《大数据导论》课后习题答案集（2022）

第 1 章课后习题

1.答案要点

数据是指所有能输入到计算机并被计算机程序处理的符号的总称，是具有一定意义的数字、字母、符号和模拟量的统称。数据的形式有很多，可以为数值、文字、声音、图像、视频或其他计算机可以识别和处理的多种形式。

信息是经过加工处理后，具有一定含义、存在逻辑关联、有时效性、对决策有价值的数据，信息使数据之间建立了相互联系。

知识是在一定背景/语境下，将数据与信息、信息与信息在应用之间建立起有意义的联系。知识不是信息的简单累加，加入以往的经验后具有判断和预期、方法论和技能等，可以解决较为复杂的问题。

智慧是人类所表现出来的一种独有的能力，主要表现为收集、加工、应用、传播知识的能力，以及对事物发展的前瞻性看法。智慧是一种推测的、非确定性和非随机的过程，要回答人们难以得到甚至无法得到答案的问题，是一种卓越的判断力和决策能力。

DIKW（Data, Information, Knowledge, Wisdom）金字塔模型刻画了人类对数据的认识程度的改变过程，它体现了与数据相联系的信息、知识和智慧这些概念，还向我们展现了数据是如何一步步转化为信息、知识、乃至智慧的方式。

DIKW 金字塔模型刻画了人类对数据认识程度的转变过程，也就是说，数据、信息、知识和智慧是人类认识客观事物过程中不同阶段的产物，既是一个从低级到高级的认识过程，也是一个从不可预知到可预知的增值过程，即数据通过还原其真实发生的背景成为信息，信息赋予其内在含义成为知识，知识通过理解变成智慧。由此可见，数据在 DIKW 金字塔中的重要作用。

2.答案要点

（1）结构化数据

结构化数据，也称作行数据，是以先有结构、后有数据的方式生成的数据，其一般特点是：数据以行为单位，一行数据表示一个实体信息，每一行数据的属性相同，如下表给出了主要农业国粮食产量与耕地情况，它们是结构化数据。

表 1-1 2013 年度主要农业国粮食产量与耕地情况

国家	粮食总产量（亿吨）	耕地面积（亿亩）	占世界耕地比例（%）
中国	5.01	18.15	8.06
美国	3.63	29.55	13.15
印度	2.16	25.5	11.32
巴西	1.33	12.9	5.76
加拿大	0.51	10.2	4.52
澳大利亚	0.31	7.65	3.45

（2）非结构化数据

非结构化数据是指数据结构不规则或不完整、没有预先定义的数据模型，很难用关系数据库的二维逻辑表来表现的数据，比如一段文本、一张图片和音频/视频信息等等都是非结构化数据。

（3）半结构化数据

半结构化数据介于结构化数据和非结构化数据之间。半结构化数据包含相关标记，用来分隔语义元素以及对记录和字段进行分层，如下表中数据就是半结构化数据。可以看出有三个地区结构，每个地区结构中，有的是两个粮食作物，有的是三个粮食作物，半结构化数据能够灵活地表达数据信息。

XML 格式数据	JSON 格式数据

<部分地区主要作物产量（万吨）>
<地区 名称=“北京”>
<小麦>18.7</小麦>
<玉米>75.2</玉米>
</地区>
<地区 名称=“河北”>
<稻谷>58.8</稻谷>
<玉米>1703.9</玉米>
<小麦>1387.2</小麦>
</地区>
<地区 名称=“广西”>
<稻谷>1156.2</稻谷>
<甘蔗>8104.3</甘蔗>
</地区>
</部分地区主要作物产量（万吨）>
	{
	"部分地区主要作物产量（万吨）":{
	"北京":{
	"小麦":18.7,
	"玉米":75.2
	},
	"河北":{
	"稻谷":58.8,
	"玉米":1703.9,
	"小麦":1387.2
	},
	"广西":{
	"稻谷":1156.2,
	"甘蔗":8104.3
	}
	}
	}

3.答案要点

（1）Volume（数量大）

大数据的重要特征之一就是数量大，随着网络和信息技术的高速发展，社交网络、移动网络、各种智能终端等都成为数据的来源，数据开始爆发式增长，存储单位由传统的GB或TB字节发展到现在的PB、ZB字节甚至更高。

（2）Variety（种类多）

现有数据来源不仅有传感器、智能设备自动产生的数据，还有人类自身的生活行为，除数字、符号等结构化数据之外，更有大量包括网络日志、音频、视频、图片、地理位置信息等半结构化或非结构化数据，数据类型庞杂。

（3）Velocity（速度快）

数据产生和更新的频率，也是衡量大数据的一个重要特征。随着现代感测、互联网、计算机技术的发展，通过高速的计算设备、探测设备或者社交工具，创建实时、动态数据已成为流行趋势。例如，Facebook每天有18亿照片上传或被传播；全国公路上安装的交通堵塞探测传感器和路面状况传感器每天都在产生大量的数据。

（4）Value（价值高）

大数据有巨大的潜在价值，但是有价值的数据往往被淹没在海量无用数据中，比如，一天24小时监控录像，可用的关键数据也许仅为1－2秒钟。每天数十亿的搜索申请中，只有少数固定词条的搜索量会对某些分析研究有用处。

（5）Veracity（真实性）

数据的真实性即数据的质量，只有真实而准确的数据才对数据管控和治理真正有意义。大数据来源于不同领域和用户，这些数据的有效性、真实性以及所提供数据的个人或单位的信誉都与原来数据产生的方式有区别，社会和企业愈发需要有效的信息治理以确保其真实性和安全性。

大数据的4V是指前4个维度，最后一个维度新提出，构成5V特征。

挑战：数据量的爆炸式增长导致对原有数据存储架构、计算模型和应用软件系统都提出了全新的挑战。在算力研究方面，由国家超级计算无锡中心研制的“神威·太湖之光”，在2016年德国法兰克福国际超算大会(ISC)公布的全球超级计算机500强榜单中夺得第一。在算法方面，需要研究有效的机器学习算法，用来有效处理大数据的实时分析，以满足用户需求。另外，大数据最大的价值在于通过从大量不相关的数据中，挖掘出对未来趋势与预测分析有价值的数据，更需要提出更为有效的算法和软件系统。

4.答案要点

大数据时代，人们对待数据的思维方式会发生如下三个变化：第一，人们处理的数据从样本数据变成全部数据；第二，由于是全样本数据，人们不得不接受数据的混杂性，而放弃对精确性的追求；第三，人类通过对大数据的处理，放弃对因果关系的渴求，转而关注相关关系。总体上大数据思维可以概括为全样思维、相关思维、容错思维。

全样思维：大数据时代，随着数据收集、存储、分析技术的突破性发展，我们可以更加方便、快捷、动态地获得研究对象有关的所有数据，而不再因诸多限制不得不采用样本研究方法，相应地，思维方式也应该从样本思维转向总体思维，从而能够更加全面、立体、系统地认识总体状况。

例如：苹果公司总裁乔布斯治疗胰腺癌方式。在与癌症抗战8年的斗争过程中，乔布斯采用了不同的治疗方式，成为世界上第一个对自身所有DNA和肿瘤DNA进行排序的人，得到了整个基因密码的数据文档，而不是只有一系列标记的一个样本。基于此，医生们能够基于

乔布斯的特定基因组成的大数据，按所需效果用药并调整医疗方案。

相关思维：大数据时代，通过大数据技术可以挖掘出事物之间隐蔽的相关关系，获得更多的认知与洞见，运用这些认知与洞见帮助我们捕捉现在和预测未来，建立在相关关系分析基础上的预测正是大数据的核心议题。通过关注线性的相关关系以及复杂的非线性相关关系，可以帮助我们看到很多以前不曾注意的联系，还可以掌握以前无法理解的复杂技术和社会动态，相关关系甚至可以超越因果关系，成为我们了解这个世界的更好视角。

啤酒和尿布的故事是沃尔玛利用大数据获益的典型案例，也是体现大数据相关关系的典型案例。沃尔玛超市的管理人员在分析销售数据时，发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然会经常出现在同一个购物篮中，这一独特的销售现象引起了高管的重视。原来，美国的妇女通常在家照顾孩子，所以她们经常会嘱咐丈夫在下班回家的路上为孩子买尿布，而丈夫在买尿布的同时又会顺手购买自己爱喝的啤酒。沃尔玛发现这一现象后，开始尝试把啤酒和尿布摆放在同一区域，让年轻的父亲可以同时找到这两件商品，并很快地完成购物，这项措施为商家带来了大量的利润。

容错思维：大数据时代，大量非结构化、异构化的数据能够得到储存和分析，这对传统的精确思维造成了挑战。也就是说，当拥有海量实时数据时，绝对的精准不再是追求的主要目标，适当忽略微观层面上的精确度，容许一定程度的错误与混杂，反而可以在宏观层面拥有更好的知识和洞察力。

2006 年，谷歌公司开始涉足机器翻译。为了训练计算机，谷歌翻译系统吸收了它能找到的所有翻译，谷歌翻译部负责人弗朗兹·奥齐指出，“谷歌的翻译系统不会像 IBM 的 Candide 一样只是仔细地翻译 300 万句话，它会掌握用不同语言翻译的、质量参差不齐的、数十亿页的文档”。尽管输入源很混乱，但对比其他翻译系统，谷歌的翻译质量相对而言最好，而且可翻译内容也更多。到 2012 年，谷歌数据库涵盖了 60 多种语言，能够接受 14 种语言的语音输入，并有很流利的对等翻译。

