5mm 12支/盒	晨光	27	33.9 01	
5	A20000 永佳牌(0)学生(0)课桌10号	晨光	34	30.9 01	
6	AP0007 学生办公(0)课桌(0)课桌	晨光	61	18.9 01	
7	A00070 永佳牌(0)学生(0)课桌(0)课桌	晨光	19	1.99 01	
8	ACT000 学生(0)课桌(0)课桌	晨光	62	6.2 01	
9	AP0000 永佳 10(0)课桌(0)课桌	晨光	30	30.9 01	
10	ACT000 学生(0)课桌(0)课桌	晨光	57	24.9 01	
11	AP0001 记事本(0)记事本	晨光	24	1.97 01	
12	AM0009 办公(0)资料(0)文件夹(0)文件夹	晨光	55	15.9 01	
13	AM0006 学生(0)课桌(0)课桌	晨光	20	24.9 01	
14	A00000 学生(0)课桌(0)课桌	晨光	31	6.9 01	
15	A0.3500 自动(0)订书机(0)订书机	晨光	17	2.4 01	
16	SP0007 学生(0)课桌(0)课桌	晨光	95	20.9 01	
17	ACT000 学生(0)课桌(0)课桌	晨光	50	2.9 01	
18	ACT000 学生(0)课桌(0)课桌	晨光	14	10.9 01	
19	A0.2000 自动(0)订书机(0)订书机	晨光	12	1.9 01	
20	A0.2000 自动(0)订书机(0)订书机	晨光	19	1.9 01	
21	AP0001 自动(0)订书机(0)订书机	晨光	95	4.99 01	
22	SP0008 学生(0)课桌(0)课桌	晨光	74	28.9 01	
23	ACT000 学生(0)课桌(0)课桌	晨光	22	2.9 01	
24	PL000 金属(0)订书机(0)订书机	晨光	92	6.5 01	
25	SP 学生(0)课桌(0)课桌	晨光	46	12.99 01	
26	ACT000 学生(0)课桌(0)课桌	晨光	29	5.9 01	
27	ACT000 学生(0)课桌(0)课桌	晨光	66	6.9 01	

图4.1.9 数据导出操作8



## 拓展练习

经历上述活动后，请同学们进行小组合作，设计一份网店的开学季促销活动海报，将促销的产品、价格、渠道、具体的促销方式等在海报中呈现。



## 拓展知识

### 分析性思维

在日常生活中，我们总能听到“新的短跑世界纪录诞生了”“计算机的性能提高了”这样的说法。从数据分析的角度来看，这些表述仍然略显不足。它们只是指出某样东西随着时间的推移而改进，那么改变是大还是小呢？与其他改变相比，结果又如何呢？

1951年问世的世界第一台商业计算机UNIVAC I能以大约10万次每秒加法运算的速度工作。相比之下，1999年英特尔公司为美国桑迪亚国家实验室（Sandia National Laboratories）制作的ASCI Red超级计算机实现了每秒万亿次浮点运算。但是，我们不是分开来研究这两个数据，相反，我们通过计算提升系数来分析二者之间的关系。这种分析使我们能将计算机的进步和其他方面的进步进行比较，并能用统一的眼光来看待它们。

使用正确的方法能进行科学的数据分析固然十分重要，但是以分析性思维来思考问题，同样是一个非常有用的好习惯。要熟练地运用数据，分析性思维必不可少，而且在我们其他方面的学习中，在我们的职业生涯中，以及我们的整个生命中，分析性思维都发挥着关键作用，可以帮助我们更深入地理解信息技术世界以及我们居住的这个世界。

## 4.2 数据分析工具

随着数据分析技术的发展，相继出现了各种实用的数据分析工具。本节将通过完成“分析网店的销售业绩”任务，引导同学们与数据近距离接触，并通过具体的活动，体验不同的数据分析方法在数据分析工具中的应用。



### 学习目标

- ★ 了解数据分析工具。
- ★ 体会常用的数据分析方法。
- ★ 体验应用数据分析工具与方法分析数据问题的过程。



销售业绩是指一个时间段或者阶段展开销售业务后的收益总结。销售业绩可以从销量、销售额等方面来考查。

电商平台的开学季促销活动结束了，佳玮很关心这段时间网店中相关商品的销售情况，她希望为此次促销活动做一个总结，并向大家展示自己的劳动成果。



### 任务 分析网店的销售业绩

#### ※ 活动1 观察在线数据分析结果

佳玮想了解网店的总体销售业绩，她可以借助哪些工具达到目的呢？在线数据分析工具可以方便、快捷地呈现分析结果，她可以借助在线数据分析工具来分析销售业绩。

#### ● 常用的在线数据分析工具

在线数据分析工具多种多样，常用的在线数据分析工具及其功能如表4.2.1所示。

表 4.2.1 常用的在线数据分析工具

工具名称	功能
生意参谋	生意参谋平台是阿里巴巴推出的官方数据产品门户，提供一站式、个性化、可定制的数据决策体验
数据罗盘	数据罗盘是京东商城为开放平台商家推出的数据工具，包括店铺分析、行业数据和京东实验室三个模块，提供流量分析、销售分析、广告分析、仓储分析、配送分析、售后分析、行业分析等多维度的数据
生e经	生e经是淘宝店铺的数据分析工具，可以帮助商家及时了解店铺的经营状况
百度指数	百度指数是以百度海量网民行为数据为基础的数据分享平台。在这里，你可以研究关键词搜索趋势、洞察网民兴趣和需求、监测舆情动向、定位受众特征

使用在线数据分析工具，我们通常需要执行三个步骤（以知己知彼网站为例）。

第一步，注册账号，登录在线数据分析工具，如图4.2.1所示。

第二步，选择相关的功能模块，如图4.2.2所示。

第三步，查看网店的销售情况，如图4.2.3所示。



图 4.2.1 在线数据分析工具操作1



图 4.2.2 在线数据分析工具操作2



图 4.2.3 在线数据分析工具操作3

### ※ 活动2 选择数据分析工具

在线数据分析工具会给我们提供分析结果，而这些分析结果又是如何得到的呢？如果我们选择的是之前导出的离线数据库中的数据，那么利用什么工具能够实现数据分析的过程？

#### ● 离线数据分析工具

能够实现数据分析过程的还有离线数据分析工具，这些工具中有进行基础分析的工具，还有进行专业分析的工具。基础的分析工具以微软的Excel软件为代表，专业的分析工具有SPSS、SAS、Stata、MATLAB等，表4.2.2列出了其中几款常用的离线数据分析工具。

表 4.2.2 常用的离线数据分析工具

工具名称	简介	适用情境	优点	缺点
Excel	Excel 是微软办公套装软件的一个重要的组成部分，它可以进行各种数据的处理、统计分析和辅助决策操作，广泛地应用于管理、统计财经、金融等众多领域	初级数据分析	基本工具、使用广泛、模块简单	功能简单、适用场景较少
SPSS	SPSS为IBM公司推出的一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务	高级数据分析	模块丰富、场景控制	功能复杂、使用者需有实践经验
MATLAB	MATLAB是美国MathWorks公司出品的商业数学软件，用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境	高级数据分析	工具丰富	功能复杂、使用者需有一定的数学基础和编程能力

如果要选择一款离线数据分析工具来分析从离线数据库中导出的数据（可以参考教科书配套资源“Goods.xlsx”中的数据），你会选择哪一款？请上网查阅表4.2.3中各工具的详细资料，之后做出你的选择，在相应工具后面画“√”并说明理由。

表 4.2.3 数据分析工具的选择

工具名称	你的选择	理由
Excel		
SPSS		

续表

工具名称	你的选择	理由
Stata		
SAS		
MATLAB		
Python		
R		

### ※ 活动3 分析环比、同比

佳伟想知道促销活动期间参与促销商品的环比、同比，我们能够用什么方法来分析环比、同比呢？如果是在数据分析工具中，如何得到这些结果呢？（可以参考教科书配套资源“Goods.xlsx”文件中的数据）

### ● 比较分析法

比较分析法又称对比分析法，指将两个或两个以上的数据进行对比，分析它们之间的差异。它是一种简便的、能快速分析出结果的分析方法。运用对比分析法能直观地分析出事务某些方面的差异或变动，并且可以通过数据量化出变动所产生的差距，从而针对差距进行调整。在进行对比分析时，一定要选择具有可比性的两者或多者进行对比，这样才有意义。常见的对比方式有以下几种，如图4.2.4至图4.2.8所示。

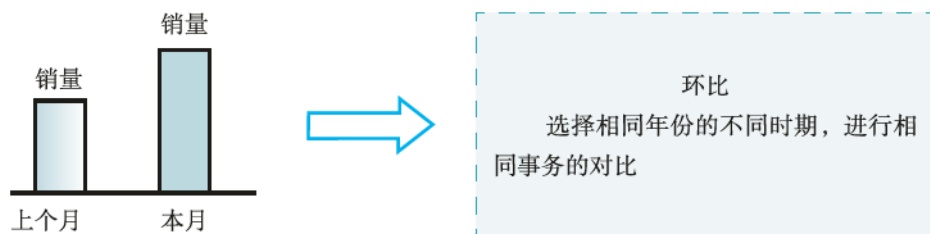


图 4.2.4 常见的对比方式1

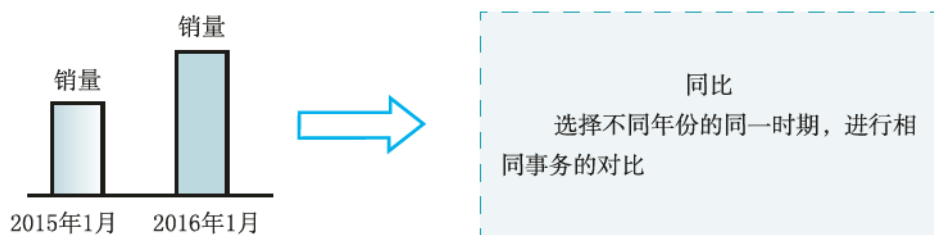


图 4.2.5 常见的对比方式2

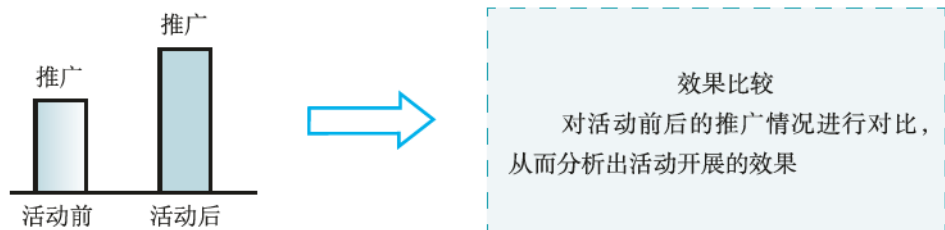


图 4.2.6 常见的对比方式3

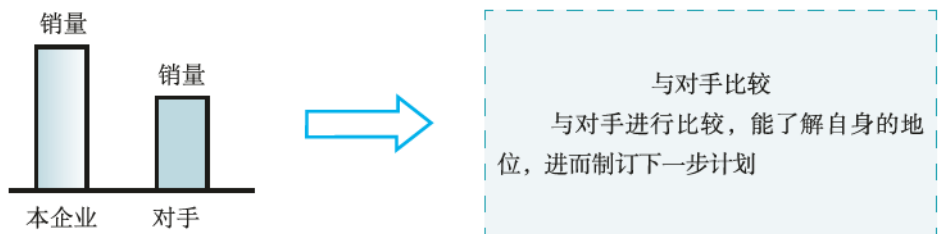


图 4.2.7 常见的对比方式4

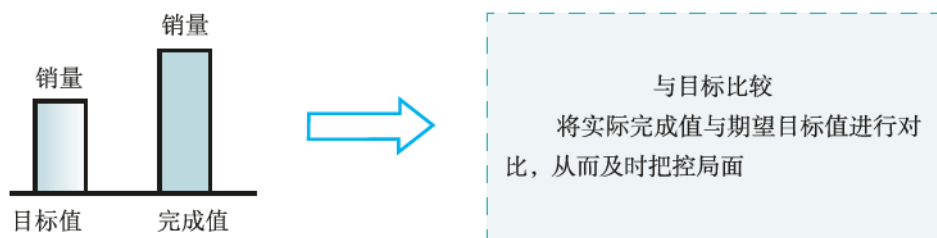


图 4.2.8 常见的对比方式5

此外，比较分析法经常与分组分析法结合起来使用，从而得到更精确的分析结果。

### ● 分组分析法

分组分析法就是依据数据分析对象的特征，按照一定的指标，将对象划分成不同的类别进行分析，以便深入了解数据分析对象的内部结构或现象之间的依存关系，从而挖掘和分析出分析对象不同的特征和性质。常用的分组分析方法有以下三种。

