



计算机类专业“互联网+”创新型精品教材

大数据平台部署与运维

大数据平台部署 与运维

主 编 宿宏毅 阿 伦 刘海龙

主
编
宿宏毅
阿
伦
刘海龙

北京出版集团
北京出版社



北京出版集团
北京出版社

图书在版编目 (CIP) 数据

大数据平台部署与运维 / 宿宏毅, 阿伦, 刘海龙主
编. -- 北京: 北京出版社, 2025.1. -- ISBN 978-7
-200-19218-6

I . TP274

中国国家版本馆 CIP 数据核字第 2025VX0775 号

大数据平台部署与运维

DASHUJU PINGTAI BUSHU YU YUNWEI

主 编: 宿宏毅 阿 伦 刘海龙

出 版: 北京出版集团

北 京 出 版 社

地 址: 北京北三环中路 6 号

邮 编: 100120

网 址: www.bph.com.cn

总 发 行: 北京出版集团

经 销: 新华书店

印 刷: 定州启航印刷有限公司

版 印 次: 2025 年 1 月第 1 版 2025 年 1 月第 1 次印刷

成品尺寸: 185 毫米 × 260 毫米

印 张: 13.5

字 数: 304 千字

书 号: ISBN 978-7-200-19218-6

定 价: 43.00 元

教材意见建议接收方式: 010-58572341 邮箱: jiaocai@bphg.com.cn

如有印装质量问题, 由本社负责调换

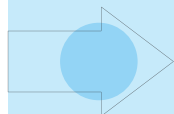
质量监督电话: 010-82685218 010-58572341 010-58572393



项目一 初识数据分析	1
任务一 初识大数据	2
任务二 大数据的发展认知	6
项目二 平台基础环境配置	15
任务一 安装操作系统	16
任务二 配置静态 IP 及远程登录	23
项目三 Hadoop 框架的部署	32
任务一 Hadoop 的认知	33
任务二 Hadoop 单机模式和伪分布模式的部署	36
任务三 Hadoop 全分布模式的部署	49
项目四 HDFS 的使用	66
任务一 HDFS 的认知	67
任务二 使用 HDFS 的 Web 界面	73
任务三 使用 Shell 管理 HDFS 文件与目录	77
项目五 初识 MapReduce	86
任务一 MapReduce 的认知	87
任务二 使用 MapReduce 实现词频统计	96
任务三 使用 MapReduce 完成电商销售数据的统计	106
项目六 HBase 的部署与使用	116
任务一 HBase 的认知	117
任务二 HBase 单机模式和伪分布模式的部署	123



任务三	HBase 完全分布模式的部署	129
任务四	利用 Shell 操作 HBase	133
项目七	Hive 的部署与使用	142
任务一	Hive 的认知	143
任务二	Hive 本地模式的部署	147
任务三	Hive 远程模式的部署	154
任务四	利用 Hive 实现数据导入	161
任务五	利用 Hive 实现词频统计	168
项目八	Spark 的部署与使用	175
任务一	Spark 的认知	176
任务二	Spark Local 的部署与操作	179
任务三	Spark Standalone 的部署与操作	183
任务四	Spark on YARN 的部署与操作	196
参考文献		207



项目一

初识数据分析

大数据已经成为现代社会的重要组成部分。它不仅改变了我们的生活方式和工作方式，也为企业和政府机构提供了更宽广的视野和更深邃的洞察力。未来，随着技术不断进步和应用领域不断拓展，大数据将会发挥更大的作用。

知识目标

- » 了解大数据的概念及内涵。
- » 了解大数据的 4V 特征。
- » 了解大数据的发展背景及未来发展方向。

能力目标

- » 能梳理出大数据技术体系的基础框架。
- » 具备数据理解与洞察能力，能洞察数据背后的业务价值。

素养目标

- » 具备批判性思维和问题解决能力，能从大量数据中提取有价值的信息，并利用这些信息做出决策或解决问题。
- » 不断学习和更新自己的知识和技能，以适应不断推陈出新的大数据技术和应用场景。

任务一

初识大数据

任务导入

当前，大数据在各个行业都已经有了广泛应用，例如在零售业，可根据历年销售、社交媒体等数据预测需求、调整库存，可根据顾客购物历史、偏好实施个性化营销，激活需求、提升客户忠诚度；在金融服务行业，可通过分析大量交易和用户行为数据，检测欺诈、保障资金安全、评估风险、优化投资组合、提升资产管理效率；在社交媒体领域，可借其分析用户行为，提供个性化内容与广告，可监测舆情，了解公众意见与情绪。随着技术发展，大数据将在各领域继续拓展应用边界。那么，什么是大数据？大数据和传统数据库有什么区别呢？