5.答案要点

数据科学是关于数据的科学，是研究探索网络空间中数据奥秘的理论、方法和技术，可以理解为基于传统的数学和统计学理论和方法，运用计算机技术进行大规模数据计算、分析和应用的一门学科。自吉姆格雷提出数据密集型发现将成为科学研究的第四范式之后，科学研究从原有的实验科学、理论科学、计算科学，发展到目前兴起的数据科学。

总体来说，数据科学主要有两方面内涵：一是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律，二是为自然科学和社会科学研究提供一种新方法，称为科学研究的数据方法，其目的在于揭示自然界和人类行为现象和规律。

6.答案要点

大数据技术是许多技术的集合体，关系型数据库、数据仓库、OLAP(On-Line Analytical Processing)、数据挖掘、商务智能等是大数据技术的组成部分。大数据处理分析还需要大数据新技术的支撑，包括分布式文件系统、分布式数据库、分布式并行编程等，不同的应用场景对应不同的大数据计算模式，典型的大数据处理技术有批处理计算、流计算、图计算以及查询分析等。

（1）批处理计算

主要针对大规模数据的批量处理，是日常数据分析工作中常见的一类数据处理需求。MapReduce 是一种处理海量数据的并行编程模式，适用于在大规模计算集群上编写离线的、大数据量的、相对快速处理的并行化程序，适合搜索引擎、Web 日志分析、文档分析处理、

机器翻译等针对文本型数据分析的应用领域。Spark 解决了 MapReduce 框架表达能力不足的缺点，适用于迭代计算，弥补了 MR 高延迟的缺陷。

（2）流计算

主要针对实时处理来自不同数据源、连续到达的流数据，经过实时分析处理，给出有价值的分析结果。流计算与离线批处理不同，它是在数据到达的同时即进行计算处理，计算结果实时输出。

（3）图计算

大数据时代，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、交通路网等。这类以图型表征的数据在大数据系统需要处理的数据量中占有相当大的比例，因此，图数据的表达、建模、存储、处理成为大数据计算体系的一种特定类型。

（4）查询分析计算

查询分析计算式一种企业常见的应用场景，主要面向大规模数据的存储管理和查询分析，用户输入查询语句，可以快速得到相关的查询结果。hive 是基于 Hadoop 的一个数据仓库工具，提供完整的 sql 查询功能。其优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，十分适合数据仓库的统计分析。

7.答案要点

大数据处理流程主要包含数据采集、数据预处理、数据存储与管理、数据挖掘分析和数据可视化等

（1）大数据采集

数据采集是指数据获取，它主要通过物联网、互联网、移动互联网以及各类业务应用平台等来获取结构化、半结构化、非结构化的海量数据，常用的数据采集方式包括：批量采集、数据抓取，智能传感设备自动采集、业务数据导入等方式，常用的大数据采集工具有 Flume、Kafka 等。

（2）大数据预处理

数据大体上都是不完整、不一致的“脏”数据，无法直接进行数据挖掘，或者挖掘效果不佳。为了提高数据挖掘的质量，需要对数据进行预处理，常用的数据预处理方法包括数据清洗、数据集成、数据变换和数据规约等。

（3）大数据存储

大数据要用存储器把采集到的数据存储起来，需要建立相应的数据库进行管理和调用。大数据存储重点解决复杂的结构化、半结构化和非结构化的大数据管理与处理技术，大数据存储通常采用分布式文件系统、关系型数据库、NoSQL 数据库以及云存储等技术。

（4）大数据分析与挖掘

大数据分析与挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取出人们事先未知的、隐含在其中的、有潜在有用知识的过程。与传统统计分析方法不同的是，大数据分析与挖掘一般没有预先设定好的主题，主要是在现有数据基础上进行基于各种算法的计算，从而起到预测效果。大数据分析与挖掘技术主要包含关联分析、聚类分析、分类外析、预测模型、回归分析等技术。此外，机器学习、深度学习等也是常用的数据挖掘算法。

（5）大数据可视化

数据可视化就是运用计算机图形和图像处理技术，将数据转化为图形图像显示出来，其根本目的是实现对稀疏、杂乱、复杂的数据深入洞察，发现数据背后有价值的信息。数据可视化并不是简单地将数据转化为可见的图形符号和图表，而是能将不可见的数据现象转化为可见的图形符号和图表，能将错综复杂、看起来没法解释和关联的数据，建立起联系和关联，

发现规律和特征，获得更有商业价值的洞见。

8.答案要点

大数据分析与挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取出人们事先未知的、隐含在其中的、有潜在有用知识的过程。与传统统计分析方法不同的是，大数据分析与挖掘一般没有预先设定好的主题，主要是在现有数据基础上进行基于各种算法的计算，从而起到预测效果。

大数据分析与挖掘技术主要包含关联分析、聚类分析、分类分析、预测模型、回归分析等技术。此外，机器学习、深度学习等也是常用的数据挖掘算法。

- 关联分析发现存在于大量数据集中的关联性或相关性，从而描述一个事物中某些属性同时出现的规律和模式，目标在于发现数据集中隐藏的相关联系。
- 聚类分析在于将数据集内具有相似特征属性的数据聚集在一起，同一个数据群中的数据特征要尽可能相似，不同数据群中的数据特征要有明显区别。
- 分类分析是根据重要数据类的特征向量及其他约束条件，构造分类函数或分类模型，目的是根据数据集的特点把未知类别的样本映射到给定类别中。
- 预测建模是根据数据集的特征，以目标结果为目的建立映射关系，预测建模有两个任务，一是分类，用于预测具有多种属性的数据类别，二是回归，用于预测连续数据及未来的变化趋势。

9.答案要点

大数据应用场景包括各行各业对大数据处理和分析的应用，不同行业的用户有不同的需求，本节仅举几个代表性行业应用场景，说明各行业如何使用大数据创造价值。

（1）农业大数据应用

农业大数据就是一切与农业相关的数据，包括上游的种子、化肥和农药等农资研发，气象、环境、土地、土壤、作物、农资投入等种植过程数据，以及下游的农产品加工、市场经营、物流、农业金融等数据，都属于农业大数据的范畴，贯穿整个产业链。大数据可以加速作物育种、实现农产品追溯、精准管理、自动识别作物病虫害、智能识别生理状态、快速数果预测产量等应用。

（2）教育大数据应用

教育大数据顾名思义就是教育行业的数据分析应用，大数据对教育行业产生了重大影响。基于大数据的个性化教学、科学化评价、精细化管理、智能化决策等，将对促进教育公平、提高教育质量、培养创新人才具有不可估量的作用。

大数据可驱动教学模式重塑、驱动评价体系重构、精准教育，对教育大数据的全面收集、准确分析、合理利用，已成为教育决策创新的重要驱动力。

（3）零售大数据应用

零售行业最有名气的大数据案例就是沃尔玛的啤酒和尿布的故事以及 Target 通过向年轻女孩寄送尿布广告而告知其父亲女孩怀孕的故事，这是典型的关联分析和精准广告案例。

零售行业大数据应用有两个层面，一个层面式零售行业可以了解用户的消费喜好和趋势，进行商品的精准营销，降低营销成本，另一个层面是可以通过客户购买记录，了解客户关联产品购买喜好，将相关的产品放到一起来增加产品销售额，

（4）金融大数据应用

金融行业拥有丰富的数据，并且数据质量和数据维度都很好，应用场景较为广泛，典型的金融行业应用场景有：银行数据应用场景、保险行业应用场景和证券数据应用场景。

（5）医疗大数据应用

医疗大数报主要包括三大类:第一类为医疗机构各类临床相关信息系统获得或产生的数据，第二类为个人健康与公共卫生数据，第三类为各类生物样本和各类组学数据。多种来源的数据导致医疗数据的种类多样,既包含数值型数据为主的生化检查数据,也包含医疗文本、医学影像、文献信息、生物信息等类型的数据，形成了多态非结构化医疗大数据。

疾病风险预测是医疗大数据最重要的应用，它通过所收集整理的个人健康信息，分析并建立生活方式、环境、遗传等危险因素与健康状态之间的量化关系来进行行可能性预测，帮助预测对象发现某些病的患病可能性和程度,从而针对患病概率我率比较大的项目采取积极有效的预防措施，以便最大限度地预防或延缓患病的发生。精准医疗是应用现代遗传技术、分子影像技术、生物信息技术结合患者生活环境和临床数据来实现精准的治疗与诊断，制订具有个性化的疾病预防和治疗方案。

（6）交通大数据应用

交通管理中大量传感器的介入势必产生大数据。在交通领域，海量的交通数据主要产生于各类交通的运行监控、服务，高速公路、干线公路的各类流量、气象监测数据，公交、出租车和客运年辆 GPS 数据等，数据量大且类型繁多，数据量也从 TB 级跃升到 PB 级。