数量分组：计算各组单位数、分组指标量在总体总量中所占的比重。

关系分组：根据紧密的因变量与自变量之间的储存关系进行分组。

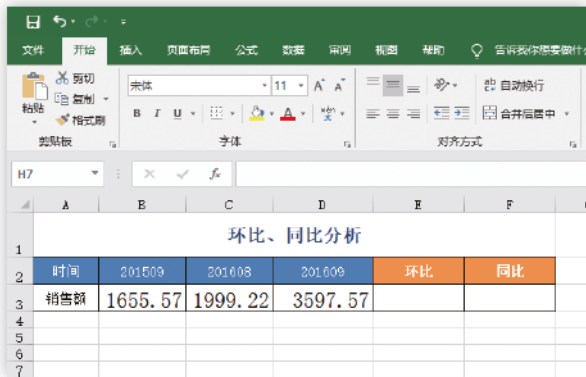
质量分组：按照不同质量分组，进行各组之间的数量分析。

这三种方式可以结合使用，但是分组时必须遵循两个原则：一是单一原则，即总体中的任何一个单位只能归属于某一个组，不能重复划分；二是不漏原则，即总体中的每一个单位，都要分配到某一个组中，不能遗漏。

## ● 分析环比、同比

Excel软件是一款常用的数据分析工具，在Excel中，可以通过计算公式帮助我们完成分析。

第一步，制作环比、同比分析表格，表格中包括同年份不同时期和不同年份同一时期的数据，如图4.2.9所示。



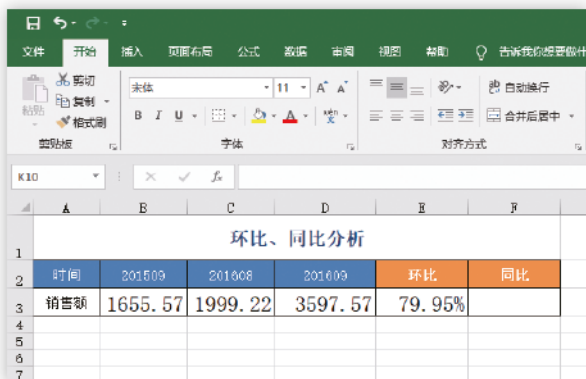
环比、同比分析					
时间	201509	201608	201609	环比	同比
销售额	1655.57	1999.22	3597.57		

图 4.2.9 数据分析操作1

第二步，选中环比数据单元格（E3），输入计算公式： $= (D3-C3)/C3*100\%$ ，设置E3单元格的数字格式为百分比，如图4.2.10所示。



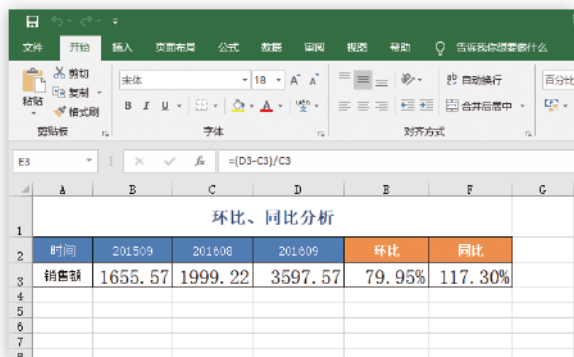
在 Excel 中，可以运用 VLOOKUP 函数和 IF 函数进行数据的快速分组。VLOOKUP 函数嵌套层数无限制，可一步到位。IF 函数在 Excel 2010 版本中最多可以嵌套 64 层。



环比、同比分析					
时间	201509	201608	201609	环比	同比
销售额	1655.57	1999.22	3597.57	79.95%	

图 4.2.10 数据分析操作2

第三步，选中同比数据单元格（F3），输入计算公式： $= (D3-B3)/B3*100\%$ ，设置F3单元格的数字格式为百分比，如图4.2.11所示。



环比、同比分析					
时间	201509	201608	201609	环比	同比
销售额	1655.57	1999.22	3597.57	79.95%	117.30%

图 4.2.11 数据分析操作3



### ※ 活动4 分析成交量与商品价格间的关系

促销期间，佳玮店铺一些商品的成交量增加了，她想知道出现这样的情况是不是与商品的价格有关系。我们需要用哪种方法来弄清楚二者之间的关系呢？如何利用数据分析工具实现目标？（可以参考教科书配套资源“Goods.xlsx”文件中的数据）

#### ● 相关分析法

相关分析法用于分析某两个变量之间的关系，在这种关系中，变量之间存在着不确定、不严格的依存关系，对于变量的某个数值，可以有另一个变量的若干数值与之对应，这若干数值围绕着它们的平均数呈现出有规律的波动。例如，地区与销量之间、商品的价格与消费者需求之间，就存在着这样的相关关系。

两个变量之间的相关程度通过相关系数来表示，相关系数的值 $r$ 在-1和1之间，可以是此范围内的任何值。正相关时， $r$ 值在0和1之间，负相关时， $r$ 值在-1和0之间， $r$ 的绝对值越接近1，两变量间的关联程度越强， $r$ 的绝对值越接近0，两变量间的关联程度越弱。

#### ● 分析相关性

在Excel中，可以通过函数来分析相关性。

第一步，从数据库中导出商品销售表，保留名称、价格和销售量三列数据，如图4.2.12所示。

	A	B	C
	宝贝名称	宝贝价格	30天销售量
1	晨光文具中性笔0.35签字笔 黑色水笔学生办公文具24支装 包邮	24.9	1549
2	晨光文具中性笔芯 水笔芯 替芯通用笔芯0.5mm/0.7mm包邮	9.9	725
3	晨光文具中性笔0.5可擦动笔字笔会议笔创意水笔 12支包邮K35	32.3	406
4	晨光文具电脑涂卡铅笔2B自动铅笔考试推荐涂卡AMP93701	1.65	322
5	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	30	317
6	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	35	294
7	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	12.9	235
8	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	1.67	219
9	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	30	212
10	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	34.9	199
11	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	8.1	187
12	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	25.92	186
13	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	12.54	184
14	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	12.98	180
15	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	3	180
16	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	1.5	177
17	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	5.33	177
18	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	45	177
19	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	33.7	176
20	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	2.7	176
21	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	25	174
22	晨光文具中性笔0.5mm水笔办公学生优品系列10支/盒包邮AGPB2001	38	173

图 4.2.12 分析相关性操作1

第二步，选中某单元格（D2），在“公式”选项卡中找到统计函数

CORREL，如图4.2.13所示。

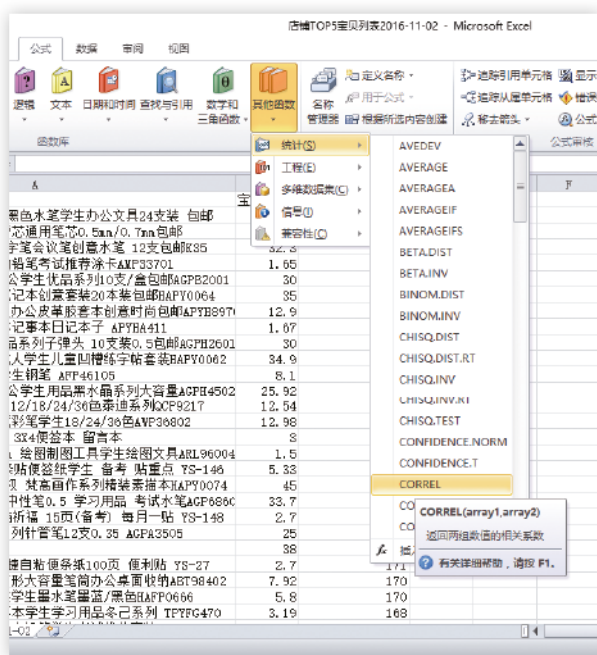


图 4.2.13 分析相关性操作2

第三步，在弹出的对话框中选择需要进行相关分析的两列数据（B2:B6,C2:C6），如图4.2.14所示，单击“确定”按钮后，会在被选中的单元格显示相关系数。

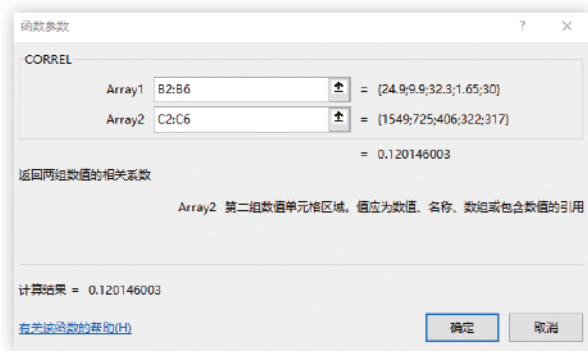


图 4.2.14 分析相关性操作3

**CORREL 函数**  
只适用于线性相关的情况下，在非线性相关情况下并不适用。如果要快速得出结果，并且是在多个维度的情况下，可以用 Excel 内置的数据分析的相关系数来得出相关矩阵。



## 拓展练习

经历上述活动后，请同学们进行小组合作，生成一份网店在开学季促销期间的销售数据表，表中要体现出促销期间的日常数据以及促销期间的总体销售情况。

## 4.3 数据可视化

进行数据分析，除了要将数据背后的“故事”挖掘出来，还要将数据包装起来。这样，既能让阅读者快速理解数据所要表达的意义，又能让阅读者有良好的视觉体验。本节将通过完成“生成网店销售业绩报表”和“撰写网店销售业绩分析报告”两个任务，让大家看到数据的美，并通过具体的活动，体验数据可视化的过程与方法。



### 学习目标

- ★ 了解常用的数据可视化方法。
- ★ 体验数据呈现与解释的过程。
- ★ 初步形成数据可视化思想。

电商平台的开学季促销结束了，佳玮的网店的交易额颇为可观。对此，有数据为证。但是，她发现枯燥乏味的数据查阅起来很费劲，从而易忽略很多重要的信息。下面我们学习如何让销售数据更直观、表达的信息更丰富，特别是要分析、比较本次促销活动与日常销售的数据，从而体会到促销的作用。



### 任务一 生成网店销售业绩报表

#### ※ 活动1 观察某店铺销售业绩的呈现方式

销售业绩就是数据分析的结果，如何将这些数据分析结果通过报表一目了然地呈现出来，需要采取一定的手段。请登录知己知彼网站，选择某店铺，重点查看“每日概况”和“销售结构分析”两个模块，思考该店铺的销售业绩是采用什么方法呈现给大家的，在表4.3.1中相应的方法后画“√”。



报表是统计数据的一种手段，通常以表格、图表等形式形成统计和显示数据。

表 4.3.1 某店铺应用的数据可视化方法

数据可视化方法	应用结果
面积和尺寸可视化	
颜色可视化	
图形可视化	
地域空间可视化	

## ● 数据可视化

所谓数据可视化，就是根据数值用标尺、颜色、位置等各种视觉信息的组合来表现数据。数据可视化是一个抽象的过程，而数据可视化的结果就是将数据转化为能够帮助人们理解数据的、具有可读性的图。数据可视化的目的是“让数据说话”。

## ● 数据可视化方法

常用的数据可视化方法有以下几种。

**面积和尺寸可视化：**用长度、高度或面积来表达数据之间的对比，让读者对数据一目了然。

**颜色可视化：**通过颜色的深浅来表达数据的强弱和大小，使读者明确哪些数据是突出显示的。

**图形可视化：**使用具有一定含义的图形来呈现数据，使数据更加生动，更便于读者理解数据所要表达的意义。

**地域空间可视化：**用地图作为大背景，把数据放在地理坐标系中，使读者直观地了解整体的数据情况以及某一区域的具体数据。

### ※ 活动2 为报表中的数据选择图表类型

图表更能表达数据分析的结果，佳玮也想利用图表，让报表中的数据更加直观。

## ● 图表

图表可以帮助人们理解抽象的数据。在各类报告和说明性文件中，用图表展现数据显得简洁、可靠。图表的类型多种多样，除了传统的柱形图、饼图、折线图等，还有气泡图、蛛网图、散点图等图形。

## ● 选择图表类型的方法

如果是为了观测、跟踪数据，就要强调数据的实时性和可运算等，此时可以生成一份可实时变化、可读性强的图表。

如果是为了强调数据的呈现度，可以生成可以进行检索的交互式图表。

如果是为了发现数据之间的潜在关联，就可以生成分布式的多维图表。

如果为了帮助用户快速理解数据的含义或变化，则可以利用漂亮的颜色、动画创建生动、明了、具有吸引力的图表。

表4.3.2所列举的是一部分具体的图表类型，以及它们的适用情境。

表 4.3.2 图表的类型与适用情境

图表的适用情境	饼图	柱形图	条形图	折线图	气泡图	蛛网图	散点图
整体—部分	√	√	√				
数据间比较		√	√		√		
走势、趋势		√		√			
数据频次		√	√	√			
数据间关系		√	√		√		√
多重数据比较						√	

