# 第 3 章　大数据技术

课件：大数据技术

## 学习目标

- 掌握大数据的概念。
- 了解大数据的特征。
- 掌握大数据在各行业的典型应用。

## 案例导读

我们的衣食住行都与大数据有关，每天的生活都离不开大数据，每个人都被大数据包裹着。大数据提高了我们的生活品质，为每个人提供了创新平台和发展机遇。

大数据通过数据整合分析和深度挖掘，发现规律、创造价值，进而建立起从物理世界到数字世界再到网络世界的无缝链接。在大数据时代，线上与线下、虚拟与现实、软件与硬件跨界融合，重塑了我们的认知和实践模式，开启了一场新的产业突进与经济转型。《大数据时代》的作者 Viktor Mayer-Schönberger 这样定义大数据，“大数据是人们在大规模数据的基础上可以做到的事情，而这些事情在小规模数据的基础上是无法完成的。”

### 【案例 1】地震预测大数据

每年，地震在全球范围内都会导致超过 1.3 万人死亡，500 万人受伤或财产受损，造成的经济损失高达 120 亿美元。多年以来，科学家们主要依靠对震频的监测来预测地震。尽管有很多潜在的地震预警信号，如大气条件的变化或大量蛇群的迁移，但基于这些信号做出的预测准确率太低，无法在现实中应用。

科学家们利用大数据技术对来自卫星和气象领域的数据进行统计分析，开启了一种全新的地震预测模式。该项技术可以帮助人类提前 30 天预测全球主要地震多发国家即将发生的震级 6 级以上的大地震，精准度已达 90%。曾提前 9 天预测到了 2015 年 3 月 3 日在印度尼西亚发生的 6.4 级地震。地震预测大数据如图 3-1 所示。

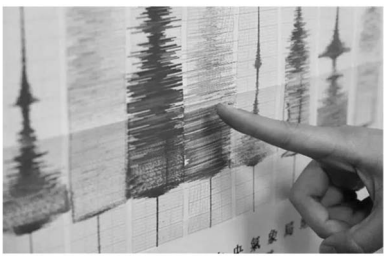

图 3-1　地震预测大数据

### 【案例 2】山东省淄博市高青县：数字特产商城带动“亮村共富”

“以数智点亮乡村，带动产业发展，推动乡村振兴”，高青县紧紧抓住用好农业数字时

代的重大机遇，立足农业资源禀赋和产业化优势，凝心聚力推进数字乡村体系建设，以数字技术改造升级农业全链条、农村各领域和农民新生活，推动农业向规模化、高端化、绿色化及智慧化转型升级。

高青县以农业农村大数据平台为基础，服务经营主体和村民。使用大数据平台，获取经营主体信息、生产信息、种植环境信息、土地利用信息、农作物长势信息，以及农业投入品、农机使用情况等数据，对农业产业的整体情况做实时、动态分析，为经营主体提供适合农作物生长及市场需求的种植建议，运用现代科技帮农民把地种好、把农产品卖好。

通过大数据分析，反映消费群体对优质农产品的购买需求和购买能力，以及喜欢的购买渠道和方式，让生产者看到优质农产品带来的经济效益，以市场和消费者认同的方式开展标准化生产，降低生产风险，提高产品价值，促进农业产业发展。高青县农业农村大数据平台如图 3-2 所示。


图 3-2　高青县农业农村大数据平台

## 【案例 3】南京高校“科技原创力”，追着害虫“跑”，用大数据预测迁飞趋势

“这个变化，对于江苏农田算是利好。”近日，南京农业大学胡高教授团队联合全国测报体系在国际著名生态学期刊《全球变化生物学》（英文名 *Global Change Biology*）上发表研究论文，揭示在全球变暖背景下，降水和风场条件的变化致使我国褐飞虱迁飞模式发生转变，为迁飞害虫的准确测报和科学防控提供了重要理论参考，为推动农业强国、助力乡村全面振兴、保障粮食安全做出积极贡献。

南京农业大学胡高教授团队基于 1978—2019 年全国 300 多个站点的稻飞虱监测数据和相关气象资料发现，自 2001 年以来，影响我国夏季盛行气流和降水时空分布的重要大气环流系统西太平洋副热带高压（简称副高）强度显著增强，位置明显西移。受此影响，我国长江以南地区夏季西南气流显著变弱、降水增加，江淮地区降水显著减弱，不利于褐飞虱的远距离迁飞，致使华南地区 7 月迁出褐飞虱的迁飞距离显著变短，长江下游地区褐飞虱迁入量显著下降。本次研究发现，由于褐飞虱迁飞模式的转变，长江下游不再成为褐飞虱 7 月份迁飞的主降区。“对于江苏包括南京来说，这个研究发现是好消息。对于害虫的防控，

依托完整的网络系统，这些年，江苏的褐飞虱虫害确实较少。”胡高说。胡高教授团队在稻田里做研究如图 3-3 所示。

图 3-3　胡高教授团队在稻田里做研究

随着信息科技的不断发展，通过网络对信息进行获取、存储、处理和传递的方式越来越普及、越来越便捷，随之产生的数据量越来越庞大，获取的数据越来越重要，一个崭新的时代正悄然来临。世界正从信息时代迈向大数据时代，数据挖掘与分析等大数据技术所展现的巨大价值，正激发大众对大数据孜孜不倦地探索。

**【案例 4】亚马逊公司利用大数据预测消费者特征**

随着互联网的快速发展和数字经济的日益繁荣，大数据和人工智能（AI）已经成为企业制定营销策略的重要辅助工具。亚马逊作为全球领先的电子商务平台，其营销策略紧密结合了大数据和 AI 技术，实现了精准的目标客户定位和个性化的营销推广。

根据消费者以往的搜索记录和消费记录等大数据，可以推算出消费者的消费偏好、经济水平、消费习惯等，甚至可以从浏览某件商品的时间长短，推断出消费者对某类商品或品牌的青睐程度，进而分析消费者购买某种商品的可能性，当可能性大于某个标准时，亚马逊公司就会预判发货。为了提高预判发货的准确性，降低物流成本，亚马逊公司采取了一些措施。例如，刚上市的畅销商品能吸引大量的消费者购买，这时往往会采用预判发货；对于经常在亚马逊网站购物且购买力较强的消费者，更倾向于预判发货。此外，还会根据消费者浏览商品的时间、购买商品的数量等数据推算其犹豫时间，对于犹豫时间较短的消费者，也会预判发货。基于大数据的消费者行为分析和市场趋势预测，亚马逊可以为用户提供个性化的推荐服务和定制化产品。例如，通过用户的购物历史和浏览行为，可以向用户推荐相关的产品和服务，提高用户满意度和忠诚度。