交通大数据的开展，可以使人们可以利用大数据分析城市交通管理中的各项信息，从而有效提高城市交通的管理效率，交通大数据的使用为城市交通管理带来了巨大的便利。一方面，交通大数据可以更好地实现对交通数据的分析整合，并对交通管理的现状和未来做出合理的分析。另一方面，通过交通大数据，城市管理人员也可以通过交通大数据的分析结果，应用在交通规划之中，更有利于城市交通管理人员对信息的整合，从而提出更好的城市交通规划方案。

智能交通的应用会使得交通体系的分析、问题的诊断、测试等都更为便捷，智能交通一方面可以提供最合适、最节约资源的交通方式，同时也使得人们对对交通体系有更为清晰的了解。

（7）社交大数据应用

社交网络主要作用是为一群拥有相同兴趣与活动的人创建在线社区,它为信息交流与分享提供了新的途径。社交网络已经得到普遍应用，作为社交网络的网站一般会有数以百万的登记用户，越来越多的人愿意在这个交互时代分享自己的见闻感受，通过手机、电脑上产生了大量数据。大数据技术可以将客户在互联网上的行为记录下来，对客户的行为进行分析，打上标签并进行用户画像。用户画像可以帮助广告主进行精准营销，将广告直接投放到用户的移动设备，其广告的目标客户覆盖率可以大幅度提高。

第 2 章课后习题

1.答案要点

大数据的来源主要有以下几个方面：

（1）系统日志数据

系统日志数据是指来自于 WEB 服务器、企业 ERP 系统、各种 POS 终端及网上支付等业务系统数据。这些数据对于企业来说是非常重要的，通过这些日志数据进行分析，可以挖掘到具有潜在价值的信息。

（2）互联网数据

互联网是大数据信息的主要来源，互联网数据主要来自于两个方面：一方面是用户通过网络所留下的痕迹（如浏览网页、发送邮件等），另一方面是互联网运营商在日常运营中生成和累积的用户网络行为数据。

（3）物联网数据

物联网是大数据的重要来源，主要通过传感器、条形码以及无线射频识别 RFID 等技术获取大数据。

（4）传统信息系统数据

一些企业会使用传统的关系型数据库或非关系型数据库来存储业务系统数据，虽然传统信息系统数据占的比例比较小，但是由于传统信息系统的数据数据结构清晰，数据的准确性比较高，所以数据往往具有较高的价值密度。

2.答案要点

常见的日志采集平台有以下几种：

（1）Chukwa

Chukwa 是一个开源的用于监控大型分布式系统的数据采集系统，构建在 Hadoop 的 HDFS 和 MapReduce 框架之上，可用于监控大规模 Hadoop 集群的整体运行情况并对它们的日志进行分析。另外，Chukwa 还包含了一个强大而灵活的工具集，可用于展示、监控和分析已采集的数据。

chukwa 从数据的产生、收集、存储、分析到展示的整个生命周期都提供了全面的支持。Chukwa 的主要部件及其功能如下：

- Agents：负责采集最原始的数据，并发送给 Collector；
- Adaptor：直接采集数据的接口和工具，一个 Agent 可以管理多个 Adaptor 的数据采集；
- Collectors：负责收集 Agent 收送来的数据，并定时写入集群中；
- MapReduce 作业：负责把集群中的数据分类、排序、去重和合并；
- HICC：负责数据的展示。

（2）Flume

Flume 是一个高可用的、高可靠的、分布式海量日志采集、聚合和传输系统。Flume 支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume 提供对数据进行简单处理的能力。

Flume 可看作是一个管道式的日志数据处理系统。数据流由事件(Event)驱动，Event 是 Flume 的基本数据单位，每个 Event 由日志数据和消息头组成，这些 Event 由外部数据源生成。Agent 是 Flume 的运行核心，Agent 是最小的独立运行单位，它是一个完整的数据收

集工具，含有三个核心组件，分别是 Source、Channel、Sink。通过这些组件，数据可以从 WEB Server 等源地址一步步流入 HDFS 系统。

（3）Kafka

Kafka 是 Linkedin 公司开发的一个分布式、支持分区的、多副本的、基于 ZooKeeper 协调的分布式日志系统，可以用于 Web/Nginx 日志、访问日志、消息服务等等。

Kafka 实际上是一个分布式发布-订阅消息系统。Producer（生产者）向某个 Topic 发布消息，而 Consumer（消费者）订阅某个 Topic 的消息，进而一旦有新的关于某个 Topic 的消息，Broker 会传递给订阅它的所有 Consumer。

（4）Scribe

Scribe 是 Facebook 开源的日志收集系统，在 Facebook 内部已经得到大量的应用。它能够从各种日志源上收集日志，存储到一个中央存储系统上，以便于进行集中统计分析处理。

Scribe 从各种数据源上收集数据，放到一个共享队列上，然后将消息推送到后端的中央存储系统上。Scribe 的主要部件及功能如下：

- Scribe：接收 Thrift Client 发送过来的数据，并根据配置文件，将不同 Topic 的数据发送给不同的存储对象用于持久化数据；
- Scribe Agent：是一个 Thrift Client，Scribe 内部定义了一个 Thrift 接口，用户使用该接口将数据发送给不同的对象。
- 存储系统：用于持久化数据。

3.答案要点

网络爬虫是按照一定规则，自动地抓取网络信息的程序或脚本。

4.答案要点

网络爬虫工作原理：网络爬虫一般是根据预先设定的一个或若干个初始网页的 URL 开始，获取初始网页上的 URL 列表，然后按照一定的规则抓取网页。每当抓取一个网页时，爬虫会提取该网页上新的 URL 并放入未抓取的 URL 队列中，接着再从未抓取的队列中取出一个 URL 再次进行新一轮的抓取，不断重复上述过程，直到队列中的 URL 抓取完毕或者满足系统其它的停止条件，爬虫才会结束。

5.答案要点

使用八爪鱼采集器，爬取豆瓣网关于影片《哪吒之魔童降世》的短评信息，内容包括用户名、星级评分、评论内容以及评论时间。

- 网址：https://movie.douban.com/subject/26794435/comments?status=P
- 提示：“评分”字段提取后需要进行进一步的设置。由于系统默认抓取的是所选元素的文本，而“评分”字段所要抓取的是所选元素的 class 属性的值，而非文本内容，因此按默认抓取，所得结果为空。单击“评分”字段后的“更多字段操作”图标按钮，进一步选择“自定义抓取方式”选项，在打开的窗口中选择“抓取元素属性值”，选择属性名称为“class”即可。