请同学们根据上一节销售业绩分析的结果，为佳玮的网店报表选择适合的图表，用来呈现分析结果。请在你选择的图表下面画“√”，并说明用途。

表 4.3.3 销售业绩图表选择结果

	饼图	柱形图	条形图	折线图	气泡图	蛛网图	散点图
你的选择							
用途							

### ※ 活动3 动手制作可视化图表

在制作可视化图表之前，佳玮要选择一个数据可视化工具，以帮助她呈现这些数据。

## ● 数据可视化工具

数据可视化工具多种多样，有基础性的入门级工具，也有专业的数据可视化工具，其中基础性的工具有Excel，专业性的工具有Tableau、水晶易表等。

Excel：是一个被广泛使用的电子表格软件，它可以通过图表工具简单方便地插入各种图表，从而完成基础性的数据可视化工作，如图4.3.1所示。

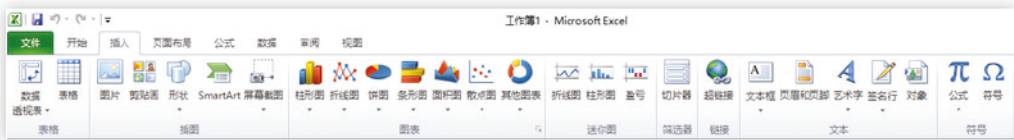


图4.3.1 Excel的图表工具

水晶易表：是一个具有交互性的可视化工具，可以呈现动态的、可交互的可视化分析结果，图4.3.2所示是由水晶易表生成的可视化图表。



图4.3.2 水晶易表的应用案例

Tableau：是一个基于可视化界面的数据可视化工具，对于想进行深入的数据分析而不想编程的人来说，这款软件非常适用。利用该软件，可以方便地与数据进行交互。Tableau Public是免费的，可以将各种图表在线发布，图4.3.3所示是发布后的Tableau Public交互图。



图 4.3.3 Tableau Public交互图

### ● 制作图表

在Tableau Public中制作商品销量和销售额图表，需要经历以下操作过程。

第一步，选择Excel连接数据源，如图4.3.4所示。

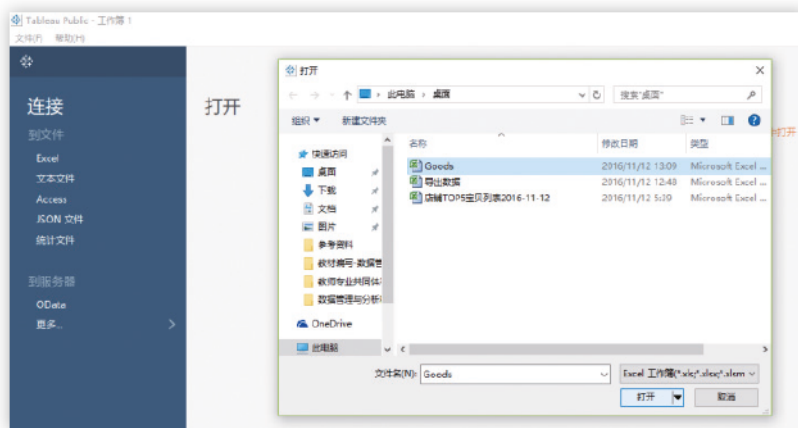


图 4.3.4 制作图表操作1

第二步，生成Tableau工作表。

第三步，将度量和维度数据分别添加到工作表中，如图4.3.5所示。

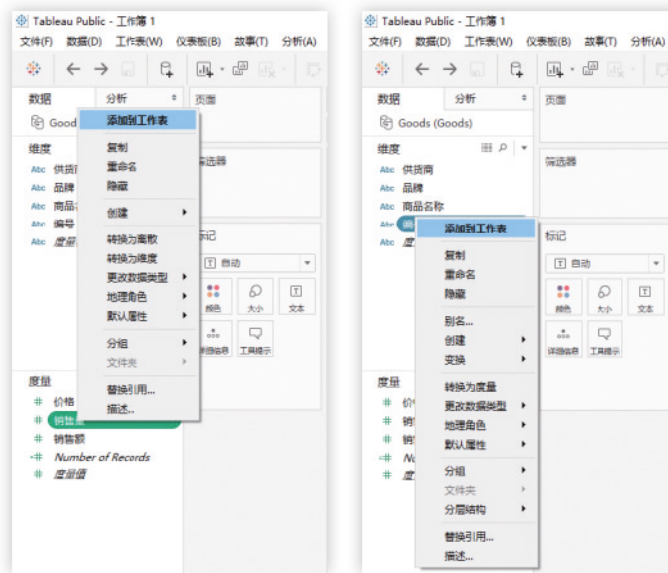


图 4.3.5 制作图表操作2

第四步，选择图形，生成图表，如图4.3.6所示。

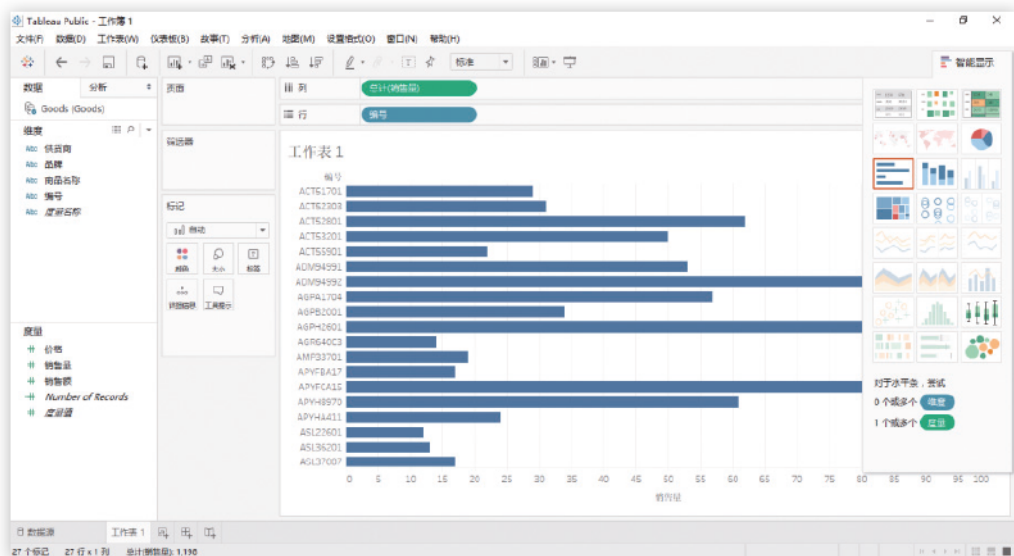


图 4.3.6 制作图表操作3

佳玮选择使用Tableau Public来制作可视化图表，你的选择是什么呢？请进行小组合作，应用离线数据库导出的数据对同比、环比、成交量与商品价格间的关系涉及的数据进行可视化呈现，并生成一份网店在开学季促销期间的销售业绩报表。（可以参考教科书配套资源“Goods.xlsx”文件中的数据）



## 任务二 撰写网店销售业绩分析报告

### ※ 活动1 明确报告撰写的思路

报告是对销售数据的整理与解释，对于如何整理和解释佳玮网店的销售数据，我们可以展开头脑风暴，充分发挥集体智慧。各小组按照下面的要求进行讨论，并记录主要观点。

讨论主题：网店销售数据报告的撰写

形式：分组讨论会

时间：15分钟

主持人：1名，主持人只主持会议，对设想不做评论

### ● 数据分析报告

数据分析报告是根据数据分析原理与方法，运用数据来反映、研究和分析某项事务的现状、问题、原因、本质和规律，并得出结论，提出解决办法的一种分析应用文本。这种文本是决策者认识事务、了解事务、掌握信息、搜集相关信息的主要工具之一，数据分析报告通过对事



务数据全方位的科学分析来评估其环境及发展情况，为决策者提供科学、严谨的依据，降低风险。

经过头脑风暴，针对报告撰写的思路，各小组基本形成了共识。在撰写数据分析报告时，要重点思考以下几个方面的问题：  
发生了什么？  
问题出在哪里？  
为什么这件事情会发生？  
需要采取什么行动？  
下一步将发生什么？  
可能发生的最好结果是什么？

### ※ 活动2 撰写数据分析报告

拿到佳玮网店的销售数据并得出分析结果后，在动手撰写数据分析报告之前，需要先确定数据分析报告的基本结构。

为了更好地完成数据分析报告，请大家帮佳玮撰写一份数据分析报告的提纲，在撰写提纲时，可以参考给出的数据分析报告模板的格式和内容。

经典的数据分析报告的结构是“总一分一总”结构，主要包括开篇、正文和结尾三大部分。

## 标 题

### 前言

主要包括分析背景、目的及思路三部分。  
为何要开展此次分析？有何意义？  
通过此次分析要解决什么问题？达到何种目的？  
如何开展此次分析？通过哪几方面开展？

### 正文

此为数据分析报告的核心部分，它将系统全面地表述数据分析的过程与结果。  
包含所有数据分析事实和观点。  
通过数据图表和相关文字综合分析。  
各部分具有逻辑关系。

### 结论与建议

结论是以数据分析结果为依据得出的，以综述性文字

进行说明。

建议是根据结论对业务所面临的问题提出改进方法。

### 附录

提供正文中涉及而未阐述的资料，是数据分析报告的补充。

## ※ 活动3 展示数据分析报告

要想向大家展示自己的数据分析报告，我们应该选择合适的呈现形式和方式。

### ● 数据分析报告的形式

因对象、内容、时间、方法等的不同，存在着不同形式的数据分析报告，需要根据实际情况来确定报告的形式。

专题分析报告是指对社会经济现象的某一方面或某一问题进行专门研究的一种数据分析报告，它为制订某项政策、解决某个问题提供参考或依据。

综合分析报告是指全面评价一个地区、单位、部门业务或其他方面发展情况的一种数据分析报告，如世界人口发展报告、企业运营分析报告等。

日常数据通报是指以定期数据分析报表为依据，反映计划执行情况，并分析其影响和成因的一种数据分析报告，它按日、月、年等时间段定期进行分析。

### ● 数据分析报告的呈现方式

人们习惯使用Word、Excel和PowerPoint等办公软件呈现数据分析报告。这三款软件各有其优势、劣势和适用范围，如表4.3.4所示。

表 4.3.4 办公软件在数据分析报告中的应用

软件名称	优势	劣势	适用范围
Word	<ul style="list-style-type: none"> <li>◆ 易于排版</li> <li>◆ 可打印成册</li> </ul>	<ul style="list-style-type: none"> <li>◆ 缺乏交互性</li> <li>◆ 不适于汇报</li> </ul>	<ul style="list-style-type: none"> <li>◆ 综合分析报告</li> <li>◆ 专题分析报告</li> <li>◆ 日常数据通报</li> </ul>
Excel	<ul style="list-style-type: none"> <li>◆ 含有动态图表</li> <li>◆ 结果实时更新</li> <li>◆ 交互性强</li> </ul>	<ul style="list-style-type: none"> <li>◆ 不适于汇报</li> </ul>	
PowerPoint	<ul style="list-style-type: none"> <li>◆ 元素丰富</li> <li>◆ 可演示汇报</li> <li>◆ 增强展示效果</li> </ul>	<ul style="list-style-type: none"> <li>◆ 不适合大篇文字</li> </ul>	<ul style="list-style-type: none"> <li>◆ 综合分析报告</li> <li>◆ 专题分析报告</li> <li>◆ 日常数据通报</li> </ul>