# 3.1 大数据的概念及由来

视频：大数据的概念及由来

## 3.1.1 大数据是什么

《华尔街日报》将大数据（big data）、智能化生产和无线网络革命称为引领未来繁荣的三大技术变革。世界经济论坛发布的报告指出，大数据为新财富，价值堪比石油。因此，世界各国纷纷将施行利用大数据夺取新一轮竞争制高点的重要举措。

大数据是指使用常用软件工具捕获、管理和处理数据所消耗的时间超出可容忍时间的数据集。大数据是一个体量庞大、数据类别繁多的数据集，并且这样的数据集无法使用传统数据库工具对其内容进行捕获、管理和处理。

Gartner 公司将大数据定义为大容量、高速度和多种类的信息资产，需要使用新处理形式来增强决策力、洞察发现力和流程优化能力。

目前对于大数据没有统一的定义，一般认为大数据是指无法在一定时间范围内使用常规软件工具进行捕获、管理和处理的数据集合，是需要使用新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。大数据泛指大规模、超大规模的数据集，因可从中挖掘出有价值的信息而备受关注，但使用传统方法无法进行有效分析和处理。

## 3.1.2 大数据是怎么来的

### 1. 大数据概念的起源

大数据的概念起源于美国，是在思科、威睿、甲骨文、IBM 等公司的倡议下发展起来的。目前，从 IT 技术到数据积累，都已经发生重大变化。

大数据的名称来自未来学家托夫勒所著的《第三次浪潮》。早在 1980 年，托夫勒就在《第三次浪潮》中热情地将大数据称颂为“第三次浪潮的华彩乐章”。《自然》杂志在 2008 年 9 月推出了名为大数据的封面专栏。从 2009 年开始，大数据成为互联网技术行业中的热门词汇。

最早应用大数据的是麦肯锡（McKinsey）公司对“大数据”进行收集和分析的设想，他们发现各种网络平台记录的海量个人信息具备潜在的商业价值，于是投入大量人力、物力进行调研，在 2011 年 6 月发布了关于大数据的报告，该报告对大数据的影响、关键技术和应用领域等方面进行了详尽的分析。麦肯锡公司在《大数据：创新、竞争和生产力的下一个前沿领域》报告中称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”麦肯锡公司的报告获得了金融界的高度重视，而后逐渐受到了各行各业的关注。

数据不再是社会生产的“副产物”，而是转变为生产资料，是可被二次乃至多次加工的原料，从中可以探索更大的价值。大数据是以数据为本质的新一代革命性信息技术，在数据挖潜过程中，能够带动理念、模式、技术及应用实践的创新。

### 2. 大数据的来源

大数据通常是大小为 PB 或 EB 级的数据集。这些数据集有各种各样的来源，如图 3-4 所示。


图 3-4　大数据的来源

（1）信息科技进步。

人们在社会网络、互联网、健康、金融、经济、交通等方面产生的各类数据，例如，病人医疗记录、视频等信息，呈现爆炸式增长的趋势，如图 3-5 所示。

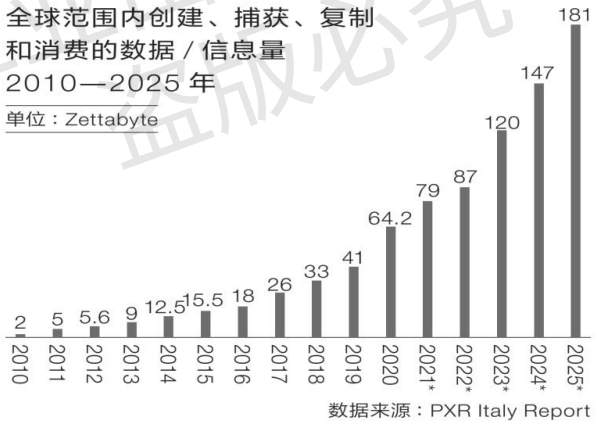


图 3-5　数据爆炸式增长

（2）互联网诞生。

物联网和社交网络的发展，以及智能终端的诞生成为了促进数据爆炸式增长的因素。数据的增长大概遵循摩尔定律。摩尔定律即在信息技术更新换代越来越迅速的情况下，集成电路上的晶体管数量会增加一倍，性能提升一倍，价格降低一半，这是电子工业历史上第一个被发现并获得公认的定律，如图 3-6 所示。随着电子技术和计算机技术的飞速发展，数据总量不断增大，例如，在医疗领域中各类数字设备、科学实验与观察所采集的数据，摄像头不断产生的数字信号，医疗物联网不断产生的人的各项特征值，气象业务系统采集设备所采集的海量数据等，都是大数据的来源。

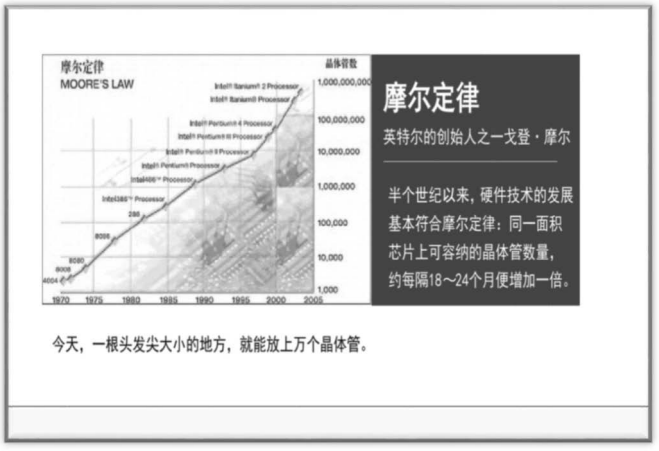


图 3-6　摩尔定律

（3）云计算技术的发展。

云计算一般由数量惊人的计算机群构成，如谷歌数据中心拥有的服务器超过 100 万台，路由器和交换机可以使谷歌数据中心的服务器进行对话，如图 3-7 所示。光纤网络速度是平时家用网速的 20 万倍，云计算可以让普通人体验每秒 10 万亿次的计算能力，如此强大的计算能力，可以模拟核爆炸、预测气候变化和市场发展趋势。