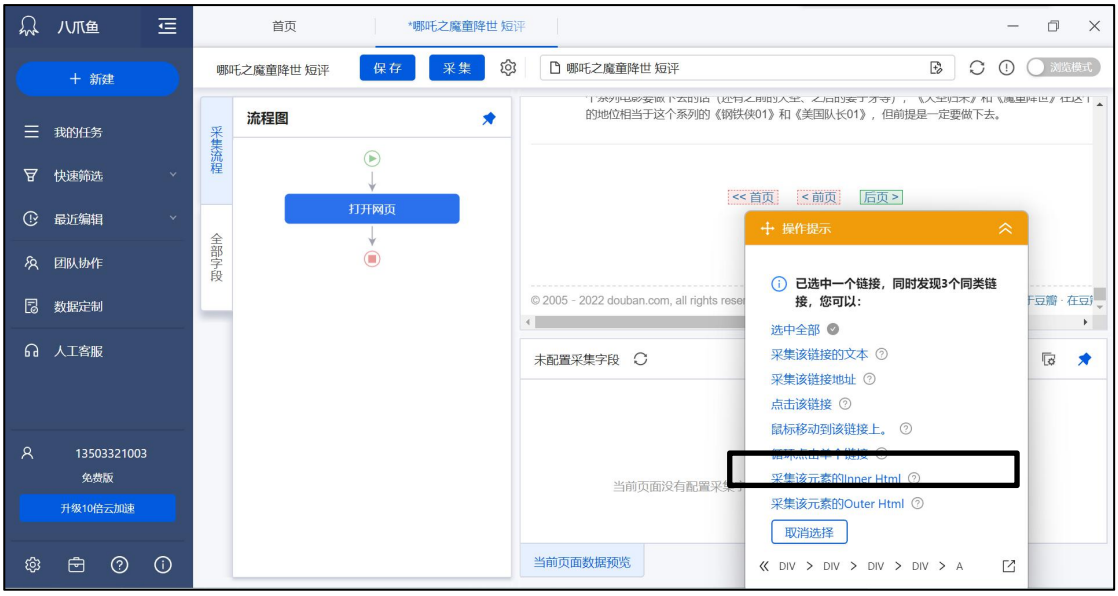
步骤 1：新建自定义任务

在八爪鱼操作界面中，单击“新建”按钮，选择“自定义任务”选项，进入新建任务窗口，将要爬取的目标网页地址输入或复制粘贴到编辑区域，单击“保存设置”按钮，如图所示。

单击“保存设置”按钮后，系统会自动加载网页内容，在网页加载过程中，需取消自动识别。

步骤 2：设置循环翻页。

拖动加载后的网页右侧滚动条至页面底部，单击“后页”按钮，在弹出的“操作提示”对话框中选择“循环点击单个链接”选项，如下图所示。



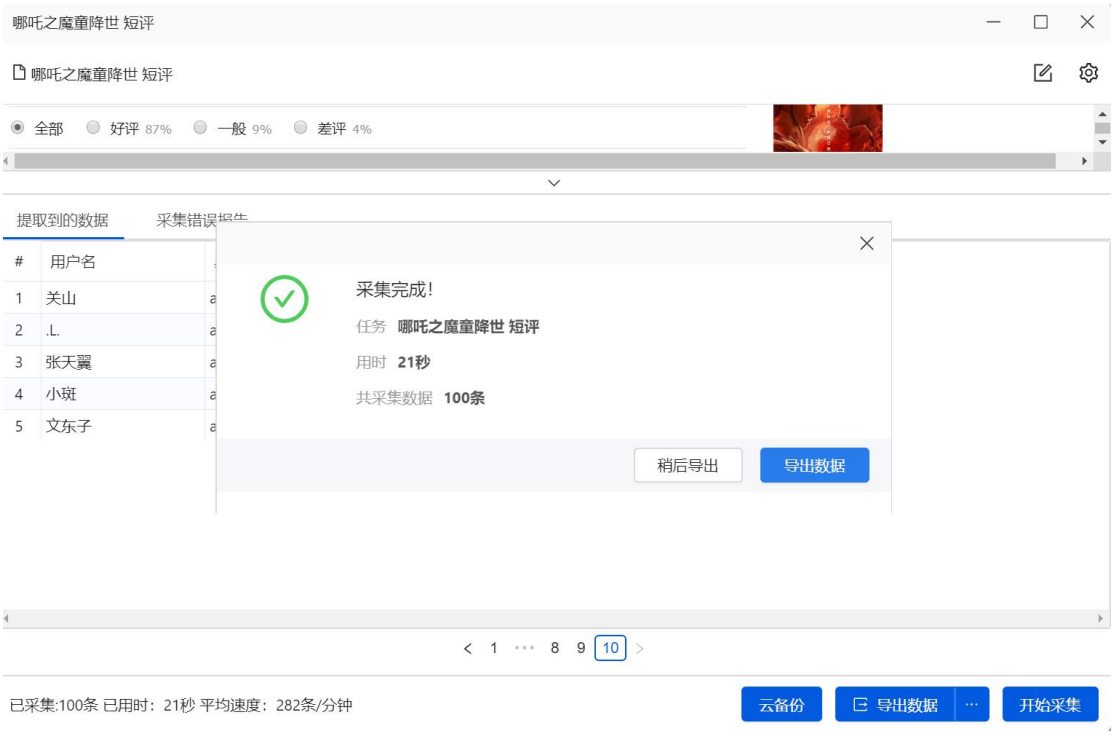
此时在操作流程图中会自动建立一个翻页循环，选中“循环翻页”框，单击右侧的设置按钮，进一步设置退出循环条件为“循环执行次数等于 5”，即爬取前 5 页数据，否则会默认爬取所有页面。

步骤 3：设置提取字段

首先提取“用户名”字段，单击页面中某条评论的用户名，在弹出的“操作提示”对话框中选择“选中全部”选项，进一步选择“采集以下链接文本”，在页面下方的数据预览框中会自动生成所要爬取的字段。单击字段名右侧的编辑按钮，将字段名修改为“用户名”。按照此方法，依次生成“星级评分”、“评论内容”、“评论日期”字段。此时，“星级评分”字段默认为空，单击字段后的“更多字段操作”图标按钮，进一步选择“自定义抓取方式”选项，在打开的窗口中选择“抓取元素属性值”，选择属性名称为“class”即可获取评分数据。提取字段全部设置完成后如下图所示：

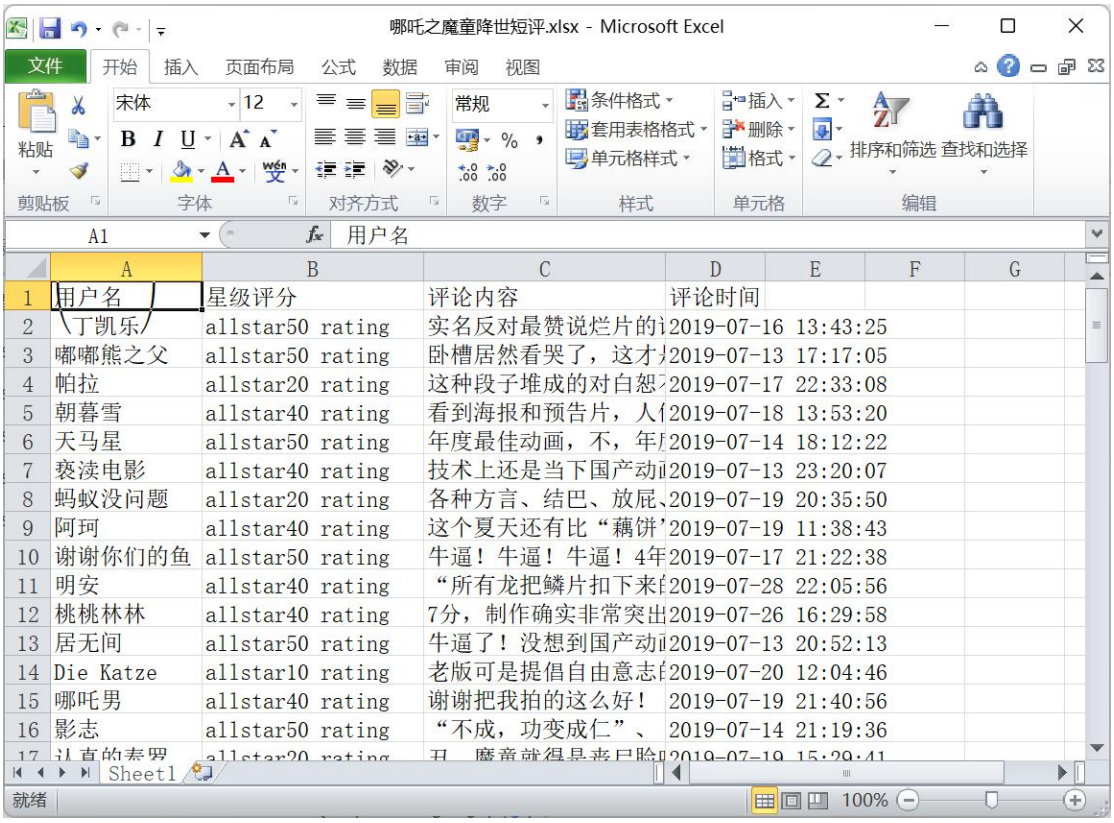
步骤 4：启动采集。

单击界面上方的“采集”按钮，并在弹出的窗口中单击“启动本地采集”，即可开始采集。本例由于设置了只爬取 5 页，因此共采集 100 条数据，如下图所示。



步骤 5：导出数据。

单击“导出数据”按钮，进一步选择导出为 Excel 文件，导出结果如下图所示。



	A	B	C	D	E	F	G
1	用户名	星级评分	评论内容	评论时间			
2	丁凯乐	allstar50 rating	实名反对最赞说烂片的	2019-07-16 13:43:25			
3	嘟嘟熊之父	allstar50 rating	卧槽居然看哭了，这才	2019-07-13 17:17:05			
4	帕拉	allstar20 rating	这种段子堆成的对白恕	2019-07-17 22:33:08			
5	朝暮雪	allstar40 rating	看到海报和预告片，人	2019-07-18 13:53:20			
6	天马星	allstar50 rating	年度最佳动画，不，年	2019-07-14 18:12:22			
7	亵渎电影	allstar40 rating	技术上还是当下国产动	2019-07-13 23:20:07			
8	蚂蚁没问题	allstar20 rating	各种方言、结巴、放屁、	2019-07-19 20:35:50			
9	阿珂	allstar40 rating	这个夏天还有比“藕饼”	2019-07-19 11:38:43			
10	谢谢你们的鱼	allstar50 rating	牛逼！牛逼！牛逼！4年	2019-07-17 21:22:38			
11	明安	allstar40 rating	“所有龙把鳞片扣下来	2019-07-28 22:05:56			
12	桃桃林林	allstar40 rating	7分，制作确实非常突出	2019-07-26 16:29:58			
13	居无间	allstar50 rating	牛逼了！没想到国产动	2019-07-13 20:52:13			
14	Die Katze	allstar10 rating	老版可是提倡自由意志	2019-07-20 12:04:46			
15	哪吒男	allstar40 rating	谢谢把我拍的这么好！	2019-07-19 21:40:56			
16	影志	allstar50 rating	“不成，功变成仁”、	2019-07-14 21:19:36			
17	认真的泰罗	allstar20 rating	丑，魔童就得是夜尽脸	2019-07-19 15:29:41			

6.答案要点如下，详见 2.3.1

原始数据中常见的问题有数值缺失、噪声数据、异常数据等，数据清洗就是通过填充缺失的数据值，光滑噪声数据、识别和删除离群点等方法，达到纠正错误、标准化数据格式、清除异常和重复数据的目的。常用技术：缺失值处理（删除有缺失的记录、忽略缺失值、填充缺失值）；消除噪声（分箱、回归、离群点分析）。