根据对销售业绩的分析结果，请你来选择报告的形式与呈现方式，在表4.3.5中相应的选项后面画“√”。

表 4.3.5 数据分析报告的形式与呈现方式选择结果

报告的形式	你的选择	呈现方式	你的选择
专题分析报告		Word	
综合分析报告		Excel	
日常数据通报		PowerPoint	



### 拓展练习

经过数据分析与图表制作，佳玮的网店的销售业绩结果已经整理出来了。请撰写一份销售业绩分析报告，并在课堂上向大家展示和汇报。



## 单元学习评价

为了满足网店未来发展的需要，佳玮准备招聘一位数据分析师。下面是她为应聘者准备的部分笔试题目，请根据本单元的学习体验，通过互评或自评来判断自己是否胜任数据分析师。

1. 如果进行数据分析，你会选择使用哪个工具？

- 在线数据分析工具     Python     Excel     Stata  
 SPSS     SAS     MATLAB     R

2. 请列举几种常用的数据分析方法，并描述这些方法适用的情境。

方法1: \_\_\_\_\_ 适用情境: \_\_\_\_\_

方法2: \_\_\_\_\_ 适用情境: \_\_\_\_\_

方法3: \_\_\_\_\_ 适用情境: \_\_\_\_\_

方法4: \_\_\_\_\_ 适用情境: \_\_\_\_\_

3. 如果想呈现网店经营数据的销售情况以及变化趋势，你会选择以下哪种图表？

- 饼图     气泡图     柱形图     蛛网图  
 条形图     散点图     折线图

4. 请针对任意一种数据分析报告呈现方式，说明其优点、缺点及自己的使用感受。

数据分析报告呈现方式: \_\_\_\_\_

优点: \_\_\_\_\_

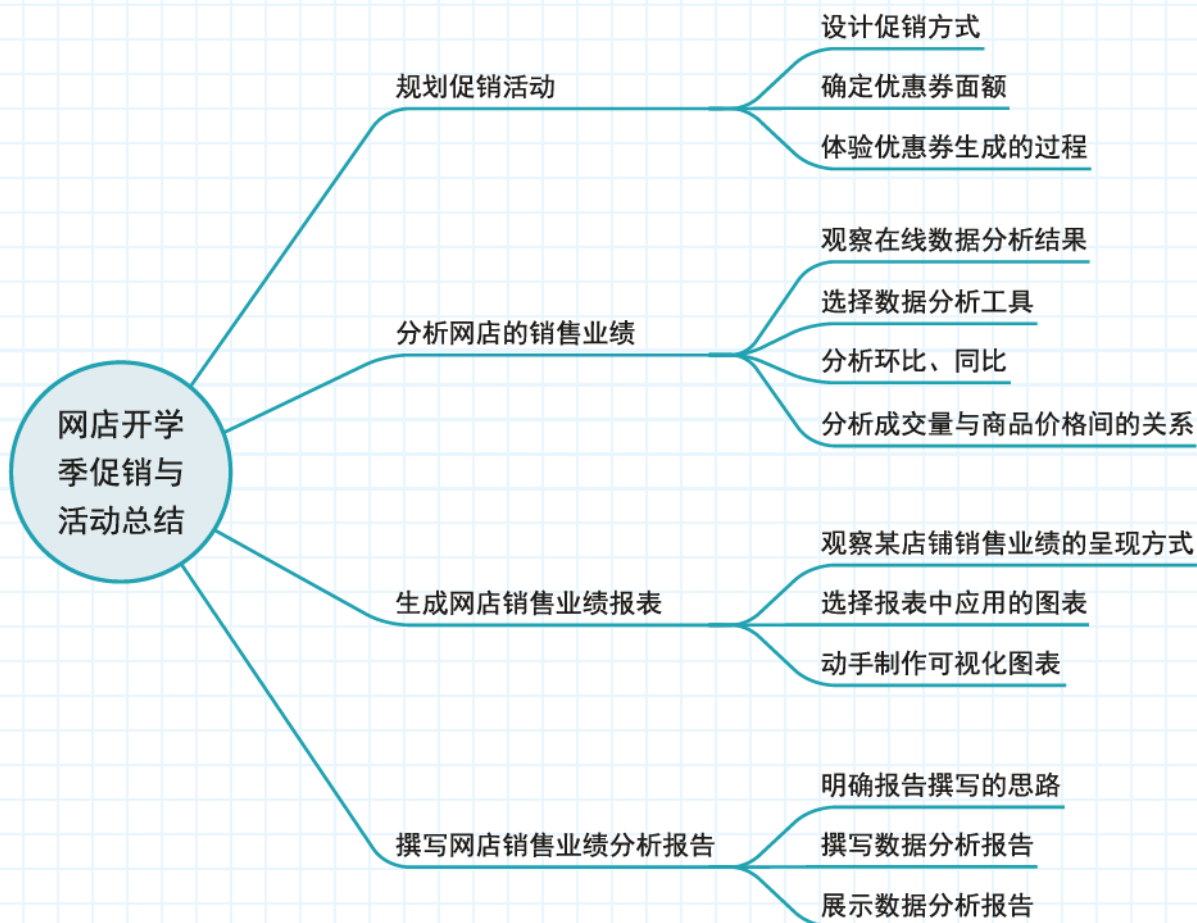
缺点: \_\_\_\_\_

使用感受: \_\_\_\_\_

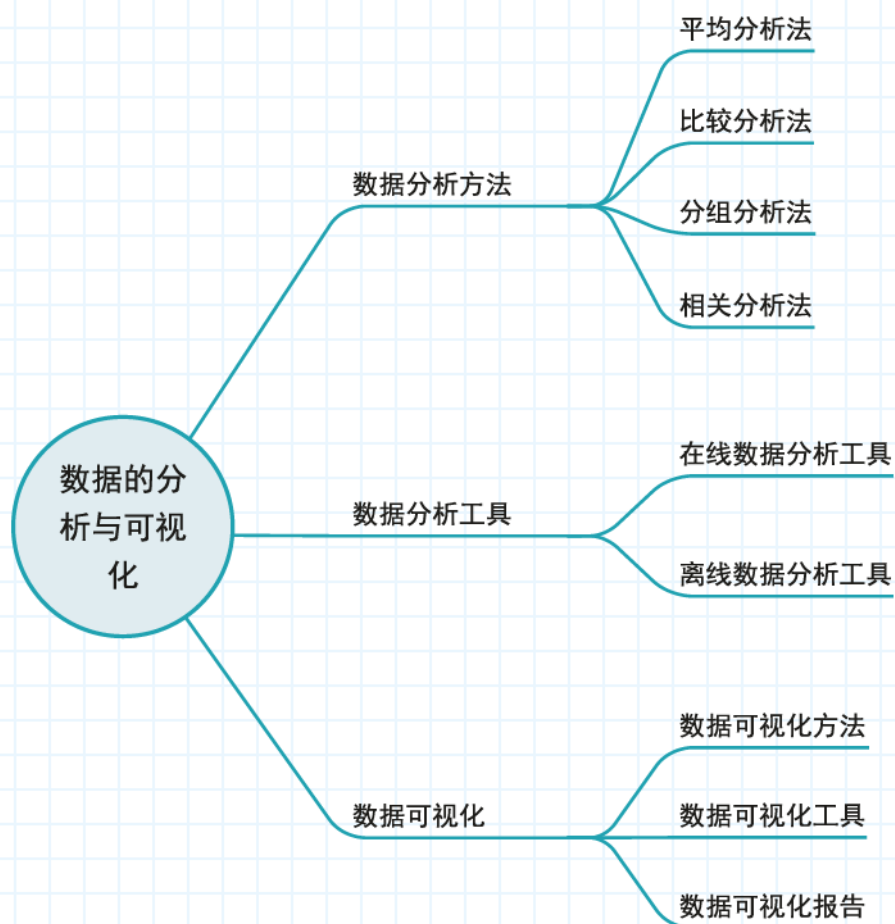
## 单元学习总结

本单元围绕“网店开学季促销与活动总结”项目展开学习。通过依次完成规划促销活动、分析网店的销售业绩、生成网店销售业绩报表和撰写网店销售业绩分析报告等任务，开展相应的主体活动，掌握数据分析方法，体验数据分析工具的使用，经历数据分析与可视化的过程，形成数据分析与可视化的思想。

### “项目—任务—活动”结构图



## 知识地图



## 第 5 单元 数据安全

一些大型企业的财务系统或商务系统通常已经运行了若干年，积累了大量至关重要的财务、客户和交易往来等数据，这些数据对于企业自身和竞争对手来说都是价值连城的。任何一个系统在运行过程中，都会受到来自各个方面的数据安全威胁，如果不及时处理，那些价值连城的数据就会有丢失、泄露的风险，继而影响整个系统的运行与后续发展。

在本单元中，我们将通过“网店数据的安全保障”项目展开学习，认识数据丢失与泄露的风险，掌握数据备份与还原的过程，形成一定的数据安全意识与能力。

要保障网店的数据安全，需要考虑一系列的问题：网店的数据安全威胁来自哪里？我们能发现数据所遭受的破坏吗？我们能够采取什么办法来降低数据安全的风险？如果数据遭受破坏，应该如何应对？为解决这些问题，我们需要完成以下两个任务：

- ◆ 查找数据安全威胁；
- ◆ 强化网店数据安全措施。

## 5.1 数据丢失与泄露的风险

如果忽视数据安全问题，数据丢失与泄露的风险就会加大，我们需要时刻防范这些风险，找到降低风险的方法。本节将通过完成“查找数据安全威胁”的任务，让大家认识数据安全问题的重要性，掌握降低数据丢失与泄露风险的办法。



### 学习目标

- ★ 认识数据安全问题及其可能产生的后果。
- ★ 掌握降低数据安全风险的基本方法。
- ★ 初步形成数据安全防范意识。

某日，佳玮的网店所在的某电商平台瘫痪了，从事故发生至恢复长达近 12 小时，在此期间，网店不能进行正常的经营活动，导致店铺的经营蒙受损失。然而，更让人忧心的事情也随之而来，信息安全部主管告诉佳玮，存放着用户数据、交易数据、库存信息等的后台数据库很可能在这次事故中遭受到了来源未知的安全威胁。佳玮想知道，数据安全威胁究竟是什么？它来自哪里？为何会有如此大的“杀伤力”？



### 任务 查找数据安全威胁

#### ※ 活动1 收集网店数据安全问题

一般情况下，电商平台事故最容易受到影响的就是后台数据库，后台数据库是否安全直接决定着前台数据是否会出现异常。佳玮想知道电商平台发生事故期间自己网店前台的各类数据是否出现了异常，以及这些异常现象属于哪类数据安全问题。



## ● 数据安全问题

数据共享是数据库的主要特点之一，而数据共享必然会带来数据的安全性问题。数据的安全问题主要有两类：一类是计算机系统运行安全；一类是数据库的数据安全。

## ● 数据安全问题的后果

如果是计算机系统运行安全问题，网络非法用户可以通过网络手段破坏计算机操作系统的正常运行，甚至导致计算机系统瘫痪。

一旦出现数据库的数据安全问题，网络非法用户则会通过木马、恶意程序等入侵数据库，对数据进行篡改、损坏，或盗取数据。

篡改数据指对数据库中的数据进行增加或删除，从而破坏数据。

损坏数据将导致无法读取数据库中的数据。

窃取数据指在未获得数据库管理权限的情况下，获得数据库中的数据。

佳玮将发生事故期间网店出现的问题进行整理汇总，发现了一些异常情况，请大家帮佳玮分析这些异常情况属于哪种数据安全问题，并填写表5.1.1。

表 5.1.1 数据安全问题情况汇总

异常情况	所属数据安全问题类别	后果
电商瘫痪的同时，无法登录网店		
登录网店后，其主页换成了卖其他产品的网页		
客户的联系方式丢失		
有3种商品还有库存，在前台却显示为已下架	篡改数据	破坏数据

### ※ 活动2 讨论网店数据安全面临的威胁

对于此次电商平台事故，佳玮感觉到了事态的严重性，但却不知道原因是什么，她很想找到“幕后黑手”。面对此次数据安全问题，下面我们来帮佳玮分析网店数据可能会面临的安全威胁。请同学们分组讨论，并分享讨论结果。

## ● 数据安全面临的威胁

### 硬件故障

一是硬盘驱动器损坏。通常情况下，如果在高温或者机械损伤等情况下，硬盘容易出现故障，里面所存储的数据会因为硬盘故障而丢失。

二是电源故障。电源作为计算机的动力来源，经常会因为电压不稳或者功率过大而导致硬盘损坏，进而会损坏硬盘或存储设备中的数据。频繁电源故障同样也严重影响应用系统的正常使用。

### 软件方面的威胁

一是病毒的影响。计算机病毒会造成计算机资源的损失和破坏，从而导致资源和财富的巨大浪费，甚至是社会性的灾难。随着信息化社会的发展，计算机病毒的威胁日益严重，反病毒的任务也更加艰巨。

二是操作系统的缺陷。数据库的安全和操作系统是否完善有很大的关系，若操作系统有重大漏洞，则很容易遭受攻击，因而数据库的安全也就不能得到保证。目前常用的操作系统存在的漏洞非常多，这也给很多别有用心之人以可乘之机。

三是数据库系统本身的安全漏洞。大多数用户在使用数据库系统的时候并没有考虑该软件自身的安全性，如果软件自身的安全存在问题，就容易出现数据泄露甚至崩溃的危险。

### 人为因素

任何数据库系统都是由人来操作的，所以人为因素是影响数据库安全的重要因素。最常见的此类安全隐患有两种：一种是内部安全，即员工的操作失误或者有意为之；另一种是来自外部的人为因素影响，如最常见的黑客行为。黑客利用工具对系统进行破解或者攻击，会对数据库安全造成重大威胁。



**数据库管理员**  
(Database Administrator, DBA), 是从事管理和维护数据库管理系统的相关工作人员的统称, 属于运维工程师的一个分支, 主要负责业务数据库从设计、测试到部署交付的全生命周期管理。数据库管理员的核心目标是保证数据库管理系统的稳定性、安全性、完整性和高性能。

佳玮网店的数据可能面临的威胁来自哪里？你做出判断的理由是什么？

---



---



---



---



---

### ※ 活动3 探究降低网店数据安全风险的办

面对来自多方面的数据安全威胁，佳玮很想知道电商平台会有什么对策。在之前的学习中，我们已经模拟构建了电商平台离线数据库。如果让你来负责后台数据库的管理，那么这个数据库的安全如何维护？有什么办法能够降低数据安全的风险呢？

#### ● 数据库的安全性

数据库的安全性是指保护数据库中的数据，防止不合法的使用所造成的数据泄露、更改或破坏。保护数据库安全，就是通过种种防范措施以防止用户越权使用数据库中的数据。

#### ● 数据库安全保护措施

数据库系统中一般采用用户标识和鉴别、存取控制机制、视图机制、审计以及数据加密等技术来保护数据库的安全。

##### 用户标识和鉴别

用户标识和鉴别是为数据库提供的最外层保护措施。用户每次登录数据库时都要输入用户标识，经系统核对后，合法的用户会获得进入系统最外层的权限。

##### 存取控制机制

存取控制机制是数据库安全的一个重要保证，它能够确保具有数据库使用权限的用户访问数据库并进行权限范围内的操作，同时令未被授权的用户无法接近数据。

##### 视图机制

在数据库中，可以为不同的用户定义不同的视图，通过视图机制把要保密的数据对无权操作的用户隐藏起来，从而自动地对数据提供一定程度的安全保护。视图机制使系统具有数据安全性、数据逻辑独立性和操作简便等优点。