任务目标

1. 能理解大数据的四大特点。
2. 会鉴别大数据在生活中的应用。

任务实施

一、大数据的概念

近年来，大数据已成为一个高频词。无论是在专业领域，还是在日常娱乐消费领域，大数据都早已进入人们的生活，跟大众休戚相关。那么，什么是大数据？大数据并不是指数据量大于某个阈值，而是指超出了传统数据库处理能力的数据集。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

也可以说，大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要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。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换句话说，如果把大数据比作一种产业，那么这种产业实现盈利的关键在于提高数据“加工能力”，通过“加工”实现数据的“增值”。

大数据必然无法用单台计算机处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。因此，大数据必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

二、大数据的特点

听起来复杂高深，但其实很简单，大数据就是海量资料、巨量资料，这些巨量资料来源于世界各地随时产生的数据，在大数据时代，任何微小的数据都可能产生不可思议的价值。在维克托·迈尔－舍恩伯格及肯尼思·库克耶编写的《大数据时代》中提到了大数据的3个特征：Volume 海量、Velocity 快速、Variety 多样。另外，随着大数据技术的成熟和应用场景的不断增加，另一个特征也越来越突出，即 Value 价值。这四项就是常说的大数据的4V特征，如图1-1所示。

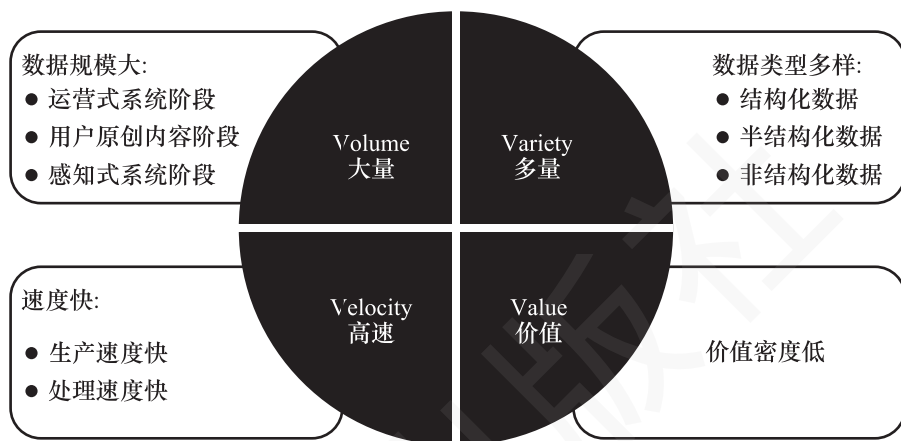


图 1-1 大数据的 4V 特征

（一）大量（Volume）

从数字化技术诞生以来，人类社会的数据生产方式经历了三个阶段：

- （1）运营式系统阶段，数据仅由少数运营人员生产并发布。
- （2）用户原创内容阶段，即自媒体时代，全民都是数据的生产者与发布者，数据生产速度大幅提升。
- （3）感知式系统阶段，即物联网时代，各类摄像头、传感器实时生产数据、记录数据，其数据生产速度较人工数据生产再一次发生质的跨越。当前，我们正处于感知式系统阶段，数据正呈现爆发性增长。

大数据的特征首先就体现为数据规模大。Map3 时代，一个小小的 MB 级别的 Map3 就可以满足很多人的需求，然而随着时间的推移，存储单位从过去的 GB 到 TB，乃至现在的 PB、EB 级别。只有体量达到 PB 级别的数据集，才能被称为大数据。1PB 等于 1024TB，1TB 等于 1024GB，那么 1PB 等于 $1024 \times 1024GB$ 的数据（存储单位换算见表 1-1）。随着信息技术的高速发展，数据开始爆发性增长。社交网络、移动网络、各种智能工具、服务工具等，都成为数据的来源。淘宝网近 4 亿的会员每天产生的商品交易数据约为 20TB；Facebook 约 10 亿的用户每天产生的日志数据超过 300TB。社会迫切需要智能的算法、强大的数据处理平台和新的数据处理技术，来统计、分析、预测和实时处理如此大规模的数据。