7.答案要点如下，详见 2.3.2

数据集成的本质是整合数据源，要考虑的主要问题包括多个数据源中字段的语义差异、结构差异、字段间的关联关系，以及数据的冗余重复等。

8.答案要点如下，详见 2.3.3

数据归约指在尽可能保持数据原貌的前提下，最大限度地精简数据量。数据归约得到的数据集比原数据集小得多。数据归约导致的较小数据集需要较少的内存和处理时间，因此可以使用占用计算资源更大的挖掘算法，但能够产生同样的（或几乎同样的）分析结果。常用技术包括维归约（主成分分析、属性子集选择）和数量归约（参数化归约、直方图、抽样）等。

9.答案要点如下， 详见 2.3.4

数据变换就是对数据格式统一化，目的是为了更好地完成数据挖掘。通常挖掘算法对数据格式有自身特定的限制，这就要求在进行数据挖掘前，将这些格式不一样的数据集进行数据格式的转换，使得所有数据的格式统一化。常用技术有数据立方聚集、数据离散化、数据规范化（最小-最大规范化、Z-score 标准化、小数定标规范化）等。

10. 答案

(1) 均值 ：15.08　中位数：15　众数：20

(2) 每个箱中的数据分别是：(3，6，7，7)，(8，9，9，10)，(10，12，12，12)，(15，15，15，18)，(18，18，20，20)，(20，20，24，24)，(30，30)

(3) 均值平滑：5.75，5.75，5.75，5.75，9，9，9，9，11.5，11.5，11.5，11.5，15.75，15.75，15.75，15.75，19，19，19，19，22，22，22，22，30，30

中值平滑：6.5，6.5，6.5，6.5，9，9，9，9，12，12，12，12，15，15，15，15，19，19，19，19，22，22，22，22，30，30

11. 答案

(1) 0.7

(2) 0.9276

(3) 0.22

第 3 章课后习题

1.答案要点

数据存储与管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的过程。主要管理技术包括关系型数据库、分布式文件系统、非关系型数据库 NoSQL、新型数据存储与管理技术 NewSQL 以及云存储等技术等。

2.答案要点

关系数据库建立在关系数据模型之上，主要用来存储结构化数据并支持数据的插入、查询、更新和删除等操作。（使用案例略）

3.答案要点如下，详见 3.1.2 小节

数据处理大致可以划分为事务型处理（OLTP：联机事务处理）和分析型处理（OLAP： 联机分析处理）两类。事务型处理一般针对的是具体业务，通过对一个或一组数据的查询和修改，为特定应用进行服务；分析型处理一般针对某个主题通过综合大量历史数据处理，服务于决策支持。区别详见表 3-1。

4.答案要点如下，详见 3.1.2 小节

数据仓库技术的出现解决了传统数据库在分析型处理、提供决策支持分析方面的不足。数据仓库的数据来自于事务型数据库，经过一系列的抽取、转换、加载的处理，变成对终端用户有用的信息，形成一个新的集成系统。与传统数据库相比，数据仓库的特点是数据的继承与分析能力，数据仓库的出现并不是要取代传统事务型数据库，而是以传统的事务型数据库为基础，建立一个用于支持管理层决策分析的综合信息分析应用系统。区别详见表 3-2。

5.答案

略

6.答案

主要包括分布式文件系统、NoSQL 数据库、NewSQL 数据库。

7.答案要点如下，详见 3.3.1 节

分布式文件系统在物理结构上由计算机集群中的多个节点构成。这些节点分为两类：一类叫“主节点”（MasterNode），或称为“名称节点”（NameNode）；另一类叫“从节点”（SlaveNode），或称为“数据节点”（DataNode）。系统将文件分成若干块进行存储，每个块通常备份多份，以保证数据的可靠性。名称节点负责文件和目录的创建、删除和重命名等，同时管理数据节点和文件块的映射关系。客户端只有访问名称节点才能找到请求的文件块所在位置，进而到相应位置读取所需文件块。数据节点负责数据的存储和读取，在存储时，由名称节点分配存储位置，然后由客户端把数据直接写入相应数据节点；在读取时，客户端从名称节点获得数据节点和文件块的映射关系，然后到相应位置访问文件块。数据节点也要根

据名称节点的命令创建、删除数据块和冗余复制。

8.答案要点如下，详见 3.3.2 节

NoSQL 是对非关系数据库的统称，它所采用的数据模型不是传统关系数据库的关系模型，而是类似键值、列族、文档等非关系模型，没有固定的表结构，通常也不存在连接操作，也没有严格遵守关系数据库的 ACID 约束。

按照存储架构设计的不同，NoSQL 数据库可分为键值数据库、列存储数据库、文档数据库和图数据库四大类。键值数据库有 Redis、Amazon DynamoDB、Aerospike 等；列族数据库（也称为列存储数据库）有 HBase、Cassandra、HyperTable 等 ；文档数据库包括 MongoDB、Couchbase、MarkLogic 等；图数据库则有 Neo4j、Infinite Graph 等。几种 NoSQL 数据库的区别见表 3-11。

9.答案要点如下，详见 3.3.3 节

关系数据库以完善的关系代数理论作为基础，有严格的标准，支持事务 ACID 四性，借助索引机制可以实现高效的查询，技术成熟，有专业公司的技术支持；主要不足在于可扩展性较差，无法较好支持海量数据存储，数据模型过于死板、无法较好支持 Web2.0 应用，事务机制影响了系统的整体性能等。

NoSQL 数据库可以支持超大规模数据存储，灵活的数据模型可以很好地支持 Web2.0 应用，具有强大的横向扩展能力等；主要不足在于缺乏数学理论基础，复杂查询性能不高，大都不能实现事务强一致性，很难实现数据完整性，技术尚不成熟，缺乏专业团队的技术支持，维护较困难等。

NewSQL 是对各种新的可扩展、高性能数据库的简称，这类数据库不仅具有 NoSQL 对海量数据的存储管理能力，还保持了传统数据库支持的 ACID 和 SQL 等特性。不同的 NewSQL 数据库的内部结构差异很大，但是它们有两个显著的共同特点都支持关系数据模型；都使用 SQL 作为其主要的操作接口。

10.答案要点如下，详见 3.3.2 节

HBase 采用表来组织数据，表由行和列组成，列划分为若干个列族。表中每个行由行键（Row Key）来标识。一个 HBase 表被分组成许多“列族”（Column Family）的集合，列族里的数据通过列限定符（列）来定位。通过行、列族和列限定符确定一个“单元格”（Cell），单元格中存储的数据被视为字节数组 byte。多个 Cell 在一起组成一个存储（Store）,划分的依据是列族。多个 Store 构成一个区域（Region），一张表可以是多个 Region，划分的依据是行键。每个单元格都保存着同一份数据的多个版本，这些版本采用时间戳进行索引，HBase 中需要根据行键、列族、列限定符和时间戳来确定一个单元格，因此，可以视为一个“四维坐标”，即行键、列族、列限定符、时间戳。

11.答案要点如下，详见 3.3.2 节

文档数据库是围绕一系列语义上自包含的文档来组织数据管理，文档没有模式，也就是说并不要求文档具有某种特定的结构。一个文档数据库实际上是一系列文档的集合，其中每个文档是一个数据记录，这个记录能够对包含的数据类型和内容进行“自我描述”，XML 文档、HTML 文档和 JSON 文档就属于此类。每个文档所包含的内容是一系列数据项的集合，每个数据项都是一个键-值对的组合，该值既可以是简单的数据类型如字符串、数字和日期等，

也可以是复杂的类型，如有序列表和关联对象。举例略。

12.答案参见

DB-Engines 网址：https://db-engines.com/

13.答案要点如下，详见 3.3.5 节

大数据 SQL 查询引擎是大数据的访问和查询接口，遵循类 SQL 的语法，服务于非关系型数据库或其它分布式处理系统。SQL 并不适用于解决所有大数据问题，例如，它并不适合用来开发复杂的机器学习算法，但是它对很多分析任务非常有用，而且它的另一个优势是工业界非常熟悉它。利用 SQL 查询引擎处理一些大数据分析问题，可以在一定程度上降低人员学习成本，缩短项目开发周期。

HiveQL 是一中简单的、类似 SQL 的查询语言，它与大部分 SQL 语法兼容，但是并不完全支持 SQL 标准，比如，HiveQL 不支持更新操作，也不支持索引和事务，它的子查询和连接操作也存在很多局限。

14.答案

(1) select 学号,姓名 from Student
where 专业=‘数据科学与大数据技术’ and 性别=‘男’

(2) select Score.学号,姓名,成绩 from Student,Score
where Student.学号=Score.学号 and 专业=‘计算机科学与技术’
and 课程=‘高等数学’and 成绩 > 85

(3) select 专业,CONVERT(DECIMAL(13.2),AVG(成绩)) as 高级语言程序设计平均分
from Student,Score where Student.学号=Score.学号 group by 专业
order by 高级语言程序设计平均分 desc

第 4 章课后习题

1．答案要点如下，详见 4.1 节

数据挖掘是大数据分析的重要技术之一。大数据分析最大的价值在于对大量不相关的各种类型的数据进行分析，挖掘出对未来趋势与模式预测分析有价值的信息和知识。数据挖掘就是人们常说的知识发现，通过对海量的、杂乱无章的、不清晰的并且随机性很大的数据进行挖掘，从而找到其中蕴含的有规律、有价值并能够理解和应用的知识。