图 3-7　谷歌数据中心

### 3.1.3　大数据的 3V 特征和 5V 特征

从字面上看，大数据这个词可能会让人觉得它只是容量非常大的数据集合而已。但容量大只不过是大数据特征的一个方面，如果只拘泥于数据量，就无法深入理解当前围绕大数据所进行的讨论。因为“用现有的一般技术难以管理”这样的状况，并不仅仅是数据量增大这一个因素造成的。

IBM 提出可以用 3 个特征相结合来定义大数据：Volume（容量大）、Variety（多样性）和 Velocity（速度快），这就是简单的 3V 特征，即容量庞大、种类丰富和速度极快的数据，后来又相继补充了 Veracity（真实性）和 Value（价值密度低）特征，如图 3-8 所示。


图 3-8　大数据的 5V 特征

（1）Volume。

最初提到数据的容量指的是被大数据解决方案所处理的数据容量很大，并且持续增长。数据容量大能够影响数据的独立存储和处理需求，同时还能对数据准备、数据恢复、数据管理的操作产生影响。如今，数据的存储数量正在急剧增长，我们存储的事物包括环境数据、财务数据、医疗数据、监控数据等。有关数据量的量级已从 TB 级转向 PB 级，并且会不可避免地转向 ZB 级。但是，随着可供企业使用的数据量不断增长，可处理、理解和分析的数据的比例却在不断下降。

（2）Variety。

数据多样性指的是大数据解决方案需要支持多种不同格式、不同类型的数据。数据多样性给企业带来的挑战包括数据聚合、数据交换、数据处理和数据存储等。

随着传感器、智能设备及社交协作技术的激增，企业中的数据也变得更加繁杂，因为它不仅包含传统的关系型数据，还包含来自网页、互联网日志文件（包括单击流数据）、搜索索引、社交媒体论坛、电子邮件、文档、主动和被动系统的传感器等方面的原始、半结构化和非结构化数据。

种类表示所有的数据类型。其中，呈爆发式增长的一些数据，如互联网上的文本数据、位置信息、传感器数据、视频等，使用企业中主流的关系型数据库是很难存储的，因为它们都属于非结构化数据。

当然，在这些数据中，有一些是一直存在并得以保存的。和过去不同的是，除了存储，还需要对这些大数据进行分析，并从中获得有用的信息，如监控摄像机中的视频数据。近年来，超市、便利店等零售企业几乎都配备了监控摄像机，其最初目的是防范盗窃，但现在也出现了通过监控摄像机的视频数据来分析顾客购买行为的案例。

（3）Velocity。

数据产生和更新的频率也是衡量大数据的一个重要特征。在大数据环境中，数据产生得很快，在极短的时间内就能聚集起大量的数据集。从企业的角度来说，数据的输入速率代表数据从进入企业到进行处理的时间。要处理快速的数据输入流，需要企业设计出弹性的数据处理方案，同时需要强大的数据存储能力。若想有效处理大数据，则需要在数据变化的过程中对它的数量和种类进行分析，而不只是在它静止后进行分析。

根据数据源的不同，处理速度也有所不同，处理速度不是一直处于高速，如核磁共振扫描图像不会像高流量 Web 服务器日志文件的生成速度那样快。无论速度如何，一分钟内

能够生成的数据都是十分庞大的，如 35 万条推文、可供浏览 300 个小时的 YouTube 视频、1.71 亿份电子邮件，以及 330GB 飞机引擎的传感器数据等。

（4）Veracity。

IBM 在 3V 特征的基础上又归纳总结了第四个“V”——Veracity（真实性）。“只有真实而准确的数据才能让对数据的管控和治理真正有意义。”IBM 软件集团大中华区业务分析洞察及智慧地球解决方案总经理卜晓军，在主题为“大数据·大洞察·大未来”的年度大数据战略发布会上的发言中这样总结。随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限性被打破，企业愈发需要有效的信息治理以确保数据的真实性及安全性。

（5）Value。

IDC（互联网数据中心）称：“大数据是一个貌似不知道从哪里冒出来的大的动力。但是实际上，大数据并不是新生事物。然而，它确实正在进入主流，并得到重大关注，这是有原因的。廉价的存储、传感器和数据采集技术的快速发展，通过云和虚拟化存储设施增加的信息链路，以及创新软件和分析工具，正在驱动着大数据。大数据不是一个‘事物’，而是一个跨多个信息技术领域的动力/活动。大数据技术描述了新一代的技术和架构，其被设计用于：通过使用高速（Velocity）的采集、发现和分析，从超大容量（Volume）的多样（Variety）数据中经济地提取价值（Value）。”

## 3.2 大数据处理

视频：大数据处理

### 3.2.1 大数据处理的基本流程

根据大数据处理的生命周期，大数据技术体系涉及大数据采集与预处理、大数据存储与管理、大数据计算模式与系统、大数据分析与挖掘、大数据隐私与安全等。大数据技术体系如图 3-9 所示。