表 1-1 存储单位换算

单位	换算公式
Byte (字节)	1Byte = 8bit
KB (KiloByte, 千字节)	1KB = 1024Byte
MB (MegaByte, 兆字节)	1MB = 1024KB
GB (GigaByte, 吉字节)	1GB = 1024MB
TB (TeraByte, 太字节)	1TB = 1024GB
PB (PetaByte, 拍字节)	1PB = 1024TB
EB (ExaByte, 艾字节)	1EB = 1024PB
ZB (ZettaByte, 泽字节)	1ZB = 1024EB

(二) 多样 (Variety)

如果只有单一的数据，那么这些数据就没有了价值，如只有单一的个人数据，或者单一的用户提交数据，这些数据则不能被称为大数据。广泛的数据来源，决定了大数据形式的多样性。例如，当前的上网用户中，年龄、学历、爱好、性格等每个人的特征都不一样，这些也就是大数据多样性的缘由，如果扩展到全国，那么数据的多样性会更强，每个地区，每个时间段，都会存在各种各样的数据。

任何形式的数据都可以产生作用，目前应用最广泛的就是推荐系统，如淘宝、网易云音乐、今日头条等，这些平台都会通过对用户的日志数据进行分析，从而进一步推荐用户喜欢的东西。

目前按照数据类型可将大数据分为三类：结构化数据、半结构化数据、非结构化数据。

(1) 结构化数据：是由二维表结构来逻辑表达和实现的数据，严格遵循数据格式与长度规范，可以存储在传统关系型数据库中，如财务系统、医疗系统、信息管理系统等，数据间具有因果关系。

(2) 半结构化数据：相对于结构化数据，其数据结构变化很大，因果关系弱。通常采用 HTML 文档、邮件、网页等形式表达。如员工简历信息，不是所有的员工简历都具有相同的字段，具有一定的灵活性。

(3) 非结构化数据：指图像、视频、音频文件，数据间没有因果关系。

(三) 高速 (Velocity)

可以从两个角度理解大数据的高速：

(1) 数据生产速度快：在物联网时代，基于传感器等数据生产方式的多样性与高速网络带宽的共同作用，高效的数据生产已成为现实。

(2) 数据处理速度快：大数据时代的应用，要求其自身对大数据具有实时分析结果的能力，以实时指导生活和生产实践。例如，当前各大 App 基于用户行为实现的用户内容精准推送能力，极大程度上决定了其在同类市场中的用户占有率。

生活中每个人都离不开互联网，也就是说，每个人每天都在向大数据提供大量的资料。并且这些数据是需要及时处理的，因为花费大量资本去存储作用较小的历史数据

是非常不划算的, 对于一个平台而言, 也许保存的数据只有过去几天或一个月之内的数据, 再远的数据就要及时清理, 不然代价太大。基于这种情况, 大数据对处理速度有着非常严格的要求, 服务器中大量的资源都用于处理和计算数据, 很多平台都需要做到实时分析。数据无时无刻不在产生, 谁的速度更快, 谁就有优势。

(四) 价值 (Value)

价值也是大数据的核心特征。在现实世界所产生的数据中, 有价值的数据所占比例很小, 价值密度较低。例如, 小区监控视频 24 小时记录小区监控画面, 需要消耗大量存储空间, 却仅在特殊情况下 (如盗窃案) 的一小段视频才有意义。

但是, 相比于传统的小数据, 大数据最大的价值在于, 通过从大量不相关的各种类型的数据中, 挖掘出对未来趋势与模式预测分析有价值的数据, 并通过机器学习、人工智能或数据挖掘深度分析, 发现新规律和新知识。当有 1PB 以上的全国所有年龄为 20~35 岁的年轻人的上网数据时, 那么它自然就有了商业价值, 如通过分析这些数据, 我们就能知道这些人的爱好, 进而指导产品的发展方向等; 如果有了全国几百万病人的数据, 对这些数据进行分析就能预测疾病的发生, 这些都是大数据的价值。