2.答案要点如下，详见 4.2.1 小节

数据集中趋势的重要指标有平均数、中位数、众数。

均值是数据集中趋势中最常用的一个测度值，用于数值型数据，它是一组数据的均衡点所在，是数据误差相互抵消后的必然结果，优点是作为一组数据的代表,比较稳定、可靠，，因为它与每一个数据都有关,反映出来的信息最充分。另外均值计算公式简单，易于理解，应用较广泛，缺点是易受极端值的影响。

中位数是将一组数据按大小顺序依次排列后,位于正中间的一个数据或正中间两个数据的平均数。优点是不受极端数据的影响，缺点是可靠性较差，因为它只利用了部分数据。

众数是一组数据中出现次数最多的数据，众数的优点也是不受极端数据的影响，缺点是可靠性较差。

3．答案要点如下，详见 4.2.2 小节

数据的极差、四分位距、方差和标准差是数据离散程度的常用指标。

极差指一组数据中最大值和最小值之差。是对整体数据离散程度的度量。优点是计算简单，含义直观，运用方便。缺点是不能反映中间数据的分布情况，不能准确描述出数据的分散程度，易受极端值的影响。

四分位距指上四分位数和下四分位数之差,四分位距反映了中间 50% 数据的离散程度。优点是不受极端值的影响，缺点是只反映了中间数据的离散程度。

方差是一组数据中，所有个体与总体均值距离的平方和的均值。优点是使用了所有数据信息，衡量了数据整体的离散情况。缺点是与处理数据的量纲不一致，无法直观比较相对波动情况。

标准差是方差的算术平方根，优点与处理数据的量纲一致，易于比较。

4．答案要点如下，详见 4.1.2 小节

数据挖掘流程可以分为以下阶段 ：确立挖掘目的、数据处理、数学建模、模型验证和评估、模型应用。

（1）确立挖掘目的：对目标做简单评估，确立所需要的数据类型，获取数据集。数据集的选取对数据挖掘起决定作用。数据挖掘通常有关联、分类、回归、聚类等方法，不同的方法代表着数据挖掘的某种目的。

（2）数据处理：在获得了原始数据之后，进行数据的清洗，数据预处理，特征选择等。

（3）数学建模：应用各种建模技术对数据进行分析和挖掘，建模往往是一个螺旋上升不断优化的过程，不是一成不变的模型。因此，需要对模型进行结果分析，如果效果不佳，

则需要调整模型参数，对模型进行优化。

（4）模型验证和评估：在完成了建模后，需要对建立的数学模型进行验证，以评估其效果。

（5）模型应用：在完成数据挖掘的工作后，可以将模型投入使用以解决问题。

5．答案要点如下，详见 4.3.4 小节

一元线性回归自变量只有一个，影响因变量的因素只有一个。多元线性回归影响因变量的因素有多个，即因变量与多个自变量相关。

线性回归即因变量可以表示为多个自变量的线性组合，表达形式为 $Y = X\beta + \varepsilon$，非线性回归模型的因变量是自变量的一次以上函数形式，回归规律在图形上表现为形态各异的各种曲线。

6．答案要点如下，详见 4.3.3 小节

利用数据挖掘技术方法构建的学习模型学习能力过强，训练集将自身特性当作所有潜在样本都会具有的一般性质，泛化能力下降，即训练误差小，泛化误差大。这种现象称为过拟合。

7．答案要点如下，详见 4.3 节

回归、分类、聚类方法是数据挖掘的三大任务，根据处理的数据特点，数据挖掘方法主要有两种 ：一种是针对带标注数据集的有监督数据挖掘方法，一种是针对没有标注数据集的无监督的数据挖掘方法。有监督的数据挖掘方法主要包括分类分析和回归分析。分类和回归分析都需要找到数据之间的依赖关系，并且进行预判输出，但分类分析输出的是离散类别值，回归分析预测的是连续值。聚类分析是无监督的数据挖掘方法，根据指定的聚类标准将数据集聚成不同的簇，簇内数据尽量相似，簇间数据尽量不同。

8.答案要点

利用数据分析工具的描述统计功能，数据汇总结果如下表，具体步骤参看本章 4.4 节

总人口(万人)	
平均	4220.226
标准误差	485.7905
中位数	3793
众数	#N/A
标准差	2704.767
方差	7315763
峰度	-0.53646
偏度	0.537758
区域	9257
最小值	287
最大值	9544
求和	130827
观测数	31

其中因为各地区人口总数没有重复值，数据唯一，所以众数为空（#N/A）

9.答案要点

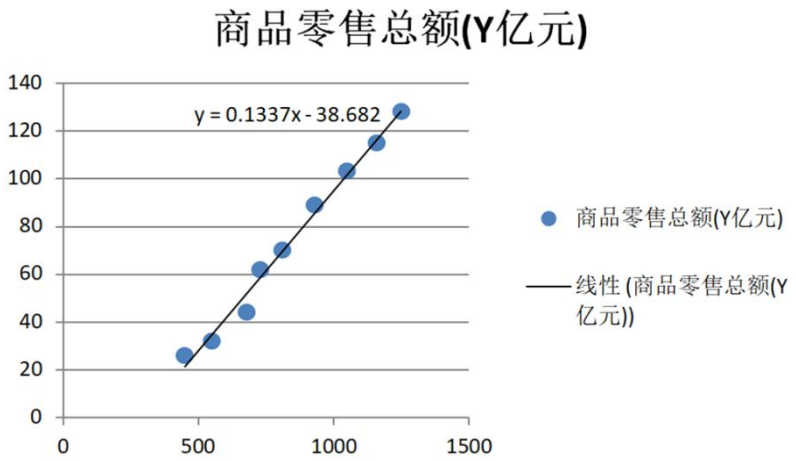
利用数据分析工具求人均收入和商品零售总额的相关系数，具体步骤参看本章 4.4 节人均收入和商品零售总额的相关系数为 0.994，说明人均收入和商品零售总额高度正向相关。

绘置散点图再添加趋势线的方法建立以商品零售总额为因变量，人均收入为自变量的一元线性回归模型如下：

$$y = 0.1337x - 38.682$$

则当人均收入 x=1300 时，y=135.128。

绘置散点图再添加趋势线的结果如下图，具体操作步骤参看本章 4.4.3 节。



第 5 章课后习题

1.答案要点

数据可视化就是将结构或非结构化的数据转换成适当的可视化图表，将隐藏在数据中的信息直接展现在人们面前，可视化的目标就是“让数据说话”，将数据的价值更好发挥出来。

2.答案要点

（1）数据分为四个大类：分类数据、量化数据、时间数据和地理数据：

（2）数据之间的关系通常分为 7 类：①简单对比、②时间序列、③相关性、④分级排序、⑤偏差性、⑥分布情况、⑦局部与整体。

（3）数据对可视化的作用：要想把数据可视化，必须要知道它表达的含义，知道它所代表的事物以及和这些事物之间的联系，这是因为可视化目的是让我们从另一个角度去探索数据本身所蕴含的价值。也就是说，在进行数据分析、数据图表绘制或数据可视化之前，我们应该要对数据类型以及数据之间的关系有所了解。在此基础上才可以选择合适的图表进行数据可视化。这样才能更好地将数据的价值发挥出来。

3.答案要点

按照数据所属的不同类型，数据可视化可以分为统计数据可视化、关系数据可视化、地理空间数据可视化、时间序列数据可视化以及文本数据可视化，而可视化可以有统计图表、标签云、热力图、地图、仪表盘等表达方式，它们能将图形更直观地展现的用户面前。

4.答案要点

（1）常用的数据可视化工具有：Microsoft Excel、Tableau、魔镜、Gephi 等；

（2）编程可视化语言有：R、python、Processing 等可视化编程语言

（3）Web 可视化技术有： D3.js、Highcharts、Echarts 等关键技术。

5. 答

略

第 6 章课后习题

1.答案要点

区别：

集中式是指由一台或多台主计算机组成中心节点，数据集中存储于这个中心节点中，并且整个系统的所有业务单元都集中部署在这个中心节点上，系统的所有功能均由其集中处理。

分布式系统就是一群独立计算机集合共同对外提供服务，但对系统用户来说，就像是一台计算机在提供服务一样。分布式意味着可采用更多普通计算机组成分布式集群对外提供服务。

应用场景：

集中式计算架构包括大型主机和超级计算机。就大型主机而言，集中式架构多用于传统的银行、电信、交通、医疗等行业，集中式架构下，包括操作系统、中间件、数据库等“基础软件”均为闭源商用系统。集中式架构的典型案例是 IOE（IBM， Oracle，EMC）提供的计算设备、数据库技术和存储设备共同组成的系统。就超级计算机而言，在计算科学领域发挥着重要的作用，用于各个领域的计算密集型任务中，包括量子力学、天气预报、气候研究、石油和天然气勘探、分子建模、核武器爆炸和核聚变模拟。