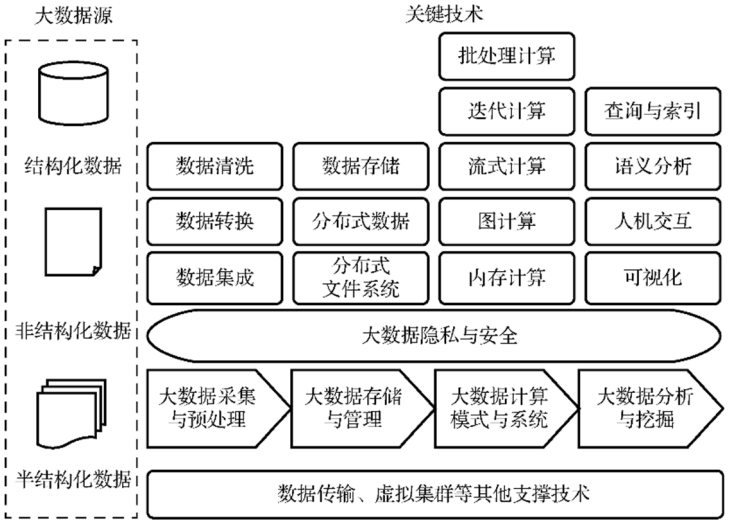


图 3-9 大数据技术体系

一般而言，大数据可以通过 4 个基本步骤进行处理，如图 3-10 所示。

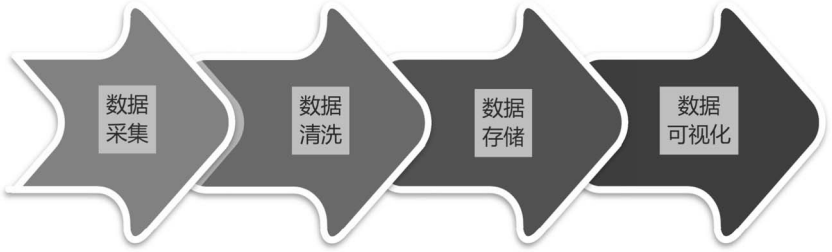


图 3-10　大数据处理的基本步骤

大数据处理的模型也可以被认为是“数据→信息→知识→智慧”的金字塔模型，这是一个量级由大到小、价值由低到高的数据模型，如图 3-11 所示。


图 3-11　大数据处理的金字塔模型



（1）数据采集。

数据采集（数据获取）是大数据处理的第一个步骤，为大数据处理收集足够的、未经加工的原始数据。数据来源包括内部自有数据和外部他营数据。

数据采集一般分别为 DPI 采集、系统日志采集、网络数据采集和其他数据采集。目前很多公司都有自己的海量数据采集工具，均满足每秒数百兆字节的采集和传输需求。

（2）数据清洗。

对海量数据进行分析时，需要将原始数据导入一个大型的分布式数据库中，并对其做一些简单的数据清洗和预处理工作。如果没有经过数据清洗，直接将原始数据交给大数据系统进行处理则可能产生错误，因此数据清洗在整个数据处理的过程中具有非常重要的地位。

（3）数据存储。

在现在的大数据处理中，海量数据的存储是一门重要的学科，其研究目标包括如何有效解决物理存储媒介的问题。数据存储一方面要求良好的物理硬件支持，从而保证海量数据可以被接纳；另一方面需要为处理完毕的数据建立方便访问的服务（如建立索引），从而保证数据可以被快速、准确地访问。数据存储与大数据应用密切相关。

大数据给存储系统带来了 3 个方面的挑战：一是存储规模大，通常会达到 PB 级甚至 EB 级；二是存储管理复杂，需要兼顾结构化、非结构化和半结构化数据；三是对数据服务

的种类和水平要求较高。目前，出现了一批用于应对大数据存储与管理挑战的新技术，具有代表性的研究包括分布式缓存（如 CARP、mem-cached）、基于 MPP 的分布式数据库、分布式文件系统（如 GFS、HDFS），以及各种 NoSQL 分布式存储方案（如 MongoDB、CouchDB、HBase、Redis、Neo4j 等）。各大数据库厂商（如 Oracle、IBM、Greenplum 等）都已推出支持分布式索引和查询的产品。

（4）数据可视化。

数据可视化是指依据图形、图像、计算机视觉及用户界面，通过对数据的表现形式进行可视化的解释。数据可视化常用的工具有 Python 中的 Matplotlib 绘图工具库、百度 ECharts、Tableau 等，Excel 中的高级图表功能也可以很好地实现数据可视化。数据分析是大数据处理的核心，但用户往往更关注结果的展示形式。如果分析结果正确，但没有采用适当的解释方法，则所得的结果很可能让用户难以理解，在极端情况下甚至会误导用户。由于大数据分析结果具有海量且关联关系极其复杂等特点，所以采用传统的解释方法基本不可行。目前常用的方法是可视化技术和人机交互技术。

可视化技术能够迅速且有效地简化与提炼数据流，帮助用户交互筛选大量的数据，有助于用户更快更好地从复杂数据中取得新的发现。用直观的图形方式向用户展示结果，已作为最佳结果展示方式之一率先被科学与工程计算领域采用。常见的可视化技术有原位分析（InSitu Analysis）、标签云（Tag Cloud）、历史流（History Flow）、空间信息流（Spatial Information Flow）、不确定性分析等。我们可以根据具体的应用需要选择合适的可视化技术，如通过数据投影、维度降解和电视墙等方法解决大数据显示问题。

### 3.2.2 大数据处理工具和技术发展趋势

#### 1. 大数据处理工具

（1）常用的大数据处理工具。

现有的大数据处理工具大多是对开源的 Hadoop 平台进行改进并将其应用于各种场景。Hadoop 完整生态系统中的各子系统都有相应大数据处理的改进产品。常用的大数据处理工具如表 3-1 所示，这些工具中的部分已经投入商业应用，还有一部分是开源软件。在已经投入商业应用的工具中，绝大部分是在开源 Hadoop 平台的基础上进行功能扩展的，或者是提供 Hadoop 平台的数据接口。

表 3-1 常用的大数据处理工具

| 种类 | | 工具示例 |
|---|---|---|
| 平台 | Local | Hadoop、MapR、Cloudera、Hortonworks、BigInsights、HPCC |
| | Cloud | AWS、GoogleComputeEngine、Azure |
| 数据库 | SQL | MySQL（Oracle）、MariaDB、PostgreSQL、TokuDB、AsterData、Vertica |
| | NoSQL | HBase、Cassandra、MongoDB、Redis |
| | NewSQL | Spanner、Megastore、F1 |
| 数据仓库 | | Hive、HadoopDB、Hadapt |
| 数据收集 | | ScraperWiKi、Needle base、bazhuayu |
| 数据清洗 | | Data Wrangler、Google Refine、Open Refine |

续表

| 种类 | | 工具示例 |
|---|---|---|
| 数据处理 | 批处理 | MapReduce、Dyrad |
| | 流式计算 | Storm、S4、Kafka |
| | 内存计算 | Drill、Dremel、Spark |
| 查询语言 | | HiveQL、PigLatin、DryadLINQ、MRQL、SCOPE |
| 统计与机器学习 | | Mahout、Weka、R、RapidMiner |
| 数据分析 | | Jaspersoft、Pentaho、Splunk、Loggly、Talend |
| 可视化分析 | | Google Chart API、Flot、D3、Processing、Fusion Tables、Gephi、SPSS、SAS、R、Modest Maps、Open Layers |

（2）基于云的数据分析平台。

目前大部分企业分析的数据量都在 TB 级，但按照数据的发展趋势，很快就会进入 PB 时代。企业的大数据分析工具和数据库也将走向云计算。基于云的数据分析平台框架如图 3-12 所示。