拓展阅读

大数据项目的实施流程

大数据项目实施有其独有的流程, 可以大致分为六个阶段: 项目规划、数据治理、项目设计、项目应用、迭代与实施、应用推广。大数据项目实施的流程如图 1-2 所示。

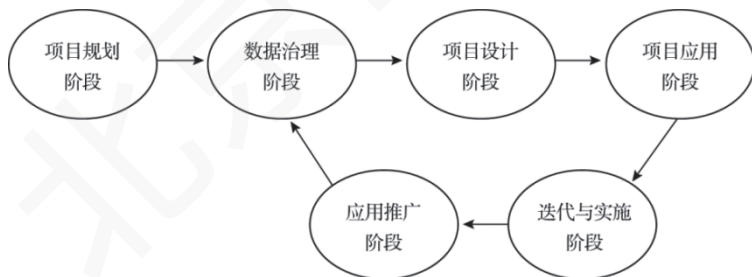


图 1-2 大数据项目实施流程

1. 项目规划阶段

该阶段为前期准备工作, 了解客户的实际情况、需求、业务模式和现有数据情况等, 然后完成数据架构规划。

2. 数据治理阶段

该阶段应进行数据采集和数据预处理, 以获得高质量的数据, 数据到位才能保障项目有效推进和执行。

3. 项目设计阶段

该阶段根据具体数据需求、数据使用模式及数据特点来设计逻辑数据模型。

4. 项目应用阶段

该阶段根据逻辑数据模型进行设计，从物理存储、文件系统、物理数据库到大数据平台进行技术选型，形成具体的功能规划。

5. 迭代与实施阶段

该阶段对大数据项目进行不断验证、修正、实施，然后根据客户试用结果评估扩展大数据的应用范围。

6. 应用推广阶段

该阶段进行有效的实施推广，数据“活”起来，才能源源不断地产生价值。

任务二

大数据的发展认知

任务导入

在数据特征的不断演变过程中，数据的资源价值逐渐得到广泛认同，这也推动了数据流通和共享需求持续增长。当前，大数据已成为社会热议的焦点话题。那么，大数据的起源何在？其历史演变又是怎样的脉络？展望未来，大数据又将呈现怎样的发展趋势？大数据的演变历程、当前状况以及未来的发展方向如何呢？

任务目标

1. 会梳理大数据前世今生的发展逻辑。
2. 能根据大数据的现状分析其未来发展趋势。

任务实施

一、大数据的前世今生

（一）大数据的发展背景

1997年，美国宇航局研究员迈克尔·考克斯和大卫·埃尔斯沃斯首次使用“大数据”这一术语来描述20世纪90年代的挑战：超级计算机生成大量的信息——在迈克尔·考克斯和大卫·埃尔斯沃斯案例中，模拟飞机周围的气流——是不能被处理和可视化的。数据集通常之大，超出了主存储器、本地磁盘，甚至远程磁盘的承载能力，这个问题在当时被称为“大数据问题”。

2008年9月4日，《自然》(Nature)刊登了一个名为“Big Data”的专辑。2011年5月，美国著名咨询公司麦肯锡(McKinsey)发布《大数据：创新、竞争和生产力的下一个前沿》的报告，首次提出了“大数据”概念，认为数据已经成为经济社会发展的重

要推动力。大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。

在中国,2012年3月,中华人民共和国科学技术部发布的“‘十二五’国家科技计划信息技术领域2013年度备选项目征集指南”把大数据研究列在首位。中国分别举办了第一届(2011年)和第二届(2012年)“大数据世界论坛”。IT时代周刊等举办了“大数据2012论坛”,中国计算机学会举办了“CNCC2012大数据论坛”。中华人民共和国科学技术部,863计划信息技术领域2015年备选项目包括超级计算机、大数据、云计算、信息安全、第五代移动通信系统(5G)等。

2014年3月,大数据首次写入中共中央政府工作报告,预示着我国大数据时代的到来。

2015年8月31日,中华人民共和国国务院正式印发《促进大数据发展行动纲要》。

为贯彻落实《中华人民共和国国民经济和社会发展第十三个五年规划纲要》和《促进大数据发展行动纲要》,加快实施国家大数据战略,推动大数据产业健康快速发展,2017年中华人民共和国工业和信息化部编制了《大数据产业发展规划(2016-2020年)》;2017年党的十九大报告中提出,要推动大数据与实体经济深度融合。

2020年,在《关于构建更加完善的要素市场化配置体制机制的意见》中,大数据被正式列为新型生产要素。

2021年,在《“十四五”大数据产业发展规划》中,提出完善大数据标准体系建设。

2021年6月,中国正式发布《中华人民共和国数据安全法》,于2021年9月1日起正式施行。

2022年10月,党的二十大报告指出,强化国家安全工作协调机制,完善国家安全防护体系、战略体系、政策体系、风险监测预警体系、国家应急管理体系,完善重点领域安全保障体系和重要专项协调指挥体系,强化经济、重大基础设施、金融、网络、数据、生物、资源、核、太空、海洋等安全保障体系建设。