分布式计算架构应用广泛，例如 Yahoo，Facebook，Amazon 以及国内的百度，阿里巴巴等众多互联网公司都以 Hadoop 为基础搭建自己的分布式计算系统，应用于海量数据的离线分析处理、大规模 Web 信息搜索 、数据密集型并行计算；Twitter 主推 Storm 分布式计算系统，应用于实时分析、在线机器学习、持续计算、分布式远程调用等领域； Spark 的主要应用于推荐系统、实时推荐、交互式实时查询，如优酷土豆广泛使用 Spark 实现视频推荐(图计算)、广告业务等。

2.答案要点

目前来说，大数据领域最为活跃的三个计算框架为 Hadoop、Spark 以及 Flink。三个框架在不同的大数据处理场景当中，表现各有优势，下面对三个框架进行对比分析。

Hadoop 专为批处理而生，一次将大量数据集输入到输入中，进行处理并产生结果。Hadoop 提供可配置的内存管理，可以动态或静态地执行此操作。MapReduce 采用面向批处理的模型，批处理静态数据，不支持处理流数据。MapReduce 计算数据流没有任何循环，每个阶段使用上一阶段的输出，并为下一阶段产生输入。Hadoop 与 Spark 和 Flink 相比，性能低。

Spark 定义是一个批处理系统，但也支持流处理。Spark 提供可配置的内存管理，从 Spark 1.6 开始已朝着自动进行内存管理的方向发展。Spark 采用微批处理，微批处理本质上是一种“先收集再处理”的计算模型。Spark Streaming 以微批处理数据流，实现准实时的批处理和流处理。尽管机器学习算法是循环数据流，但 Spark 将其表示为有向无环图(DAG)。Spark 支持微批处理，但流处理效率不如 Flink。

Flink 为流和批处理提供了一个运行时。Flink 有自己的内存管理系统，提供自动内存管理。Flink 是真正的流引擎，使用流来处理工作负载，包括流，SQL，微批处理和批处理。Flink 采用连续流式流传输模型，实时对数据进行处理，而不会在收集数据或处理数据时出现任何延迟。Flink 在运行时支持受控循环依赖图，支持机器学习算法非常有效。Flink 使用本机闭环迭代运算符，尤其在支持机器学习和图形处理方面，表现优异。

3.答案要点

Hadoop是Apache软件基金会下用Java语言开发的一个开源分布式计算平台，实现在大量计算机组成的集群中对海量数据进行分布式计算。Hadoop框架核心包括：分布式文件系统HDFS、分布式计算框架MapReduce和资源管理系统YARN。

4.答案要点

HDFS 具有高容错性、高可靠性、高可扩展性、高吞吐量等特征，主要表现在以下几方面：(1)硬件故障。HDF 核心的设计目标就是检测硬件故障和自动快速恢复，它可以实现持续监视、错误检查、容错处理和自动恢复，从而保证数据的完整性。(2)数据访问。HDFS 提高了数据吞吐量，满足了批量数据处理的设计要求。(3)简单一致性模型。HDFS 提供“一次写入、多次读取”的服务，文件一旦创建、写入、关闭之后就不需要修改了，只能被追加或读取，简单化了数据一致的问题，便于提供高吞吐量的数据访问。(4)大数据集。HDFS 支持处理超大规模文件，通常可以达到是 GB 甚至 TB 级，一个由数百台机器组成的集群可以支持千万级别的文件。

HDFS 用于大数据领域的数据存储。依据以上特点， HDFS 适用于大文件、大数据处理，处理数据达到 GB、TB、甚至 PB 级别的数据；适合流式文件访问，一次写入，多次读取；文件一旦写入不能修改，只能追加。

5.答案要点

在 HDFS 中，采用数据块作为最基本的存储单位，默认大小为 128MB（数据块最早默认大小是 64 MB，从 2.7.3 版本开始，官方关于 Data Blocks 的说明中，block size 变成了 128 MB）。

6.答案要点

HDFS 采用主从(Master/Slave)结构模型，一个 HDFS 集群是由一个名称节点(NameNode，NN)和若干个数据节点(DataNode，DN)组成(最新的 Hadoop 版本有多个 NameNode 的配置)。NameNode 作为主服务器，管理文件系统命名空间和客户端对文件的访问操作。DataNode 负责数据的存储和读写。

7.答案要点

在 MapReduce 中，一次计算主要分为 Map（映射）和 Reduce（归约）两个阶段，输入和输出由 HDFS 分布式文件系统进行存储。Map Reduce 处理流程包括以下基本步骤。第一步，数据分片。将输入文件切分成多个分片，每一个分片都会复制多份到 HDFS 中，即 MapReduce 通过数据分片的方式将一个大任务拆分成若干个小任务。第二步，数据映射。完成数据分片后，MapReduce 通过 InputFormat 从文件的输入目录中读取数据记录，然后一个 Mapper 处理一个数据分片，每个 Mapper 对相应分片中的每一行记录进行解析处理，根据用户自定义的映射规则重新组织生成一系列的<Key,Value>对作为输出的中间结果。第三步，数据混洗。Shuffle 从 Mapper 处获取中间结果，并通过排序、合并、归并等操作将这些中间结果按照相同的 Key 汇集排序生成<Key,Value-list>形式的中间结果，从而把无序的<Key,Value>变成有序的<Key,Value-list>，便于 Reducer 并行处理。第四步，数据归约。Reducer 获取一系列的<Key,Value-list>中间结果后，按照用户定义的逻辑进行汇总和映射，得到最终计算

结果，并由 OutputFormat 把结果输出到文件系统。

8.答案要点

Hadoop 是一个开源分布式计算平台，实现在大量计算机组成的集群中对海量数据进行分布式计算。Hadoop1.0 框架核心由 HDFS 和 MapReduce 两个系统组成，而 Hadoop2.0 的框架核心由 HDFS、MapReduce 和 YARN 三个系统组成。Hadoop1.0 和 Hadoop2.0 最大的区别就是新增了 YARN 作为资源管理系统。YARN 是一个纯粹的通用的资源管理调度框架，为上层应用提供统一的资源管理和调度，支持多种计算框架并存。不仅限于 MapReduce 一种框架，也可以为其他框架使用，如 Tez、Spark、Storm 等。YARN 的引入为 Hadoop 集群的利用率、资源的统一管理和数据的共享等方面带来了极大的提升。

9.答案要点

MapReduce1.0 计算框架主要由三部分组成：编程模型、数据处理引擎和运行时环境。它的运行时环境由一个 JobTracker 和若干个 TaskTracker 两类服务组成，其中 JobTracker 负责资源管理和作业调度，TaskTracker 负责接收来自 JobTracker 的命令并执行分配的任务，定期向 JobTracker 发送心跳信息和资源使用情况等。

Hadoop2.0 将 JobTracker 中的资源管理和作业控制分开，把 MapReduce1.0 体系结构重新设计生成 YARN 和 MapReduce2.0。MapReduce2.0 运行于资源管理框架 Yarn 之上，运行时环境不再由 JobTracker 和 TaskTracker 等服务组成，而是变为通用的资源管理系统 Yarn 和作业控制进程 ApplicationMaster，其中 Yarn 负责资源管理的调度，ApplicationMaster 负责作业的管理。

Yarn 系统的引入，克服了 MapReduce1.0 架构中集群只有一个 JobTracker 存在单点故障隐患、节点压力大、不易于扩展等问题。同时 YARN 支持多种计算框架并存，相比于 MapReduce1.0 架构只支持 MapReduce 作业，避免了应用中根据不同的需求同时搭建 Hadoop、Spark 等多个集群，并解决了由此造成的集群管理复杂、资源利用率低、跨集群数据共享成本增加等问题。

10.答案要点

Spark 是借鉴了 MapReduce 并在其基础上发展起来的，继承了其分布式计算的优点并改进了 MapReduce 明显的缺陷，二者的区别如下：

(1) Spark 把运算的中间数据存放在内存，不再需要读写 HDFS，迭代计算效率更高，能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法；MapReduce 的中间结果需要保存到磁盘，这样必然会有磁盘 IO 操作，影响性能，因此适合日志分析挖掘等较少的迭代的长任务需求。

(2)Spark 通过 RDD 实现高效容错。RDD 是一组分布式的存储在节点内存中的只读性质的数据集，这些集合是弹性的，某一部分丢失或者出错，可以通过整个数据集的计算流程的血缘关系实现重建；MapReduce 的容错可能是重新计算，成本较高。

(3)Spark提供了更丰富的计算API，除了Map和Reduce操作之外，还延伸出了Filter、FlatMap、Count、Distinct等操作，另外还有流式处理Sparkstreaming、图计算GraphX等等；MapReduce只提供了Map和Reduce两种操作，不支持流计算以及其他模块。

(4) Spark框架及其生态相对复杂，很多时候spark作业需要根据不同业务场景的需要进行调优以达到性能要求；MapReduce框架及其生态相对简单，对性能的要求相对较弱，但运行较稳定，适合长期后台运行。

总之，Spark 生态更为丰富，功能更为强大、性能更佳，适用范围更广；MapReduce 更简单、稳定性好、适合离线海量数据计算。

11.答案要点

Spark框架包括三层：资源管理层、Spark核心层和服务层。

底层的资源管理层主要提供资源管理功能，工作一般由YARN、Mesos、Standalone等资源管理器完成。资源层主要涉及两种角色——Cluster Manager（集群管理器）和Worker Node（工作节点），Spark用户的应用程序在一个工作节点上有一个Executor（执行器），执行器内部通过多线程的方式并发处理应用的任务。