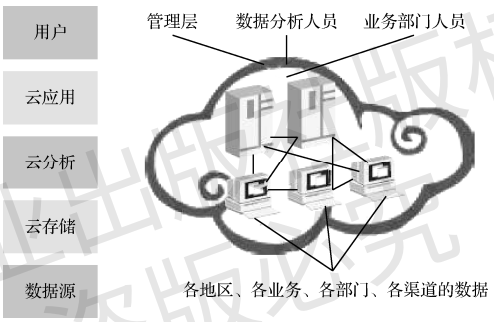


图 3-12　基于云的数据分析平台框架

云计算可以为大数据带来哪些变化呢？

云计算为大数据提供了可以弹性扩展、相对便宜的存储空间，以及计算资源，使得中小企业也可以像亚马逊一样通过云计算完成大数据分析。云计算可容纳的资源庞大且分布较为广泛，是使得异构系统较多的企业能够及时、准确地处理数据的有力方式，甚至是唯一方式，它为大数据处理方式带来了变化。

### 2. 技术发展趋势

目前，与大数据相关的技术和工具非常多，它们被用作大数据采集、存储、处理和呈现的有力武器，为企业提供了更多的选择。随着大数据的不断演进，其各个环节的技术发展呈现新的趋势，如表 3-2 所示。

表 3-2　大数据技术发展趋势

| 主要技术 | 发展趋势 |
|---|---|
| 采集与预处理 | ✧ 数据源的选择和高质量原始数据的采集方法<br>✧ 多源数据的实体识别和解析方法<br>✧ 数据清洗和自动修复方法<br>✧ 高质量数据的整合方法<br>✧ 数据演化的溯源管理 |

续表

| 主要技术 | 发展趋势 |
| --- | --- |
| 存储与管理 | ✧ 大数据索引和查询技术<br>✧ 实时/流式大数据存储与处理 |
| 计算模式与系统 | ✧ Hadoop改进后与其他计算模式和平台共存<br>✧ 混合计算模式成为大数据处理的有效手段 |
| 数据分析与挖掘 | ✧ 更加复杂<br>✧ 大规模分析与挖掘<br>✧ 大数据实时分析与挖掘<br>✧ 大数据分析与挖掘的基准测试 |
| 可视化分析 | ✧ 原位分析<br>✧ 人机交互<br>✧ 协同与众包可视分析<br>✧ 可扩展性与多级层次问题<br>✧ 不确定性分析和敏感分析<br>✧ 可视化与自动数据计算挖掘结合<br>✧ 面向领域和大众的可视化工具库 |
| 数据隐私与安全 | ✧ 进一步完善NoSQL<br>✧ APT攻击研究<br>✧ 社交网络的隐私保护<br>✧ 数字水印技术<br>✧ 风险自适应访问控制<br>✧ 数据采集、存储、分析3个过程“三权分立” |
| 其他 | ✧ 大数据高效传输架构和协议<br>✧ 大数据虚拟机集群优化研究 |



## 3.3 大数据的应用

### 1. 商品零售大数据

视频：大数据的应用

阿里巴巴公司根据淘宝网上中小企业的交易状况筛选出了财务健康和讲究诚信的企业，并对它们发放无须担保的贷款。零售企业会监控顾客在店内的走动情况及其对商品的查看情况，大数据将这些信息与交易记录相结合展开分析，从而对销售哪些商品、如何摆放货品及何时调整售价给出意见。此类方法已经帮助某零售企业减少了17%的存货，同时在保持市场份额的前提下，增加了高利润率自有品牌商品的比例。

### 2. 消费大数据

亚马逊“预测式发货”的新专利，可以通过对用户数据的分析，在用户正式下单购物前，提前发出包裹。这项技术可以缩短发货时间，从而增强消费者网上购物的意愿。从下单到收货的时间延迟，可能会降低用户的购物意愿，从而导致他们减少网上购物的频率。所以，亚马逊会根据之前的订单和其他因素，分析用户的购物习惯，从而在实际下单前便将包裹发出。该专利文件提出，虽然包裹会提前从亚马逊发出，但在用户正式下单前，这

些包裹仍会暂存在快递公司的转运中心或卡车里。为了确定要运送哪些货物，亚马逊会参考之前的订单、商品搜索记录、愿望清单、购物车，甚至用户的鼠标指针在某件商品上悬停的时间。

### 3. 证监会大数据

实际上，早在 2009 年，上海证券交易所（简称上交所）就曾有过利用大数据设置“捕鼠器”的设想。通过建立相关模型，设定一定的预警指标，在相关指标达到某个预警点时，监控系统即可自动报警。而此次在“马乐案”中亮相的深圳证券交易所（简称深交所）的大数据监测系统，更是引起了广泛关注。深交所设置了 200 多个指标用于监测估计，一旦出现股价偏离大盘走势的情况，深交所就会利用大数据查探异动背后是哪些人或机构在操控。

### 4. 金融大数据

阿里巴巴的“水文模型”会按照小微企业的类目、级别等统计商户的相关“水文数据”。例如，过往每到某个节点，某店铺的销售情况就会进入旺季，销售额会增长，其对外投放的资金额度也会上升。结合这些“水文数据”，系统可以判断出该店铺的融资需求；结合该店铺的以往资金支用数据及同类店铺的资金支用数据，可以判断出该店铺的资金需求额度。

### 5. 金融服务大数据

大连商品交易所（简称大商所）依托气象数据创新金融服务。自 2022 年，在服贸会“环境服务 • 双碳经济论坛”上，“中央气象台–大商所温度指数”在多年成功试运行的基础上正式发布后，大商所联合中央气象台及相关金融机构、产业主体，积极推进该指数在保险和场外衍生品等方面的应用，现已有多款挂钩“中央气象台–大商所温度指数”的创新保险产品陆续落地，在水产养殖、电力销售、居民生活等方面形成了多种应用场景，并通过场外衍生品构建利益相融机制，实现金融与实体企业的融合应用。这既可以帮助经营主体应对天气变化带来的负面影响，又可以为保险公司规避赔付风险开辟新路径，还可以为推动气象数据要素在更广范围内的应用奠定基础。

### 6. 制造业大数据

在摩托车生产商哈雷·戴维森公司位于宾夕法尼亚州约克市的翻新摩托车制造厂中，软件在不停地记录着各种制造数据，如喷漆室风扇的速度等。当软件“察觉”到风扇速度、温度、湿度或其他变量偏离规定数值时，就会自动调整相应的结构。哈雷·戴维森公司使用软件寻找制约公司每 86s 完成一台摩托车制造工作的原因，通过研究数据发现，产生瓶颈的原因是安装后挡泥板的时间过长。通过调整工厂配置，成功提高了安装该配件的速度。

### 7. 医疗大数据

继世界杯、高考、景点和城市预测之后，百度又推出了疾病预测产品。最新的百度灵医智惠医疗大数据解决方案（见图 3-13）已帮助多家三甲医院进行数据整理及分析，充分挖掘数据潜力。