(二) 大数据技术的发展历程

1. 大数据基础技术

大数据技术的体系庞大且复杂,其基础的技术包含数据采集、数据预处理、分布式存储、No SQL 数据库、数据仓库、机器学习、并行计算、可视化等多个技术门类 and 不同的技术层面。这里列举其中一些。

- (1) 文件存储: Hadoop HDFS、Tachyon、KFS。
- (2) 离线计算: Hadoop Map Reduce、Spark。
- (3) 流式、实时计算: Storm、Spark Streaming、S4、HeronK-V、NoSQL。
- (4) 数据库: HBase、Redis、MongoDB。
- (5) 资源管理: YARN、Mesos。
- (6) 日志收集: Flume、Scribe、Logstash、Kibana。
- (7) 消息系统: Kafka、StormMQ、ZeroMQ、RabbitMQ。

- (8) 分布式协调服务: ZooKeeper。
- (9) 集群管理与监控: Ambari、Ganglia、Nagios、Cloudera Manager。
- (10) 数据挖掘、机器学习: Mahout、Spark MLlib。
- (11) 数据同步: Sqoop。
- (12) 任务调度: Oozie。
- (13) 查询分析: Hive、Impala、Pig、Presto、Phoenix、Spark SQL、Drill、Flink、Druid、Kylin。

2. 大数据主流技术的发展历程

大数据主流技术的发展历程如表 1-2 所示。

表 1-2 大数据主流技术的发展历程

时间	事件
2002 年 10 月	Doug Cutting 和 Mike Cafarella 创建了开源网页爬虫项目 Nutch
2003 年 10 月	Google 发表 Google File System 论文
2004 年 10 月	Google 发表 MapReduce 论文
2006 年 2 月	Hadoop 项目正式启动以支持 MapReduce 和 HDFS 的独立发展
2006 年 11 月	Google 发表 Bigtable 论文, 这最终激发了 Hbase 的创建
2007 年 10 月	第一个 Hadoop 用户组会议召开, 社区贡献开始急剧上升
2008 年 1 月	Hadoop 成为 Apache 顶级项目
2008 年 6 月	Hadoop 的第一个 SQL 框架——Hive 成为 Hadoop 的子项目
2008 年 11 月	Apache Pig 的最初版本发布
2009 年 10 月	首届 Hadoop World 大会在纽约召开
2010 年 5 月	HBase 脱离 Hadoop, 成为 Apache 顶级项目
2010 年 9 月	Hive (Facebook) 脱离 Hadoop, 成为 Apache 顶级项目
2010 年 9 月	Pig 脱离 Hadoop, 成为 Apache 顶级项目
2010 ~ 2011 年	扩大的 Hadoop 社区忙于建立大量的新组件 (Crunch、Sqoop、Flume、Oozie 等) 来扩展 Hadoop 的使用场景和可用性
2011 年 1 月	Zoo Keeper 脱离 Hadoop, 成为 Apache 顶级项目
2012 年 3 月	重要功能 HDFS Name Node HA 被加入 Hadoop 主版本
2012 年 8 月	另外一个重要的企业适用功能 YARN 成为 Hadoop 子项目
2012 年 10 月	第一个 Hadoop 原生 MPP 查询引擎 Impala 加入 Hadoop 生态
2014 年 2 月	Spark 逐渐代替 Map Reduce 成为 Hadoop 的默认执行引擎, 并成为 Apache 基金会顶级项目
2017 年 12 月	继 Hadoop 3.0.0 的 4 个 Alpha 版本和一个 Beta 版本后, 第一个可用的 Hadoop 3.0.0 版本发布
2018 年以后	越来越多的组织开始将传统上独立的数据湖和数据仓库融合起来, 促进了跨多个数据源大数据的分析和查询。

（三）大数据的应用现状

大数据的主要作用是辅助决策。通过对大数据进行分析，能够总结经验、发现规律、预测趋势，这些都可以辅助决策。有人把数据比喻为蕴藏信息能量的矿产。像传统矿产一样，“矿产”本身不产生价值，必须和其他具体的领域、行业相结合，给使用需求提供帮助之后，才具有价值。例如，煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，每种煤炭资源都可以应用到不同的行业中；而露天煤矿、深山煤矿的挖掘成本又不一样。价值含量、挖掘成本比数量更为重要。对于很多行业而言，利用好这些大规模数据是赢得竞争的关键。我们掌握的数据信息越多，决策才能更科学、精确、合理。