中间的核心层主要提供内存计算框架，实现Spark的基本功能，包含任务调度、内存管理、错误恢复、与存储系统交互等模块。Spark Core中还包含了对弹性分布式数据集的API定义。Spark 核心是建立在统一的抽象弹性分布式数据集RDD之上的，这使得Spark的各个组件可以无缝地进行集成，能够在同一个应用程序中完成大数据处理。

上层的服务层主要是面向特定类型的计算服务，提供一站式解决平台，如SQL查询（Spark SQL）、实时流处理（Spark Streaming）、机器学习（MLLib）以及图计算（GraphX）等。

12.答案要点

RDD的操作算子分为两种类型:Transformation(转化)和Action(行动)。

Transformation操作就是从一个RDD产生一个新的RDD。Action操作会对 RDD 计算出一个结果，并把结果返回到驱动器程序中，或把结果存储到外部存储系统（如 HDFS）中。在RDD执行过程中，真正的计算发生在行动操作。当RDD执行转化操作时，实际计算并没有被执行，只有当RDD执行行动操作时才会促发计算任务提交，从而执行相应的计算操作。

13.答案要点

RDD的血缘关系即DAG拓扑排序的结果，描述了一个RDD是如何从父RDD计算得来。

在血缘关系中，下一代的RDD依赖于上一代的RDD。RDD通过血缘关系记住了它是如何从其他RDD中演变过来的。当某个RDD的部分分区数据丢失时，它可以通过血缘关系获取足够的信息来重新运算和恢复丢失的数据分区。这种依赖关系设计，使Spark具有了天生的容错性，大大加快了Spark的执行速度，从而带来性能的提升。

14.答案要点

大数据处理技术一般包括：大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用。针对于大数据计算平台，下面从大数据存储、管理、计算引擎三方面分别介绍大数据处理技术的发展历程及应用。

（1）数据存储系统

HDFS 源于 Google 在 2003 年发表的 GFS(Google File System)论文，是 Hadoop 体系的基础，负责数据的存储与管理。 它提供一次写入多次读取的机制，具有高容错性、流式处理、处理海量数据等优点。

Hbase 源于 Google 在 2006 年发表的 Bigtable: ADistributed Storage System for Structured Data 论文。它是一个分布式的、面向列的开源数据库，采用 HDFS 作为其底层数据存储，同时利用 MapReduce 处理 Hbase 中保存的海量数据，实现了数据存储与并行计算的完美结合，满足了大数据应用中快速随机访问海量数据（PB 级）并及时响应用户的需求。

Hive 最初是应 Facebook 每天产生的海量新兴社会网络数据进行管理和机器学习的需求而产生和发展的。它是一个基于 Hadoop 的数据仓库工具，可以对 Hadoop 中的大规模数据进行数据整理、查询和分析存储。Hive 提供了类似于 SQL 的查询语言 Hive SQL，使得不熟悉 MapReduce 的用户可以很方便地利用 SQL 语言查询、汇总和分析数据，降低了 Hive 的学习门槛。

Sqoop 项目开始于 2009 年，最早是作为 Hadoop 的一个第三方模块存在，后来为了让使用者能够快速部署，也为了让开发人员能够更快速的迭代开发，并在 2013 年独立成为 Apache 的一个顶级开源项目。Sqoop 用于数据在关系数据库（如 MySQL、Oracle 等）与 HDFS、Hive、Hbase 间的相互导入导出，便于关系数据库和 Hadoop 之间的数据迁移。

随着互联网技术的发展，人们对网络日志潜在的信息越来越重视，所以就想把这些日志收集起来，然后进行分析。但是大量的日志产生的位置比较分散，存储的目的地也很不一致，这就导致了数据采集的复杂性。Flume 就是满足这种日志采集需求的一个软件框架，用于分布式的海量日志采集、聚合和传输。

Pig 是由 Yahoo 公司开源设计的一个基于 Hadoop 的大规模数据分析平台，用于分析较大的数据集，可以在 Hadoop 中执行所有的数据处理操作，需要程序员使用 Pig Latin 语言编写相应的数据分析脚本，有利于不熟悉 Java 语言的程序员进行大数据分析处理。

（2）资源调度管理系统

Zookeeper 源自 Google 于 2006 年 11 月发表的 Chubby 论文，是为大型分布式系统提供开源、高效、可靠的协同工作系统。它可以解决分布式环境下的数据管理问题（如统一命名、状态同步、集群管理、配置同步等），运行在计算机集群上，用于管理 Hadoop 操作，因此，Hadoop 的很多组件都依赖它。

Oozie 是一个开源的工作流调度系统，可以调度 MapReduce、Pig、Hive、Spark 等不同类型的单一或具有依赖性的作业。当一个业务分析场景中需要多个 Hadoop 工作协同完成时，Oozie 可以把它们按照指定的顺序协同运行起来。

YARN 是在第一代 MapReduce 基础上演变而来的资源管理器，是为了解决原始 Hadoop 扩展性较差、不支持多计算框架的问题。YARN 是一个通用的资源管理系统，为上层各类计算框架（如 MapReduce、Spark、Storm 等）提供资源管理和调度，它将资源管理与作业调度/监控分离，提高了集群的利用率，为集群的资源统一管理和数据共享等方面带来了巨大好处。

（3）计算引擎或计算模型

MapReduce 源于 Google 在 2005 年发表的 MapReduce: Simplified Data Processing on Large Clusters。它是一种基于磁盘的分布式并行批处理计算模型，用于大规模批处理静态数据，不支持处理流数据。

Spark 于 2009 年诞生于美国加州大学伯克利分校的 AMP 实验室，它是基于内存计算的大数据并行计算框架。在 Spark 生态圈中包含了 Spark SQL、Spark Streaming、GraphX、MLlib 等组件，可用于构建大型的、低延迟的数据分析应用程序，在互联网企业中应用非常广泛。

Flink 起源于 2008 年柏林理工大学一个研究性项目 Stratosphere，2014 年 Stratosphere 成为 Apache 孵化项目，从 Stratosphere 0.6 开始，正式更名为 Flink。Flink 是混合框架，它完全支持流处理，批处理被作为一种特殊的流处理，而且只有 Flink 实现了毫秒级低延迟的实时流数据计算。

第 7 章课后习题

多选题答案：

1．ABC　　2．ABCD　　3．ABCD

4．ABCD　　5．ABCD　　6．ABC

简述题答案：

1．答案要点如下，详见 7.1.1 小节

农业大数据：指运用大数据的理念、技术和方法来指导现代农业的发展，解决农业或涉农领域关于数据采集、计算、存储以及生产的实际应用技术。

它对于现代化农业生产有什么样的意义：利用农业大数据实现农业产业可持续发展和产业结构优化，加快农业自动化、信息化、智能化进程，需要依托农业大数据及相关大数据处理分析技术。建设农业大数据支撑平台，全面、及时、前瞻性地反映农业发展动态，预测农业未来发展方向，可为政府、企业及农业从业人员提供决策管理支持。

2．答案要点如下，详见 7.2.1 小节

教育大数据对教育产生影响的主要方面：（1）改变教育研究中对数据价值的认识，推动产业智能化升级。（2）帮助学校、教师全面的了解学生精准实施教学管理。（3）帮助学生进行个性化、高效学习。

3．答案要点如下，详见 7.3.3 小节

社交大数据在识别黑灰色产业从业者时的主要过程：一方面，通过学习社交平台的社交信息及话题训练用户分类模型，从而建立黑灰产业从业者画像，根据该画像从平台用户识别黑灰产从业者；另一方面结合用户行为特征，对平台用户危险行为预测，为执法部门提供线索。

4．答案要点如下，详见 7.4 节

旅游大数据：指运用大数据的理念、技术和方法来指导旅游业的发展，解决旅游业及相关领域关于数据采集、计算、存储以及生产的实际应用技术。

它对游客、旅游行业及政府管理的意义：实现游客个性化定制旅游、提升旅游体验，帮助旅游行业提升旅游服务水平、扩展旅游营销手段、升级旅游管理和探索旅游创新，方便政府探查检测公共服务能力水平、推动协同管理、为区域经济发展整合优化资源配置。

5．答案要点如下，详见 7.5.1 小节

交通大数据的主要应用方向有：助力“智慧城市”建设，推动城市交通数据治理，进行交通拥堵预测及精准干预，保障用户便捷安全畅通出行，提升城市管理水平与服务水平等；助力“智慧铁路”建设，根据铁路系统及沿线监测数据，分析诊断故障，进行安全预警，保障铁路安全运营。例如杭州综合交通云服务平台、城市交通拥堵治理、京张“智慧”高铁等。

6．答案要点如下，详见 7.6 小节

金融大数据在风险评控方面的主要应用有：制定风险管理策略与风险管理模型，实现有效的风险管理；根据不同的评分模型对客户风险进行评价，规避一系列可能的风险。如“度小满金融”、“蚂蚁金服”等小额助农贷款业务，通过用户个人画像分析降低申请风险；使用行为风险评分模型对低分段采取止付、降额等措施提前防范风险，降低可能发生的坏账损失；使用欺诈风险评分模型识别欺诈申请、伪卡交易和非法套现等。