图 3-13　百度灵医智惠医疗大数据解决方案

大数据使更多的医疗监测产品更广泛地被应用。例如，通过社交网络来收集数据的健康类 App，也许在数年后，它们收集的数据能够让医生的诊断变得更为准确；社交网络为许多慢性病患者提供了临床症状交流和诊治经验分享平台，医生借此平台可以获得部分临床效果统计数据；基于对人体基因的大数据分析，可以实现对症下药的治疗手段；公共卫生部门可以通过全国联网的患者电子病历库，快速检测传染病并进行全面的疫情监测，通过集成的疾病监测和响应程序可实现快速响应。

### 8. 交通大数据

2021 年，百度地图上线了“未来出行”功能，该功能是根据百度飞桨深度神经网络和丰富的交通大数据，通过对用户出行大数据与实时和历史交通大数据进行智能分析，使用户获得当前或未来出行的最佳规划路线。用户只需打开百度地图，在搜索目的地后，选择驾车模式规划路线，之后点击屏幕右侧的“未来出行”按钮，即可查看不同时间段内的预估通行时间。根据个人需求，用户还可以选择不同的路线，或通过预计到达时间来反推什么时间出发不用担心拥堵、迟到问题。百度基于地图应用的 LBS 预测涵盖范围更广。例如，春运期间预测人们的流动趋势，为火车线路和航线的设置提供数据支持；节假日期间预测景点的人流量，为人们进行景区选择提供建议；平时通过百度热力图来为用户呈现城市商圈、动物园等地点的人流情况，为用户的出行选择和商家的选点选址提供参照。交通运输部门可根据不同时点、不同道路的车流量预测情况，进行智能的车辆调度或应用潮汐车道，用户则可以根据预测结果选择拥堵概率更低的出行路线。

### 9. 公安大数据

大数据挖掘技术的底层技术最早是英国军情六处研发用来追踪恐怖分子的技术。使用大数据技术可以筛选犯罪团伙，如排查与锁定的犯罪嫌疑人乘坐同一班列车、住同一个酒店的人是否是其同伙。过去，刑侦人员要证明这一点，需要拼凑不同的线索来排查疑犯。

通过对大量相关数据的挖掘分析，可显示某一区域内的犯罪率及犯罪模式。大数据可以帮助警方定位到最容易受不法分子侵扰的区域，从而创建犯罪高发地区的热点图和时间表。这不但有利于警方精准分配警力、预防打击犯罪，也能帮助市民了解所在区域的犯罪

情况、提高警惕。大数据还能应用于审计反腐工作，如图 3-14 所示。


图 3-14　大数据审计反腐

### 10. 文化传媒大数据

与传统电视剧有别，《纸牌屋》是一部根据“大数据”制作的作品。制作方 Netflix 是美国最具影响力的影视网站之一，在美国本土有约 2900 万名订阅用户。Netflix 的成功之处在于其强大的推荐系统 Cinematch，该系统将用户视频点播的基础数据，如评分、播放、快进、观看时长、播放终端等存储在数据库中，之后通过数据分析，推断出用户可能喜爱的影片种类，并为其提供定制化的推荐内容。

Netflix 发布的数据显示，用户在 Netflix 上每天产生 3000 多万个行为，如暂停、回放或快进等。同时，用户每天还会给出 400 多万个评分，发出 300 多万次搜索请求。于是，Netflix 决定根据这些数据来制作一部电视剧，投资过亿美元制作《纸牌屋》。Netflix 发现，用户中有很多人仍在点播 1991 年的 BBC 经典老片《纸牌屋》，这些观众中有许多人喜欢大卫·芬奇，并且观众大多爱看奥斯卡奖得主凯文·史派西的电影。由此 Netflix 邀请大卫·芬奇作为导演，凯文·史派西作为主演，翻拍了《纸牌屋》这一政治题材剧。2013 年 2 月《纸牌屋》上线后，Netflix 的用户数量增加了 300 万，达到了 2920 万。

### 11. 航空大数据

Farecast 已经拥有惊人的、约 2000 亿条的飞行数据记录，它被用来推测当前的机票价格是否合理。作为一种商品，同一架飞机上每个座位的价格本不应该有差别。但实际上，价格却千差万别，其中缘由只有航空公司了解。Farecast 能够预测当前的机票价格在未来一段时间内的走势。这个系统需要分析所有特定航线机票的销售价格，以确定票价与提前购买天数的关系。Farecast 票价预测的准确率已高达 75%。使用 Farecast 预测工具购买机票的旅客，平均每张机票可节省 50 美元。

### 12. 人体健康大数据

慢性病发生前人体会有一些持续性异常。从理论上来说，如果大数据掌握了这样的异常情况，便可以进行慢性病预测。结合智能工具，慢性病的大数据预测已变为可能。可穿戴设备和智能健康设备可收集人体健康数据，如心率、体重、血脂、血糖、运动量、

睡眠情况等。如果这些数据足够精准且全面，并且开发出了可以形成算法的慢性病预测模式，那么在未来，设备工具或许就可以预测用户是否有罹患某种慢性病的风险。KickStarter 上的 MySpiroo 便可收集哮喘病人的吐气数据，医生可根据该数据诊断病人病情的变化趋势。

### 13. 体育赛事大数据

世界杯期间，Google、百度、微软、高盛等公司都推出了比赛结果预测平台。百度的预测结果最为亮眼，预测全程 64 场比赛的准确率为 67%，进入淘汰赛后的准确率为 94%。互联网公司取代章鱼保罗试水赛事预测意味着未来的体育赛事结果可能会被大数据预测掌控。

Google 世界杯预测是基于 OptaSports 的海量赛事数据来构建最终的预测模型的。百度则是收集过去 5 年内全世界 987 支球队（含国家队和俱乐部队）的 3.7 万场比赛数据，同时与中国彩票网站乐彩网、欧洲必发指数数据供应商 Spdex 进行数据合作，导入博彩市场的预测数据，建立一个囊括 199 972 名球员和 1.12 亿条数据的预测模型，并在此基础上进行结果预测。

### 14. 灾害大数据

气象预测是最典型的灾害预测。如果可以利用大数据预测地震、洪涝、高温、暴雨等自然灾害，便可以减灾、防灾、救灾、赈灾。过去的数据收集方式存在有死角、成本高等问题，但物联网时代可以借助传感器、摄像头和无线通信网络，进行实时的数据监控收集，再利用大数据预测分析，做出更精准的自然灾害预测。