随着经济的发展和科技的进步，大数据受到了越来越多的关注，因为应用了大数据技术的行业和领域不仅发展速度大幅提升，并且还大大拓展了自身的发展空间。大数据的应用除给人们的生产生活带来便利外，还能有效促进经济发展，提升综合国力。

1. 金融行业

大数据最早是从金融领域开始应用的，金融领域因为其自身的发展特性，存在着海量的数据，随着经济水平的不断提升，金融领域面临的数据分析难度越来越高，传统的分析方式已经难以满足金融行业的分析需求，大数据时代的到来为金融领域带来了福音，金融领域通过大数据技术对其海量的用户数据加以分析，从用户的信用、历史记录等多个维度分析用户的多维情况，进而对用户定向的服务，这样的定制服务往往能带动企业自身获得更加有益的发展。

2. 电商行业

在电商领域，因为大数据的应用，使其运营模式发生了一定的变化，电商平台多利用数据对消费者的需求进行分析，从而预测消费者未来的消费情况，进而预测未来的销售情况，使企业在经营决策上有数据可依并可获得更多的经济效益，在一定程度上提高了企业的竞争力。在国际方面，Amazon 平台采用的推荐算法被认为是非常成功的。在国内，比较大型的电子商务平台网站有淘宝网（包括天猫商城）、京东商城、当当网、苏宁易购等。

3. 在线社交

数据对于在线社交也有着非常重要的影响。通过来源于网络的大数据，可以让用户在即时通信上得到最大的便捷，如微博、共享空间等。在线社交的大数据对用户的活动能够进行直接反应，形成非常有效的用户信息交互。通过大数据的使用，在社交平台上能够根据用户自身的情况将用户分组，向用户推荐自己关心的信息或与自己相关的群组，能够从社交结构上对用户信息进行综合分析并提供对用户非常有价值的相关信息。对用户群体之间相互关系的有效分析，能够最大限度地帮助用户找到自己感兴趣的话题或事物。社交网络中大数据的应用，最直接的体现就是对用户数据的收集和处理，从而进行相关信息的推荐，以及在线教育等。Facebook 保存着两类最宝贵的数据：一类是用户之间的社交网络关系，另一类是用户的偏好信息。Facebook 推出了一个称为 Instant Personalization 的推荐 API (Application Programming Interface, 应用程序接口)，它能

根据用户好友喜欢的信息，给用户推荐他们的好友最喜欢的物品。很多网站都使用了 Facebook 的推荐 API 来实现网站的个性化。

4. 交通行业

大数据在交通方面的应用也较为广泛，随着互联网技术的不断成熟，智慧城市这一概念在很多城市得到了积极的倡导和推广。智慧城市就是利用大数据技术来协助管理城市，在城市管理方面，首先利用摄像头、传感器等各类传导设备实时对城市情况进行监控，采集的定时数据进而被上传到智慧城市平台中，平台准确地将城市整体的信息分析出来，使城市的管理工作更加智能、科学、规范。

5. 医疗行业

大数据在医疗领域的应用主要体现在医药研发和商业模式方面。传统的药物使用的副作用分析往往基于临床的跟踪和分析，而大数据时代，药物副作用方面的样本采集可以获得很大的扩展，如避开传统方法的弊端，使药物副作用分析在操作上更加简单。同时，在药品研究上，通过大数据可以分析人们对于医疗用品的需求趋势，相关的企业可以从市场的需求出发，制订科学、合理的生产计划，使企业的发展更加顺畅。

6. 物联网

物联网不仅能够为大数据提供非常有效的数据支持，同时也能够为大数据的应用提供足够的市场空间。物联网中的所有货品，都能够产生数据并通过数据交换进行市场消费。但是在物联网中的货品无论是数量还是种类都非常多，对于物联网的具体要求及应用物联网的方式也有极大的差异。例如，在物流企业正常的营运过程中，需要对物联网提供的大数据进行有效的应用，从而对货品配送车辆进行跟踪与调配。在物流企业中，通常会在配送车辆上安装传感器或 GPS（Global Positioning System，全球定位系统）等进行车辆定位，这样不仅能够全程监控配送过程，同时也能够在配送出现意外时进行及时处理。并且这种跟踪定位，也能够第一时间为车辆的行进选择最快捷的路线，让配送工作变得更加高效、快速。