##### 审计

审计功能就是把用户对数据库的所有操作自动记录下来放入审计日志（Audit Log）中，数据一旦被非法存取，数据库管理员可以利用审计跟踪的信息，重现导致数据库现有状况的一系列事件，找出非法存取数据的人、时间和内容等。由于任何系统的安全保护措施都不可能无懈可击，蓄意盗窃、破坏数据的人总是想方设法打破控制，因此审计功能在维护数据安全方面是非常有效的。

## 数据加密

对高度敏感的数据，如财务、军事、国家机密等数据，除了以上安全性措施外，还应该采用数据加密技术。数据加密是防止数据在存储和传输过程中失密的有效手段。加密的基本思想是根据一定的算法将原始数据变换为不可直接识别的格式，从而使得不知道解密算法的人无法获得数据的内容。

以上各项措施中会应用到哪些具体的方法？请分组探究，并把探究的结果汇总到表5.1.2中。

表 5.1.2 降低数据丢失与泄露风险的具体方法

分组	措施	具体方法
小组1	用户标识和鉴别	
小组2	存取控制机制	
小组3	视图机制	
小组4	审计	
小组5	数据加密	

### ※ 活动4 离线数据库的授权与存取控制

在之前的学习中，我们分小组构建了SQL Server离线数据库，而且小组中的成员都是这个数据库的用户，他们对数据库中的各个表都拥有查询、修改、插入和删除等全部权限。但是在数据库运行过程中，经常会出现几个人同时对一个表进行操作的情况，这样就会导致数据库中的数据不安全，为了保障数据库的安全性，最常用的办法就是控制用户的权限。

## ● 用户权限

用户权限主要包括数据对象和操作类型两个要素。数据对象是数据库的组成部分，包括数据表、字段、视图等几种；操作是将新记录提交至数据库、将记录从一个状态变迁到另一个状态的机制，具体的操作类型包括添加、删除、修改等。表5.1.3列出了不同类型的数据对象的操作类型。

表 5.1.3 数据对象的类型和操作类型

数据对象	操作类型
数据表	ALTER, INDEX, SELECT, INSERT, UPDATE, DELETE, ALL PRIVILEGES
字段	SELECT, INSERT, UPDATE, DELETE, ALL PRIVILEGES
视图	SELECT, INSERT, UPDATE, DELETE, ALL PRIVILEGES

## ● 定义权限

定义用户的存取权限称为授权，通过授权能够规定用户可以对哪些数据进行什么样的操作，也可以通过收权，将用户原来的权限收回。

## ● SQL Server 授权和收权语句

### 授权语句（GRANT）

格式：

GRANT <权限1> [,<权限2>,…]

[ON <对象类型> <对象名>]

TO <用户1> [,<用户2>,…]

[WITH GRANT OPTION];

功能：将指定数据对象的指定权限授予指定的用户。

### 收权语句（REVOKE）

格式：

REVOKE <权限1> [,<权限2>,…]

[ON <对象类型> <对象名>]

FROM <用户1> [,<用户2>,…];

功能：把已经授予指定用户的指定权限收回。

在已构建的离线数据库OfflineDatabase中有Goods、Supplier和Customer三个表，某组组长想对组员的用户权限做一些变更，便在数据库中进行了如下三个操作：

```
REVOKE UPDATE(customerID)
ON TABLE Customer
FROM user3;
GRANT SELECT ON TABLE Goods TO PUBLIC;
GRANT CREATETAB ON DATABASE OfflineDatabase TO user1;
```



CREATE TAB :

允许拥有者在数据库中创建新表。

请大家分析这三个操作让数据库用户的哪些权限发生了变化？



### 拓展练习

某数据库中有两个表，它们的关系模式分别如下：

职工（职工编号，姓名，年龄，职务，工资，部门编号）

部门（部门编号，名称，经理名，地址，电话号码）

请用 SQL Server 的 GRANT 和 REVOKE 语句，完成以下授权定义或存取控制功能：

1. 用户王明对两个表有SELECT权限；
2. 用户李勇对两个表有INSERT和DELETE权限；
3. 用户刘星对职工表有SELECT权限，对“工资”字段具有更新权限；
4. 用户张玲具有对两个表的所有权限，并具有给其他用户授权的权限；
5. 用户杨兰具有从每个部门的职工中SELECT工资的权限。

## 5.2 数据备份与恢复

数据一旦遭到破坏，损失将不可避免。如果能够备份数据，并在关键时刻恢复数据，就可以将损失降到最低限度。

### 学习目标

- ★ 了解数据备份与恢复技术。
- ★ 学会制订数据备份策略，掌握数据备份与恢复的基本方法。
- ★ 初步形成数据保护意识。

在经历了这次电商平台事故之后，佳玮跟电商平台沟通，希望能够将网店的数据恢复到事故之前的状态。一天之后，她的网店数据恢复正常了，她非常好奇电商平台的相关工作人员是如何做到的。

### 任务 强化网店数据安全措施

#### ※ 活动1 讨论数据安全技术与策略

为了将数据安全的风险降到最低，电商平台的数据库管理员在数据遭受破坏之前就要想好万全之策。除了在数据库使用前就采取安全防范措施外，在数据库使用过程中应如何加强安全防范工作呢？

#### ● 数据备份

数据备份是指为防止系统出现操作失误或因系统故障导致数据丢失，而将全部或部分数据从一个存储介质复制到其他存储介质上的过程。

#### ● 数据备份的目的

- 一是重要业务资料的自动化存储管理。
- 二是简化备份的复杂性，节省人力资源，提高工作效率。

三是确保数据存放安全并被有效管理。

四是避免人为错误及自然灾害的破坏，提高数据资料的正确性。

五是快速、准确地进行数据恢复，提高服务品质并提升形象。

## ● 数据库备份

数据库备份是数据备份的重要组成部分，指的是制作数据库结构、对象和数据的副本，以便在数据库遭到破坏的时候能够恢复数据库。

## ● 数据恢复

数据恢复也称数据还原，是指将备份到存储介质上的数据再恢复到计算机系统中，它与数据备份是一个相反的过程。数据恢复措施在整个数据安全保护中占有相当重要的地位，因为它关系到系统在经历灾难后能否迅速恢复运行。

## ● 数据库恢复

数据库恢复指的是将数据库备份加载到系统中，让数据库由故障状态转变为无故障状态的过程。

电商平台的数据安全技术同样适用于我们构建的离线数据库，为了保障数据库在任何情况下都能够安全稳定地运行，我们应该首先制订相关的数据备份策略，以保证数据的备份与恢复技术的顺利实施。

## ● 数据备份策略

数据备份策略是针对数据备份与恢复过程而建立的应对策略，根据相关的策略，可以指导数据备份与恢复的过程。在建立数据备份策略之前，要对问题进行有针对性的分析，对操作系统的资源配置、数据库系统的运行要求等进行全面的分析，选择科学的策略，以保障在数据安全事件发生以后不对数据造成任何影响，而且数据库中的文件还能够恢复。

下面给出了两条数据备份的策略，请根据离线数据库的运行现状，思考还能制订出哪些策略，将小组讨论的结果填在空白处。

1. 注重存储介质的管理。
2. 培养及时备份的习惯。



## ※ 活动2 备份离线数据库

电商平台的数据安全技术同样适用于离线数据库，并能够进一步加强离线数据库的安全保护工作。请在数据备份策略的指导下，为佳玮网店的离线数据库进行备份。

### ● 数据备份方法

常用的数据备份方法有完全备份、增量备份、差异备份和日志备份等四种。

#### 完全备份

按备份周期（如一天）对整个系统所有的文件（数据）进行备份，为数据库中的所有内容制作一个副本。

#### 增量备份

增量备份的都是已更新过的数据，每次备份的数据只是相当于上一次备份后增加的和修改过的内容。例如，系统在星期日做了一次完全备份，然后在以后的六天里每天只对当天新的或被修改过的数据进行备份。

#### 差异备份

差异备份指在完全备份后将新增加或修改过的数据进行备份。它与增量备份的区别是每次备份都把上次完全备份后更新过的数据进行备份。例如，在星期日对数据库进行完全备份后，其余六天中的每一天都将当天所有与星期日完全备份时不同的数据进行备份。

#### 日志备份

将从最近一次日志备份以来所有的事务日志备份到备份设备中。利用日志备份进行恢复时，可以指定恢复到某一时间点或某一事务。

### ● 数据备份方式

数据备份的方式一般有两种，即定时备份和实时备份。

定时备份是在固定的时间间隔进行数据备份的方式，其缺点是不能保证数据的零丢失。实时备份是在任意时间间隔进行数据备份的方式，其优点是可以保证数据的零丢失，如增量备份和差异备份使用的就是实时备份方式。

### ● 数据备份操作

当前最为常用的数据库管理系统已经提供了一系列性能优良的备份恢复工具，备份恢复操作的实施在数据库管理系统中就能实现。以SQL

Server 2016为例，对数据库 OfflineDatabase进行完全备份的过程分为以下三步。

第一步，为数据库选择“备份…”任务，如图5.2.1所示。

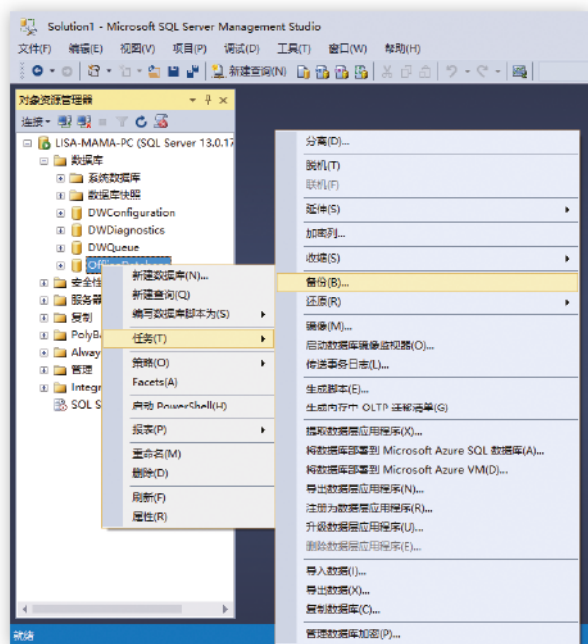


图 5.2.1 数据备份操作1

第二步，在弹出的窗口中选择“完整”备份类型，如图5.2.2所示。

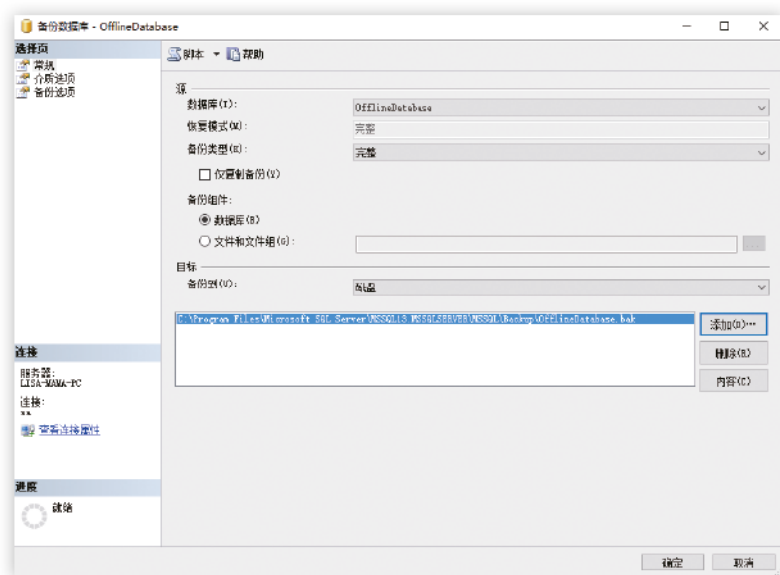


图 5.2.2 数据备份操作2

第三步，通过“添加”按钮，选择备份文件的存放路径，单击“确定”按钮，完成备份过程，如图5.2.3所示。

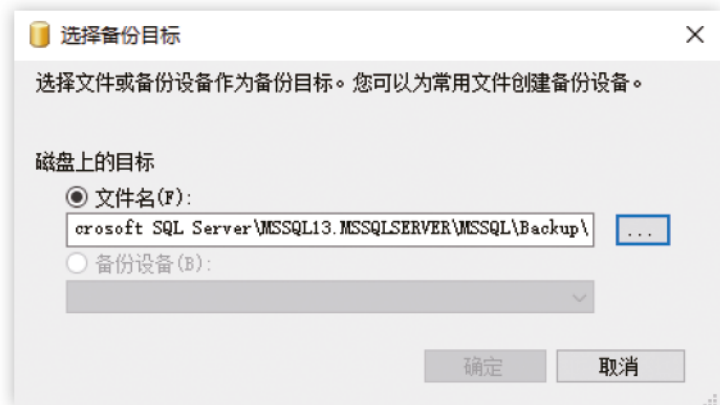


图 5.2.3 数据备份操作3

根据上述完全备份的操作过程，我们可以自己尝试进行差异备份和日志备份，两者与完全备份的操作过程一致，只是在备份类型中要进行相应的选择。

### ※ 活动3 恢复离线数据库

电商平台之所以能够在发生电商平台事故后为佳玮的网店快速地恢复了原来的数据，是采用了一定的数据恢复方法，而这种方法也正适用于我们刚刚做过备份的离线数据库。

## ● 数据恢复方法

数据恢复的基本方法有以下三种。

### 全盘恢复

将备份到存储介质上的指定系统的信息全部转储到它们原来的地方。全盘恢复一般应用于服务器发生意外灾难时导致数据全部丢失、系统崩溃或是有计划的系统升级、系统重组等情况，也称为系统恢复。