以气象卫星数据为例，气象卫星虽然是用来获取与气象要素相关的各类信息的，但是在森林草场火灾、船舶航道浮冰分布等方面，气象卫星也能发挥跨行业的实时监测服务的价值。气象卫星、天气雷达等设备监测到的非常规遥感遥测数据中包含的信息十分丰富，有可能从中挖掘出新的应用价值，从而拓展气象行业的业务领域和服务范围。例如，可以利用气象大数据为农业生产提供服务。美国硅谷有家专门从事气象候数据分析处理的公司，它从美国气象局等数据库中获取数十年来的天气数据，之后将各地降雨量、气温和土壤状况与历年农作物产量的相关度做成精密图表，用来预测各地农场来年的产量和适宜种植的品种，同时向农户提供个性化保险服务。气象大数据应用还可以在林业、海洋、气象灾害等方面拓展新的业务领域。

### 15. 环境变迁大数据

大数据除了可以进行短时间内微观的天气、灾害预测，还可以进行长期的、宏观的环境和生态变迁预测。森林和农田面积缩小、野生动植物濒危、海岸线上升、温室效应等问题都是地球面临的“慢性问题”。人类越多地了解地球生态系统及天气形态变化的数据，就越容易模拟出未来环境的变迁，进而阻止有害的转变发生。大数据提供了预测工具，帮助人类收集、存储和挖掘更多的地球数据。

除了上面列举的 15 个领域，大数据还可以被应用于房地产预测、就业情况预测、高考分数线预测、选举结果预测、奥斯卡大奖预测、保险投保者风险评估、金融借贷者还款能力评估等方面，使人类具备可量化、有说服力、可验证的洞察未来的能力。

美国的 Viktor Mayer-Schönberger 在《大数据时代》一书中提到：“未来，数据将会像土地、石油和资本一样，成为经济运行中的根本性资源。”

总之，未来的信息世界是“三分技术、七分数据”，得“数据”者得“天下”。

## 3.4　大数据的成长及挑战

视频：大数据的成长及挑战

在大数据时代，数据存在多源异构、分布广泛、动态增长、先有数据后有模式等诸多特点。正是这些不同于传统数据的特点，使得大数据时代的数据管理面临新的挑战。目前大数据处理和分析工具较为落后，问题较为严重：在大数据背景下，传统的数据分析软件都是失效的。利用目前的主流软件工具，无法在合理的时间内撷取、管理和处理数据，并将其整理成能够帮助企业经营或为主管部门决策提供支持的数据。

### 3.4.1　大数据的成长

IT（Information Technology）时代（信息时代）与 DT（Data Technology）时代（数据时代）是承前启后的两个时代。信息时代是数据时代的基石与前奏，数据时代是信息时代的传承与发展。数据时代在以一种全新的方式颠覆人们工作、生活和娱乐的模式。

（1）互联网技术推动了大数据的泛在化。

通常来讲，互联网发展经历了研究网络、运营网络和商业运营网络 3 个阶段。互联网的重要性不仅体现在其规模庞大上，而且体现在其能够提供全新的全球信息服务基础设施上。此外，互联网彻底改变了人类的思维模式、工作和生活方式，促进了社会各行业的发展，成为了时代的重要标志之一。互联网产生的数据量不断增加，尤其是电子政务、社交媒体、网上购物等应用是实时提供数据的，需要处理的网络数据越来越多，在数据处理、传输与应用方面就出现了新的问题。这种发展趋势加上其他网络数据源的普及，使大数据的泛在化成为必然结果。

（2）存储技术支撑了大数据的大容量化。

从世界上的第一台计算机出现以来，计算机的存储设备就在不断更新，从水银延迟线、磁带、磁鼓、磁芯到现在的半导体存储器、磁盘、光盘和纳米存储器，存储容量不断扩大，而存储器的价格却在不断下降。自 2005 年亚马逊公司推出云服务平台后，一种新型的网络存储方式——云存储，便逐渐应用推广，用户可以通过其获取更大的存储容量。云存储允许用户访问云中的存储资源以扩大用户的存储容量，用户可以随时随地借助任何连接到网络的设备轻松连接云端并读取数据。

（3）计算能力加速了大数据的实时化。

信息产业的发展正如摩尔所预言的那样，定期推出了具有不断优化能力的操作系统和性能更加强大的计算机。硬件厂商每开发一款运算能力更强的芯片，软件服务商就会开发一款更便捷的操作系统，这极大地提高了信息处理的速度。尤其是超级计算机和云计算的产生，使得对数据的计算能力极大地增强，为大数据的实时化处理提供了可能。

### 3.4.2 挑战与机遇

尽管大数据给人类的生产生活带来了翻天覆地的变化，但是受数据质量、分析技术和接受程度的制约，大数据在新时代需要面临许多的挑战，同时也有许多的机遇。

（1）数据的挑战与机遇。

在实际应用中，大数据的获取较为困难，同时数据质量也难以保证。通常仅针对某几个具体指标进行数据收集，如果长期依赖于部分维度的数据进行分析，那么预测结果会因为数据的不全面而产生偏差。在庞大的物联网中，设备具有一定的损坏率，从而导致收集到的数据有一些错误或偏差较大，同时采集数据的终端传感器如果存在误差，那么也会导致数据的准确性降低。此外，数据在网络中的传输具有一定的误码率，尽管误码率非常低，但如果长期不进行数据校验，或少部分关键性信息发生错误，就会对数据分析结果产生较大影响。

但也要看到针对某些特定领域的总体决策方案，大数据使得“全样本”数据的获取成为可能，而传统的“小数据”分析所需要的数据假设前提将不复存在。同时，呈指数级增长的非结构化数据和实时流数据，使得大数据的数据处理对象发生了极大的变化。通过速度极快的数据采集、挖掘与分析，可以从异构、多源的大数据中获取高价值信息，从而提供实时精准的预警预测，形成辨别决策的“洞察力”，这将是大数据给予的最好机遇，也将是大数据系统的发展方向。

（2）技术的挑战与机遇。

目前，数据挖掘与分析的算法可采用机器学习的方式。机器学习依赖于大数据不断地迭代学习，并不断地修正训练模型的参数，其局限性是难以创造新的知识，只能挖掘数据固有的规律和联系。学习效果的好坏取决于学习模型的选择，良好的学习模型能收获较好的学习效果；若学习模型选择不当，则即使计算迭代的次数再多，也难以得到理想的结果。同时，在利用大数据驱动决策时，需要将决策问题模型化，做出一些合理性假设，忽略影响较小的因素，抓住关键问题和主要矛盾。在这个过程中，某些合理性假设未必合理，这将导致决策结果出现偏差。