7. 政府部门

大数据也可以应用于政府部门，政府部门自身就具有海量的数据，依靠传统的数据分析方式已经难以满足互联网时代下剧增的数据量的分析需求。政府部门通过使用大数据，可以提高自身的管理能力、转变管理的方式，从而打造服务型的政府。

大数据技术在社会的发展中具有重要的作用，现阶段，大数据技术在发展中依旧还有一些挑战，这些挑战随着大数据、物联网、区块链等方面的技术发展在不断得到完善。在社会的不断发展中，大数据技术将进入更多的行业和领域，科技的不断发展，也使大数据技术越来越完善，为经济发展贡献更大的力量。

（四）大数据的应用场景

大数据技术在各行各业中的应用见表 1-3。

表 1-3 大数据的应用场景

行业	应用
金融行业	麦肯锡研究指出，金融业在大数据价值潜力指数中排第一。例如： (1) 基于用户行为大数据，构建用户画像，实现个性化智慧营销、创新服务等 (2) 基于内部多元异构数据和外部征信数据，优化风控管理 (3) 通过交叉领域的数据共享分析，优化产品创新
医疗行业	(1) 基于海量的病例、报告、治愈方案、药物方案等，优化医疗方案，提供最佳治疗方法 (2) 监控居民人体数据，早治疗、早康复，有效预防、预测疾病
零售行业	(1) 通过对区域人口、消费、习惯、产品认知等进行大数据分析，得到更精准的市场定位 (2) 分析消费者的消费行为、价值取向等，挖掘零售市场新需求 (3) 根据区域用户购物行为，优化物流仓储网络，提高物流效率，降低物流成本
娱乐行业	根据用户行为，推送符合用户喜好的影视、文本、广告等内容
城市管理	(1) 利用道路监控系统及车载 GPS 大数据，实现实时道路路况监控等指挥交通业务 (2) 基于各类传感器，实现景区、城市环保监测及决策
物流行业	大数据分析可用于路线规划、货物追踪、运输效率优化等，提高物流效率和降低成本。
能源行业	大数据分析可用于能源消耗分析、能源供应优化、智能电网管理等，提高能源利用效率和可持续发展。
教育行业	大数据分析可用于学生评估、个性化教育、教育资源优化等，提高教育质量和学习效果。
社交媒体	大数据分析可用于用户画像、社交推荐、舆情监测等，提供个性化服务和决策支持。

二、大数据的未来发展

(一) 大数据未来前景

大数据技术目前正处在落地应用的初期，从大数据自身发展和行业发展的趋势来看，大数据未来的前景还是不错的，具体如下。

1. 大数据自身能够创造出更多的价值

大数据相关技术紧紧围绕数据价值化展开，数据价值化将开辟出广大的市场空间，重点在于数据本身将为整个信息化社会赋能。随着大数据技术的落地应用，大数据的价值将逐渐得到体现。目前在互联网领域，大数据技术已经得到了较为广泛的应用。

2. 大数据推动科技领域的发展

大数据技术的发展正在推动科技领域的发展，大数据的影响不仅仅体现在互联网领域，也体现在金融、教育、医疗等诸多领域。在人工智能研发领域，大数据也更加显露出其重要的作用，尤其在机器学习、计算机视觉和自然语言处理等方面，大数据正在成为智能化社会的基础。

3. 大数据产业链逐渐形成

经过近些年的发展，大数据相关产业已经初步形成了一个较为完整的产业链，包括数据采集、整理、传输、存储、分析、呈现和应用，众多企业参与到大数据产业链中，并形成了一定的产业规模，相信随着大数据的不断发展，相关产业规模会进一步扩大。

4. 产业互联网将推动大数据技术落地

当前互联网正在从消费互联网向产业互联网过渡，产业互联网将利用大数据、物联网、人工智能等技术来赋能广大的传统产业，可以说产业互联网的发展空间非常大。而大数据技术则是产业互联网发展的一个重点，大数据技术能否落地到传统行业，关乎产业互联网的发展进程，所以在产业互联网阶段，大数据技术将逐渐落地，也必然落地。

（二）大数据相关人才需求及岗位需求展望

进入数字时代、贯彻新发展理念，全方位、深层次激活数据要素潜能、释放数据要素价值将驱动大数据产业高质量发展。2018~2023 年，中国大数据产业规模呈现逐年增长的趋势；2020 年底，中国大数据产业规模为 6388 亿元，较 2018 年增长 2003.5 亿元；2021 年，中国大数据产业规模为 7512.3 亿元；2022 年，中国大数据产业规模达 1.57 万亿元，同比增长 18%。大数据产业助力数字产业创新发展，成为推动数字经济发展的的重要力量。