### 个别文件恢复

将个别已备份的最新版文件恢复到原来的地方。只要浏览备份数据库或目录，找到该文件，启动恢复功能，系统将自动恢复指定的文件。

### 重定向恢复

重定向恢复指将备份的文件（数据）恢复到另一个不同的位置或系统上。重定向恢复可以是整个系统恢复，也可以是个别文件恢复。

以SQL Server 2016为例，对数据库 OfflineDatabase 进行恢复的过程分为以下三步。

第一步，为数据库选择“还原—数据库”任务，如图5.2.4所示。

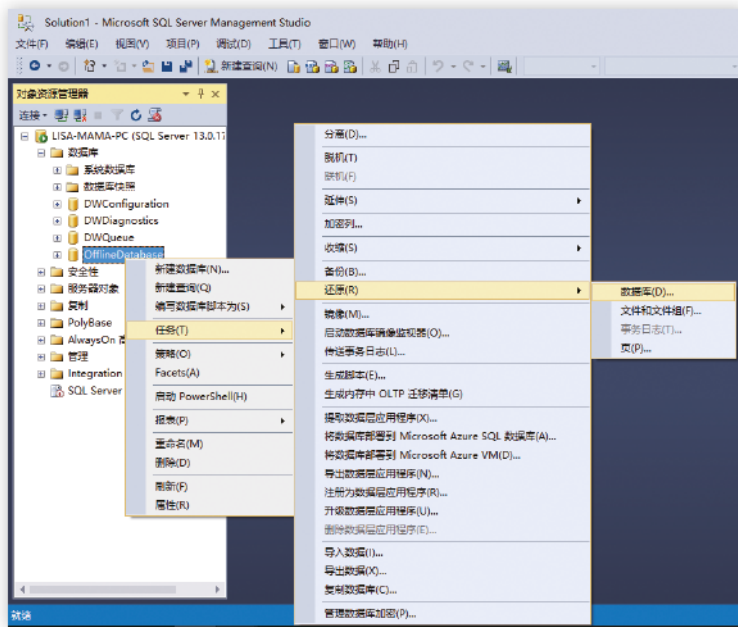


图 5.2.4 数据恢复操作1

第二步，选择“目标数据库（OfflineDatabase）”和“要还原的备份集”，如图 5.2.5 所示。

第三步，通过“选项”选择“结尾日志备份”，单击“确定”按钮，开始恢复数据，并在恢复后显示成功还原了数据库，如图 5.2.6 所示。

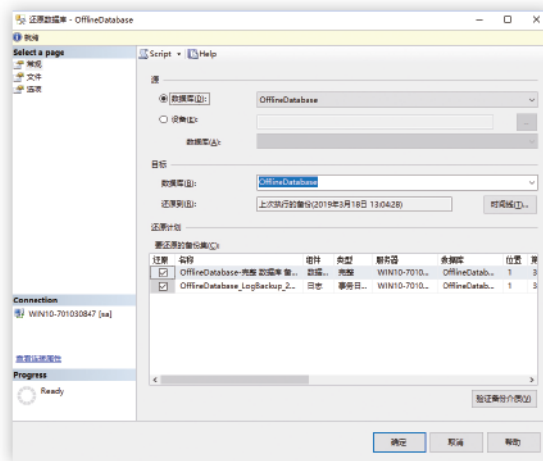


图 5.2.5 数据恢复操作2

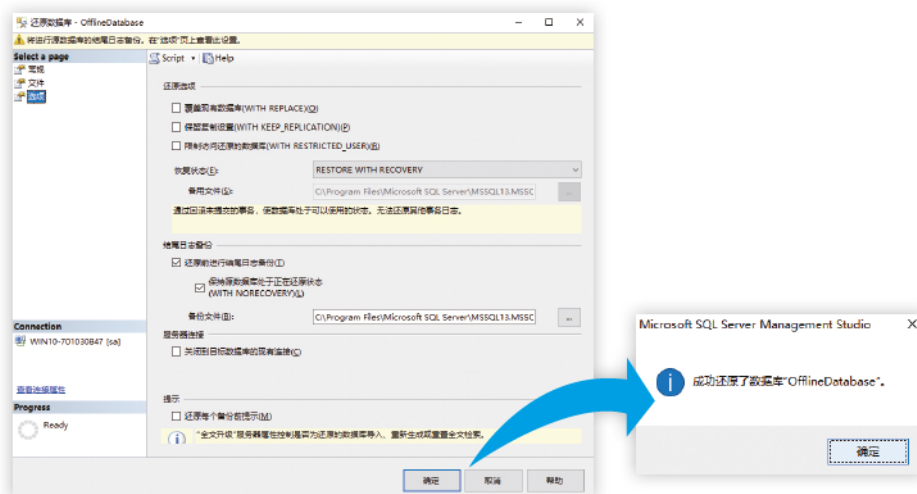


图 5.2.6 数据恢复操作3

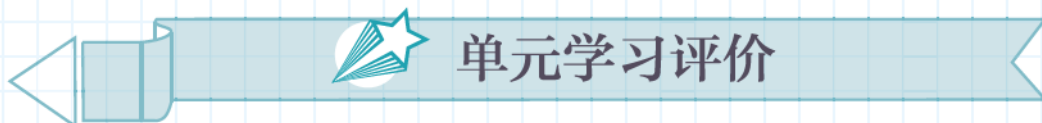




## 拓展阅读

### Oracle数据库的数据备份策略

在Oracle 数据库中，主要的备份策略分为四种：第一，注重介质的管理；第二，工作模式设成自动归档；第三，采取多个物理键；第四，培养备份的习惯。从注重介质的管理角度来看，对介质的相关管理主要表现在两个方面：第一，需要将数据库内容进行信息分类管理；第二，在介质管理的基础上能够提供更为科学合理的计划安排，防止数据再次出现丢失。从备份的主要目的角度来看，为了不影响相关数据的使用，需要将数据丢失的情况进行科学有效的处理。例如，在将数据进行存储的过程中，千万不要选择相同的计算机地址，此举既是为了防止计算机软件以及计算机硬件故障等问题，也是为了防止造成更多损失。



## 单元学习评价

风险是客观存在的，对数据安全进行风险评估可以防患于未然。请结合本单元的学习体验，检查自己是否能对佳玮的网店做风险评估。

1. 请举例说明数据安全面临哪些潜在威胁。

威胁1: \_\_\_\_\_ 举例: \_\_\_\_\_

威胁2: \_\_\_\_\_ 举例: \_\_\_\_\_

威胁3: \_\_\_\_\_ 举例: \_\_\_\_\_

2. 在管理网店的后台数据库时，需要通过相关的账号和密码来登录数据库系统。数据库系统中的Supplier表涉及网店15个供货商的数据，数据库管理员定义了15个视图，每个视图只包含一个供货商的数据，并只允许每个供货商的销售主管查询和修改其所属供货商的视图。

请根据上面的描述，说明数据库采用了哪些安全保护措施。

---

---

---

3. 网店的数据库每周日零点进行一次完全备份，每天零点进行一次差异备份，每小时进行一次日志备份。假设该网店的数据库在2018年11月21日（星期三）3:30崩溃了，作为数据库管理员，你将如何恢复数据库，并使数据损失最小？

---

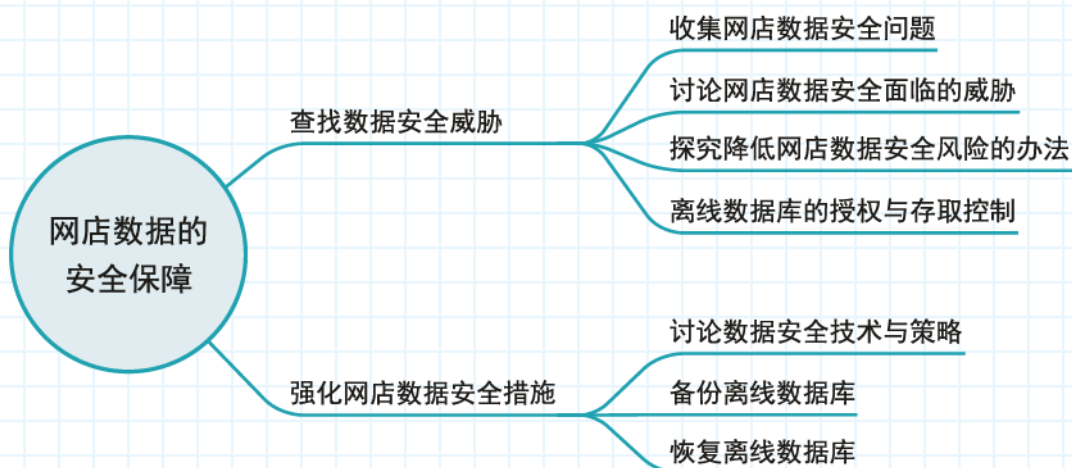
---

---

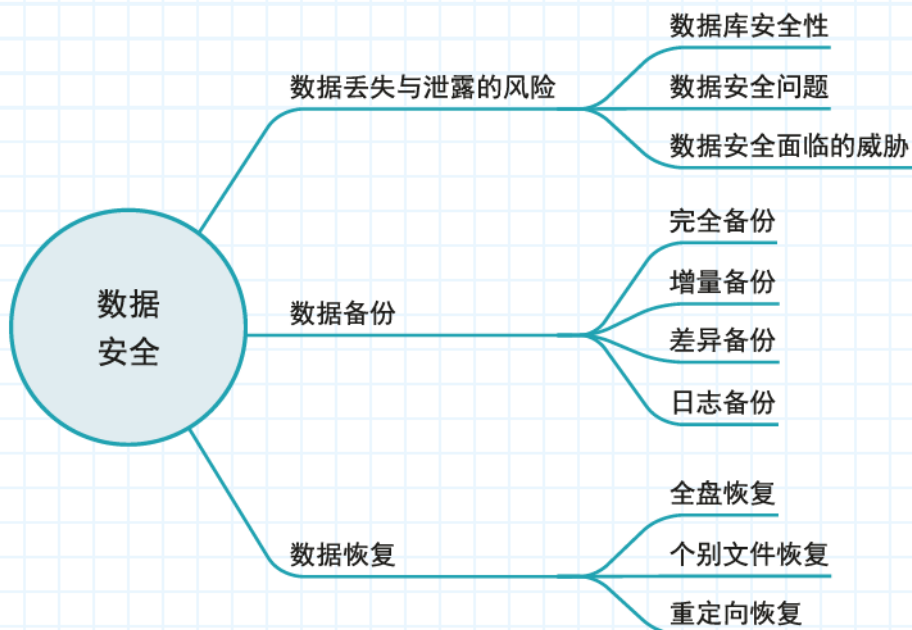
## 单元学习总结

本单元围绕“网店数据的安全保障”项目展开学习。通过依次完成查找数据安全威胁和强化网店数据安全措施等任务，开展相应的主体活动，认识数据丢失与泄露的风险，掌握数据备份与还原的过程，形成数据安全意识与能力。

### “项目—任务—活动”结构图



### 知识地图





## 第 6 单元 初识数据仓库与数据挖掘

今天的我们正沉浸在数据的海洋里，却比之前任何时候都渴望从这些海量数据中“挖掘”出隐含的、先前未知的、对我们的决策有价值的信息。我们想让数据“说话”，来帮助我们进行决策。但是，如果不借助强大的数据管理与分析技术，这些海量数据就只是简单的符号。正因为如此，数据仓库和数据挖掘才应运而生。

在本单元中，我们将通过“挖掘网店客户的潜在需求”项目展开学习，了解数据仓库的概念与基本特征，认识数据仓库在决策分析过程中的重要性，了解数据挖掘的基本概念、一般流程及其基本功能，认识数据挖掘对现实生活的重要意义。

利用数据分析技术帮助挖掘网店客户的潜在需求，需要考虑一系列问题，如原有的分析环境是否能够支持决策？利用什么样的数据分析技术来帮助决策？能在哪些方面进行决策？在实际应用中是如何体现的？为解决这些问题，我们需要完成以下三个任务：