大数据的出现使得传统数据的存储管理和挖掘分析技术难以适应时代发展的要求，这需要大数据研究者和使用者应用新的管理分析模式，从非结构化数据和流数据中挖掘价值、探求知识。大数据对存储的需求，加速了 HDFS、BigTable 等技术的发展；大量的并发数据事务处理，催生了 NoSQL 数据库；众多的数据需求分析处理，发展了 MapReduce、Hadoop 等大数据处理技术。此外，大数据与人工智能、地理信息、图像处理等多个研究领域交叉融合，彰显了基于数据驱动的大数据技术的美好前景。

（3）用户的挑战与机遇。

大数据驱动模式不同于以往依赖于相关领域专家和领导者的经验驱动模式，其分析与决策功能可辅助专家和领导者做出决定。但是，大数据应用需要建立大数据仓库和大数据系统，前期需要投入较高的经济成本，运营程度的好坏也会影响其在分析决策过程中的效果。

大数据产生的效益与机遇是不可小觑的。目前，各行业企业只是刚刚进入大数据的应用阶段，使用大数据辅助决策对绝大部分行业来说都是新时期竞争优势的创造源泉。调查显示，数据驱动型企业在生产率和盈利水平等方面普遍优于同行业竞争者。数据驱动的系统在处理特定问题时，可以做出更优的决策，如金融领域的某些系统基于大数据可以做出

占比较高的投资决策。从现在至可预见的将来，能更好地运用大数据的组织和企业将迸发出更多的创新性，可以更好地维持决策的灵活性。整个社会对数据驱动应用和决策的依赖性会越来越高。

（4）大数据隐的私与安全。

近年来，手机应用、智能摄像头、Wi-Fi 等工具泄露用户隐私的现象时有发生。如今，支撑智能时代的大数据、云计算、人工智能等技术，既是创新发展的助推器，又是滋生网络安全问题的催化剂。在智能时代，新技术既是帮凶，又是克星。信息安全的攻防战永无止境。

在密码技术层面，应将密码技术与数据标识相结合，通过信任管理、访问控制、数据加密、可信计算、密文检索等措施，构建集传输、分析、应用于一体的数据安全体系，解决隐私保护性差、身份假冒等问题。

英国励讯集团全球副总裁 FlavioVillanustre 认为，在数据流通方面，建议通过匿名方式使脱敏数据去掉标签；也可以通过“差别隐私”机制，在数据中加入一些“噪声”，以保护数据不被外部识别。

在用户数据保护方面，企业作为数据的收集者、控制者，既要做“运动员”又要做“裁判员”，显然难以解决问题。因此数据保护不能仅靠企业自律，还要让法律推动内生机制的生成。

## 思政园地

### 素养目标

- ✧ 对大数据技术的学习，能够培养学生的数据感，使其从数据的视角观察和认识世界。
- ✧ 使学生能够从现实系统中提取数据或使用信息技术采集所需的数据。
- ✧ 培养学生理解数据分布的耦合关系，即共现关系、近邻关系、依赖关系、链接关系、相关关系和因果关系等。
- ✧ 培养学生的数据权利意识，使其充分认识数据对于自身隐私和生活的重要性，积极保护自身权益。
- ✧ 引导学生正确使用大数据平台，遵循互联网准则和法律规范。

### 思政案例

大数据是如何精准识别大山里的贫困户的

大数据是如何精准识别大山里的贫困户的，请扫描右侧二维码观看视频。

数据“孤岛”、数据壁垒是大数据发展的“痛点”，也是扶贫工作的难点。让记者带大家走进贵州精准扶贫大数据支撑平台，看它如何对贫困户进行精准画像、精准识别。（视频来源：新华视频）

### 自我检测

#### 一、单选题

1．从大量数据中提取知识的过程通常称为_____。

A．数据挖掘　　B．人工智能

C．数据清洗　　D．数据仓库

2．下列论据中，能够支撑“大数据无所不能”的观点的选项是_____。

A．互联网金融打破了传统的观念和行为

B．大数据存在泡沫

C．大数据具有非常高的成本

D．个人隐私泄露与信息安全担忧

3．大数据的起源是_____。

A．金融　　B．电信

C．互联网　　D．公共管理

4．大数据正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，并从中发现新知识、创造新价值、提升新能力的_____。

A．新一代信息技术

B．新一代服务业态

C．新一代技术平台

D．新一代信息技术和服务业态

5．当前社会中，最为突出的大数据环境是_____。

A．互联网　　B．物联网

C．综合国力　　D．自然资源

6．大数据的 5V 特征中的 Volume 是指_____。

A．价值密度低　　B．速度快

C．多样性　　D．容量大

7．第一个提出大数据概念的公司是_____。

A．微软　　B．Google

C．Facebook　　D．麦肯锡

8．大数据最显著的特征是_____。

A．容量大　　B．多样性

C．速度快　　D．价值密度低

9．对大数据在社会综合治理中的作用，以下理解不正确的是_____。

A．大数据的运用能够维护社会治安

B．大数据的运用能够加强交通管理

C．大数据的运用能够杜绝抗生素的滥用

D．大数据的运用有利于走群众路线

10．在大数据时代，数据使用的关键是_____。

A．数据收集　　B．数据存储

C．数据分析　　D．数据再利用

## 二、多选题

1．在医疗领域中是如何利用大数据的？_____

A．临床决策支持　　B．个性化医疗

C．社保资金安全　　D．用户行为分析

2．下列关于大数据的说法中，错误的是_____。

A．大数据具有体量大、结构单一、时效性强的特征

B．处理大数据需采用新型计算架构和智能算法等新技术

C．大数据的发展离不开云计算技术的支持

D．大数据的发展相应地带来了数据安全问题

3．属于大数据的 5V 特征的有_____。

A．容量大　　B．商业价值高

C．速度快　　D．多样性

## 三、判断题

1．Google 流感趋势充分体现了数据重组和扩展对数据价值的重要意义。（　　）

2．对于大数据而言，最基本、最重要的要求就是减少错误、保证质量。因此，大数据收集的信息是精确的。（　　）

3．在大数据的范畴内，应该把用户视为互联网中的数据分子，独立、细致地对其行为进行特征分析，充分挖掘大数据的价值，变数据为“资产”。（　　）

## 四、讨论题

通过学习本章内容，结合实际情况总结一下大数据对我们的生活有哪些影响？