- ◆ 选择适合决策的分析环境；
- ◆ 剖析数据挖掘的一般流程；
- ◆ 利用数据挖掘辅助决策。

## 6.1 数据库和数据仓库

随着计算机与网络技术的迅猛发展，各行各业都积累了大量的业务数据。只有经过有效的存储和科学的管理，再借助功能强大的数据分析技术，这些数据才能被转化为对企业有价值的信息。通过前面的学习我们知道，数据是被存储在数据库中的，那么数据库和数据仓库是什么关系？数据仓库是如何存储和管理数据的？本节将围绕这些问题展开学习。



### 学习目标

- ★ 了解数据仓库的概念和基本特征。
- ★ 认识数据仓库在决策分析过程中的重要性。

佳玮的网店已经小有名气了，她打算深入了解客户的潜在需求，拓展网店业务。但是，利用数据库构建的分析环境已经不能满足实际需求，数据分析师告诉佳玮，数据仓库构建的分析环境更适合她。那么，数据仓库是什么？数据仓库和数据库有什么关系？数据仓库有什么特征？这些问题困扰着她，现在让我们和佳玮一起来了解一下。



### 任务 选择适合决策的分析环境

#### ※ 活动1 了解数据仓库的由来

根据数据分析师的提示，佳玮查找了一些资料。请帮助佳玮梳理资料，并回答下面的问题。



分析环境指分析应用数据源的数据环境。



**操作型处理**也称联机事务处理(OLTP),是指对数据库联机的日常操作,通常是对一个或一组记录的查询和修改,如火车售票系统。



**分析型处理**也称联机分析处理(OLAP),是指对数据的查询和分析操作,通常是对海量的历史数据进行查询和分析,如金融风险预测预警系统。

## ● 数据仓库的由来

随着信息处理量的不断增多,客户需要多角度处理海量信息,并从中获取支持决策的信息。这时,面向事务处理的操作型数据库就显得有些力不从心了。因此,威廉姆·恩门(W.H.Inmon)于1993年提出了数据仓库的概念。数据仓库是为了构建新的分析处理环境而出现的一种数据存储和组织技术。它将相关数据从事务处理环境中提取出来,按照决策支持的需要重新进行组织,建立单独的分析数据环境。数据仓库的建立将操作型处理和分析型处理区分开来。传统的数据库技术为操作型处理服务,数据仓库为分析型处理服务,二者各司其职,泾渭分明。因此,越来越多的企业认识到数据仓库能够带来效益,逐步在原有的数据库基础之上建立起了自己的数据仓库。

请学习数据仓库的发展历史,根据自己的理解,回答下面的问题。

1. 数据仓库为什么会产生?

---



---

2. 数据仓库与数据库有什么关系?

---



---

### ※ 活动2 查阅数据仓库的资料

登录公开课网、网易云课堂等网课资源平台,搜索数据仓库的相关资料。根据查阅的结果,从设计目的、面向对象、数据内容、数据特征、数据结构、数据访问量这六个维度,在表6.1.1中补充数据库和数据仓库的区别。

表 6.1.1 数据库和数据仓库的区别

维度	数据库	数据仓库
设计目的	为了捕获数据而设计	为了分析数据而设计
面向对象		
数据内容		
数据特征		
数据结构		
数据访问量		

## ● 数据仓库的基本特征

数据仓库的基本特征有六个，分别是面向主题、集成、不可更新、随时间变化、数据量大和硬件要求高。

结合上述对数据库和数据仓库的分析，分小组进行交流、讨论，对数据仓库的基本特征进行描述，并在班级中共享小组成果，补充表6.1.2。

表6.1.2 数据仓库的基本特征描述

数据仓库的基本特征	数据仓库的基本特征描述
面向主题	
集成	
不可更新	
随时间变化	
数据量大	
硬件要求高	

### ※ 活动3 确定适合决策的分析环境

结合数据仓库的基本特征及其与数据库的区别，回答下面的问题。

1. 你认为利用数据仓库能否构建出适合决策的分析环境？为什么？

---



---

2. 你认为利用数据仓库构建的环境是否比利用数据库构建的环境更适合用于决策分析？都体现在哪些方面？

---



---

## 6.2 初识数据挖掘

数据挖掘是辅助决策的重要手段。它的目的是帮助决策者寻找数据中隐含的信息，发现先前忽略的或者不知道的要素，而这些要素对于预测和决策行为是十分有用的。那么什么是数据挖掘？数据挖掘的一般流程是什么？本节将围绕上述问题展开学习。



### 学习目标

- ★ 理解数据挖掘的基本概念。
- ★ 认识数据挖掘的一般流程。

佳玮通过学习了解到，数据仓库能够帮助她建立适合决策的分析环境。同时，数据分析师告诉佳玮，利用数据挖掘这一数据分析技术，可以更好地帮助她进行决策。那么，什么是数据挖掘？数据挖掘如何帮助人们进行决策？



### 任务 剖析数据挖掘的一般流程

#### ※ 活动1 构想数据挖掘的一般流程

通过之前的学习，佳玮已经知道数据仓库能够帮助她建立适合决策的分析环境，但是如果要进行决策还需要利用数据挖掘技术。如果你是数据分析师，请你构想一下，该如何进行数据挖掘呢？

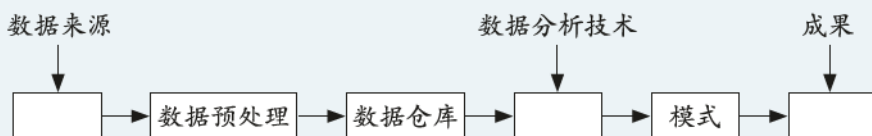
### ● 数据挖掘

数据挖掘是从大量数据中发现并提取隐藏其中的、人们事先不知道的、但又可能有用的信息和知识的一种新技术。数据挖掘在诸多领域都

得到了广泛应用，尤其是在电子商务领域中。它能够对客户进行精细化分类，分析客户的特征与喜好，从而掌握客户的潜在需求，在更高程度上满足客户的个性化需要。

结合对数据挖掘概念的理解，进行头脑风暴，尝试构想数据挖掘的一般流程并说明理由，必要时可请老师帮忙。

1. 构想数据挖掘的一般流程。



2. 请简单描述数据挖掘的一般流程。

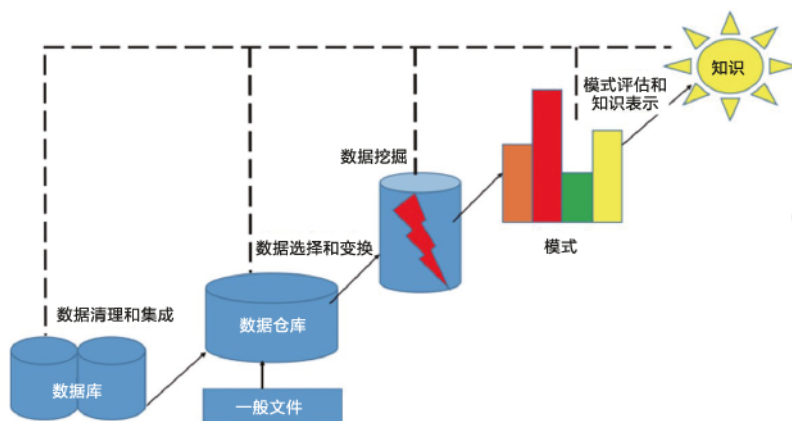
---



---

## ● 数据挖掘的一般流程

一般来说，人们把数据挖掘的流程等同于数据中的知识发现的过程。因此，与知识发现的过程一样，数据挖掘的一般流程主要包括七个步骤，分别是数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估、知识表示，如图6.2.1所示。



**模式：**对研究的实体进行必要的简化，并用适当的表现形式或规则把它的主要特征描述出来。

图 6.2.1 数据挖掘的一般流程

数据挖掘流程的各个步骤及相关描述如表6.2.1所示。

表6.2.1 数据挖掘流程的步骤及描述

步骤	相关描述
数据清理	消除噪声数据
数据集成	将多种数据源组合在一起
数据选择	从数据库中提取与分析任务相关的数据
数据变换	通过汇总或聚集操作，把数据变换并统一成适合挖掘的形式
数据挖掘	使用智能方法提取数据模式
模式评估	根据某种兴趣度量，识别代表知识的真正有趣的模式
知识表示	使用可视化和知识表示技术，向用户提供挖掘的知识



**兴趣度量**主要包括客观的度量和主观的度量。客观的度量主要考虑数据的统计显著性。主观的度量综合考虑原始数据和用户的领域知识。

### ※ 活动2 完善构想的数据挖掘流程图

分小组讨论，思考如何修改并完善构想的数据挖掘流程图，回答下列问题。

1. 你所构想的数据挖掘流程是否是这样的？

---



---

2. 如果不是，你认为是在哪部分出现了错误？根据所学的知识，修改并完善你所构想的数据挖掘流程图。

---



---

## 6.3 数据挖掘的应用

数据挖掘的历史虽然不长，但是它的发展却非常迅速，在诸多领域都被广泛地应用。那么，数据挖掘为什么会发展得这么迅速？它究竟能够做什么？本节将结合案例，了解数据挖掘的基本功能，并以数据挖掘在电子商务中的应用为例，通过对现有网络商城的体验，了解数据挖掘的具体作用。



### 学习目标

- ★ 理解数据挖掘的基本功能。
- ★ 认识数据挖掘对现实生活的重要意义。

在了解数据挖掘的基本概念之后，佳玮清楚了数据挖掘能够帮助人们进行决策。但是同时她还有一些困惑，数据挖掘能够在哪些方面帮助人们进行决策呢？请与佳玮一探究竟。



### 任务 利用数据挖掘辅助决策

#### ※ 活动1 了解数据挖掘的基本功能

佳玮已经了解了数据挖掘的一般流程，但如何利用数据挖掘帮助人们进行决策呢？有哪些案例利用了数据挖掘技术？

#### ● 数据挖掘的基本功能

数据挖掘具有关联规则、聚类、分类和预测四大基本功能。

通过阅读不同的案例，了解数据挖掘到底能够在哪些方面帮助人们进行决策。



案例1 有数据表明年龄在25~35岁的年轻父亲在给婴儿买纸尿裤的同时，有30%~40%的父亲会为自己买啤酒。于是，沃尔玛在进行商品布局时，将啤酒与纸尿裤放在了一起，这就是曾经轰动一时的啤酒与纸尿裤故事。

沃尔玛为什么要把啤酒和纸尿裤这两个不相关的商品放到一起呢？其实啤酒与纸尿裤的故事用到的就是数据挖掘的关联规则。

## ● 关联规则

关联规则是反映一个事件和其他事件之间依赖或关联的知识。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。这就是为什么电信公司、保险公司和美容公司等服务业会设计各种套餐来实现捆绑促销。

案例2 俗话说：“物以类聚，人以群分。”银行会根据储蓄金额、刷卡消费金额、诚信度等变量对客户进行分类，从而找出“黄金客户”。这样，他们就可以通过制定一些更吸引人的服务来留住客户。例如，一定额度和期限的免息投资服务、某商城的贵宾打折卡、过生日的时候送上一个小蛋糕等。

银行是如何找出“黄金客户”的呢？答案是使用数据挖掘中的聚类功能。

## ● 聚类

聚类就是将数据对象划分为若干个类，在同一类中的对象具有较高的相似度，而不同类中的对象差异较大。在聚类前并不知道所研究的对象有多少类，聚类的过程就是通过相似性的度量，使对象聚集成若干个类，各个类的成员具有共同的或相似的特征。

案例3 信用卡公司根据客户的信誉程度，对每个信誉等级建立分类分析模型，如信誉良好的客户是那些收入在5万元以上、年龄在40到50岁之间的人士。得出这个分类分析模型之后，信用卡公司就可以根据这个分类分析模型对新的用户（根据其收入和年龄）进行分类，从而判断一个新的持卡人的信誉等级是什么。



聚类与分类的区别在于事先有无定义类别。

信用卡公司是如何判断持卡人的信誉等级的呢？这里用到的是数据挖掘中的分类功能。

## ● 分类

分类就是指通过分析训练集中的数据，为每个类别建立分类分析模型，然后用这个分类分析模型对数据库中的其他记录进行分类。



**训练集：**指在建立一个模型时给定的数据类集或概念集。例如，在案例3中，训练集指的是客户在银行中产生的所有数据集。

**案例4** 当用户频繁地浏览网络商城里的某类商品时，系统就会根据用户的行为预测用户的潜在需求，如用户需要什么商品、用户购买的可能性有多大、用户还可能对哪些相关的产品感兴趣等。

系统是如何预测用户的潜在需求的呢？这里使用的是数据挖掘中的预测功能。

## ● 预测

预测是指根据时间序列型数据，由历史的和当前的数据来推测未来的数据，也可以认为是以时间为关键属性的关联知识。

### ※ 活动2 搜集应用数据挖掘功能辅助决策的案例

选择一个你感兴趣的数据挖掘功能，上网查阅资料，找出一个你认为最具有代表性的案例，填写下面的分析报告，并与同学进行分享交流。

#### 数据挖掘功能应用案例分析报告

##### 【搜集的案例及简单描述】

---



---

##### 【案例分析要点】

1. 案例体现了数据挖掘中的什么功能？

---



---

2. 案例是如何体现数据挖掘的功能的？

---



---

在了解了数据挖掘的基本功能之后，佳玮对自己的网店更有信心了。她迫不及待地想要体验一下数据挖掘的这些功能，希望通过对现有网络商城的观察与体验，给自己的网店经营带来一些借鉴。

### ※ 活动3 体验各大网络商城中的数据挖掘功能

每逢电商活动日，各大网络商城都会向顾客推送各种服务信息。请上网体验并分析这些服务信息都体现了数据挖掘的哪些功能，完成表6.3.1。

表 6.3.1 网络商城体验列表

序号	网络商城	体验结果
范例	京东	体验的功能是：_____ 预测功能 _____ 举例：_____ 商品推荐 _____ 简单描述： <u>能够根据用户的浏览路径和点击频率，预测用户的潜在购买需求，并进行个性化的商品推荐。</u>
1		体验的功能是：_____ _____ 举例：_____ _____ 简单描述：_____ _____
2		体验的功能是：_____ _____ 举例：_____ _____ 简单描述：_____ _____
3		体验的功能是：_____ _____ 举例：_____ _____ 简单描述：_____ _____



## 单元学习评价

通过本单元的学习，我们了解了数据仓库的概念和基本特征，认识了数据挖掘的一般流程及其对现实生活的重要意义。你能否胜任数据挖掘的相关工作？请参与小组交流、讨论并反思，开展小组评价或自我评价。

1. 数据仓库的基本特征是\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_、\_\_\_\_\_。
2. 关于联机事务处理（OLTP）和联机分析处理（OLAP）的区别，描述不正确的一项是（ ）。
  - A. OLAP主要是关于如何理解聚集的大量不同的数据，它与OLTP应用程序不同
  - B. 与OLAP应用程序不同，OLTP应用程序包含大量相对简单的事务
  - C. OLAP的特点在于事务量大，但事务内容比较简单且重复率高
  - D. OLAP是以数据仓库为基础的，但其最终数据来源与OLTP一样均来自底层的数据库系统，两者面对的用户是相通的
3. 请画出数据挖掘的流程图，并说一说数据挖掘对你的学习和生活有什么帮助。
4. 请帮助佳玮设计“周年庆”促销方案，每组选出一个最棒的，在班级中展示。

### “周年庆”促销方案

商品分类：\_\_\_\_\_

商品布局：\_\_\_\_\_

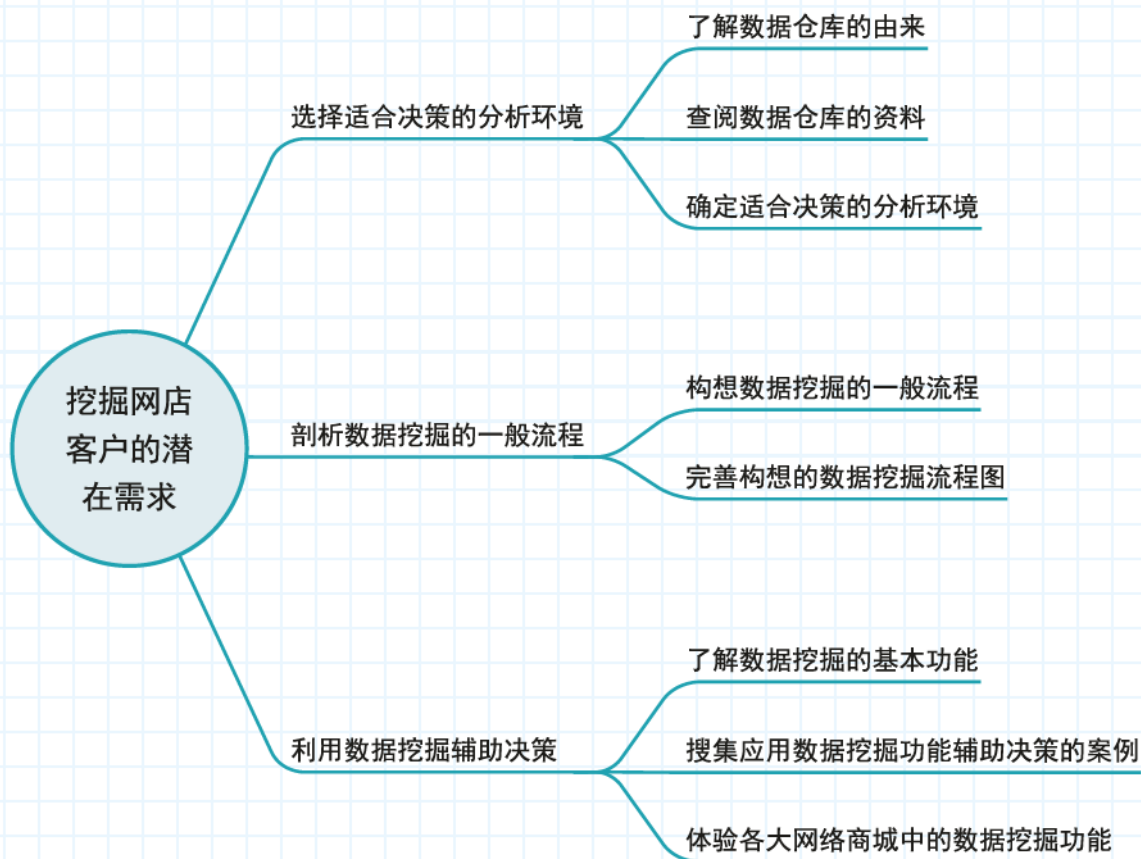
个性化商品推荐：\_\_\_\_\_

用户等级划分：\_\_\_\_\_

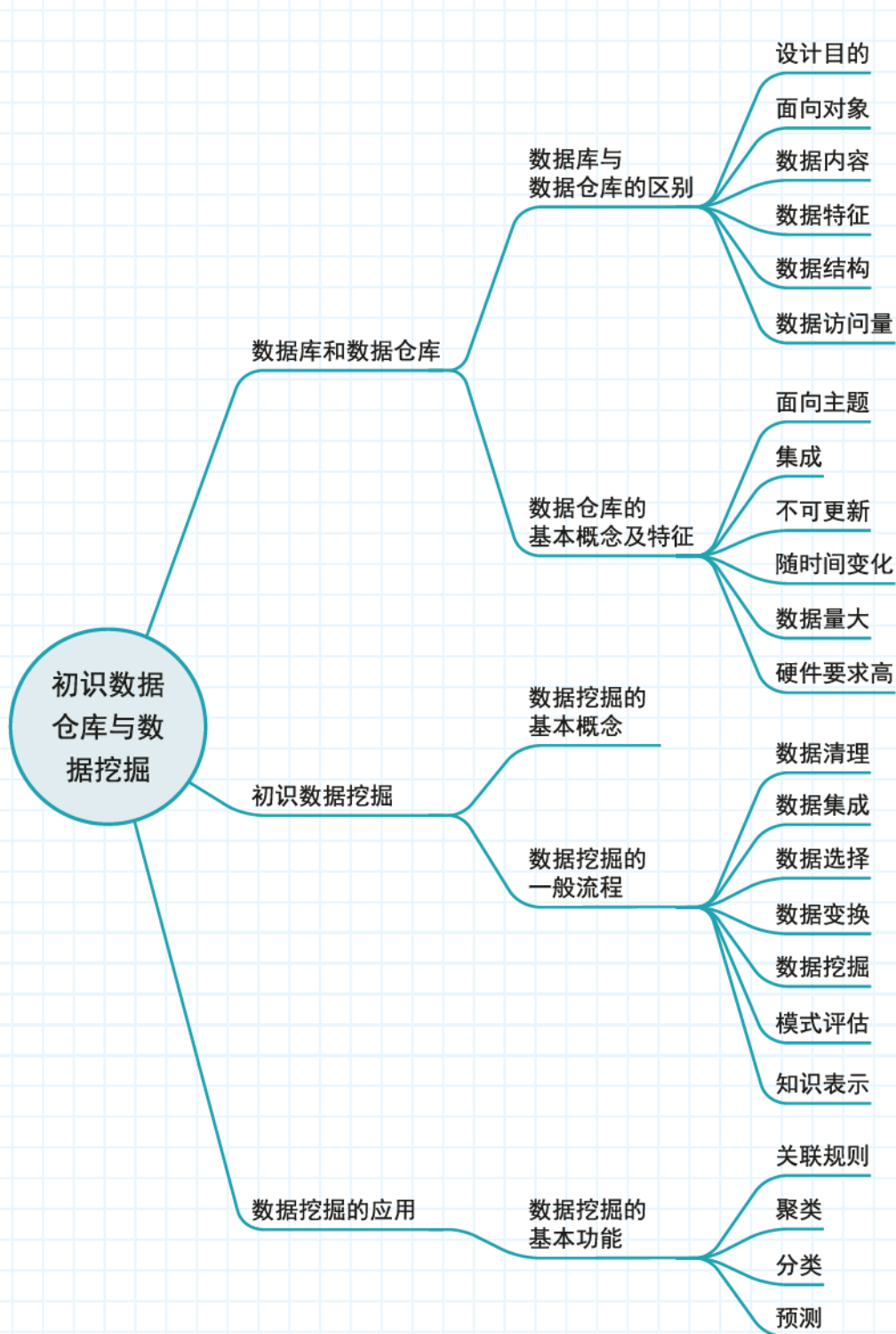
## 单元学习总结

本单元围绕“挖掘网店客户的潜在需求”项目展开学习。通过依次完成选择适合决策的分析环境、剖析数据挖掘的一般流程和利用数据挖掘辅助决策等任务，开展相应的主体活动，了解数据仓库的概念与特征，知道数据仓库与数据库的关系，了解数据挖掘的基本概念、一般流程及其主要功能，认识数据挖掘对电子商务的重要意义，从而最终形成对数据分析与管理技术的完整认识。

### “项目—任务—活动”结构图



## 知识地图



中英文术语对照表（按页码排序）

中文	英文全称或缩写	页码
数据管理	Data Management	3
数据分析	Data Analysis	3
机器学习	Machine Learning	3
需求分析	Requirement Analysis	6
需求调研	Requirement Investigation	7
业务需求	Service Requirement	7
需求确认	Requirement Confirmation	8
需求的优先级	Priority of Requirement	8
销售业绩	Sales Revenue	9
问题树	Problem Tree	10
方案评估	Scheme Evaluation	14
网络爬虫	Web Crawler	23
结构化数据	Structured Data	31
半结构化数据	Semi-structured Data	31
非结构化数据	Un-structured Data	31
数据定义语言	DDL (Data Definition Language)	49
数据操作语言	DML (Data Manipulation Language)	49
数据库管理系统	DBMS (Database Management System)	49
ActiveX数据对象	ADO (ActiveX Data Objects)	55
数据访问对象	DAO (Data Access Object)	55
Java数据库连接	JDBC (Java Database Connectivity)	55
对象链接嵌入数据库	OleDb (Object Linking and Embedding Database)	55
开放数据库连接	ODBC (Open Database Connectivity)	55
远程数据对象	RDO (Remote Data Objects)	55
结构化查询语言	SQL (Structured Query Language)	56

续表

中文	英文全称或缩写	页码
数据库事务	Database Transaction	60
4P理论	4P ( Product, Price, Place, Promotion )	70
七问分析法	5W2H ( Why, What, Who, When, Where, How, How much )	70
宏观环境分析法	PEST ( Politics, Economy, Society, Technology )	70
STP理论	STP ( Segmenting, Targeting, Positioning )	70
态势分析法	SWOT ( Strengths Weaknesses Opportunities Threats )	70
客单价	Per Customer Transaction	70
统计产品与服务解决方案软件	SPSS ( Statistical Product and Service Solutions )	78
矩阵实验室软件	MATLAB ( Matrix Laboratory )	78
统计与数据科学软件	Stata ( Software for Statistics and Data Science )	78
统计分析系统软件	SAS ( Statistical Analysis System )	78
同比	On Year-on-year Basis	79
环比	Month-on-month Ratio	79
数据分析师	Data Analyst	93
数据库管理员	DBA ( Database Administrator )	99
审计日志	Audit Log	100
风险评估	Risk Assessment	112
数据仓库	Data Warehouse	115
联机事务处理	OLTP ( Online Transaction Processing )	116
联机分析处理	OLAP ( Online Analytical Processing )	116
数据挖掘	Data Mining	118
数据中的知识发现	KDD ( Knowledge Discovery in Database )	119
关联规则	Association Rules	121
聚类	Cluster	121
分类	Classification	121
预测	Prediction	121



## 后 记

为全面落实立德树人根本任务，着力发展学生的核心素养，根据《普通高中课程方案（2017年版）》的精神，我们按照《普通高中信息技术课程标准（2017年版）》的要求对高中信息技术教科书进行了修订。

本书的修订由解月光、张晓卉、于颖、王海、褚丹、徐庆鑫等直接参与，李艺教授、曹雪利教授、刘志勇副教授审阅了本书修订稿并提出了宝贵意见。在此，我们对所有关心、支持本书编写与修订的专家、学者和老师们表示衷心的感谢。

本书选用了一些图片和文字资料，对相关作者和出版社，我们一并表示诚挚的谢意。

编者

2019年6月