大数据产业是战略性新兴产业和知识密集型产业，大数据企业对大数据高端人才和复合人才的需求旺盛。大数据的主要就业方向如下。

1.Hadoop 大数据开发方向

大数据开发工程师的市场需求旺盛，目前 IT 培训机构的重点对应岗位为大数据开发工程师、大数据运维工程师、爬虫工程师、数据分析师等。

2. 数据挖掘、数据分析和机器学习方向

数据挖掘、数据分析和机器学习的起点高、难度大，市面上只有很少的培训机构在做。对应岗位为数据科学家、数据挖掘工程师、机器学习工程师等。

3. 大数据运维和云计算方向

大数据运维和云计算方向更偏向于 Linux、云计算学科，对应岗位为大数据运维工程师。

根据 2020 年大数据人才需求与就业岗位分析，排名靠前的十大岗位如下：大数据架构师、大数据工程师、系统研发人员、数据产品经理、应用开发人员、数据分析师、数据科学家、机器学习工程师、数据挖掘分析师、数据建模师。

拓展阅读

大数据运维工程师

随着大数据产业进入高速发展期，大数据运维工程师已成为近年来的热门职业之一。

1. 工作职责

大数据运维工程师的具体工作职责如下。

- (1) 负责 Hadoop、HBase、Spark、Kafka、Redis 等大数据生态圈的组建。
- (2) 负责处理大数据集群的各类异常和故障，并能区分故障等级，优先处理影响

实时性业务的故障。

(3) 以可控的方式, 高效地完成变更工作, 包括配置管理和发布管理。

(4) 负责大数据平台的容量规划, 扩容。

(5) 负责大数据平台的定期检查, 进行性能分析与调优。

(6) 协助优化大数据平台架构, 支持平台能力和产品不断迭代。

2. 能力素养

大数据运维工程师的能力素养见表 1-4。

表 1-4 大数据运维工程师的能力素养

能力素养	具体要求
部署大数据平台	掌握 Linux 操作系统的基础知识
	掌握 Hadoop、Spark 等集群基础知识
	掌握数据库理论的基础知识
	掌握计算机网络理论的基础知识
	能够安装部署大数据系统, 能够对网络、软硬件进行配置
监控大数据平台	能够安装和使用集群监控工具, 监控大数据平台运行情况
	能够对大数据平台的故障进行定位、分析和解决
	能够将维护日志、故障、问题记录等形成运维报告
优化大数据平台	能够对运维报告进行分析, 根据分析结果对软硬件环境及运维工作机制进行调优
自动化运维能力	具备自动化运维能力, 能够设计和实施自动化运维流程和工具, 提高运维效率和质量。
系统管理能力	具备扎实的系统管理知识, 能够熟练配置、管理和监控大数据集群
	了解操作系统、网络和存储系统的原理
	了解操作系统、网络和存储系统的管理方法
故障排除能力	具备故障排除能力, 能够快速定位和解决大数据系统和应用的故障
	熟悉常见的故障排查工具
	熟悉常见的故障排查方法
安全管理能力	了解大数据安全的基本原理和方法, 能够进行系统安全配置和管理
	具备对数据进行加密、权限管理和审计等执行安全措施的能力。
团队合作能力	具备良好的团队合作能力, 能够与开发人员、数据科学家等其他团队成员紧密合作, 共同解决问题和完成任务。
学习能力和创新意识	具备良好的学习能力和创新意识, 能够不断学习和跟进最新的大数据技术和行业动态, 提供创新的解决方案。

行业动态

大数据产业发展和国家战略: 大数据产业是指大数据的产业集群、产业园区, 涵盖大数据技术产品研发、工业大数据、行业大数据、大数据产业主体、大数据安全保障、大数据产业服务体系等组成的大数据工业园区。

2021年7月13日,中国互联网协会发布了《中国互联网发展报告(2021)》,在大数据领域,2020年我国大数据产业规模达到了718.7亿元,增幅领跑全球数据市场。我国大数据企业主要分布在北京、广东、上海、浙江等经济发达省份。

2021年11月30日,中华人民共和国工业和信息化部发布的《“十四五”大数据产业发展规划》提出,到2025年我国大数据产业测算规模突破3万亿元,年均复合增长率保持在25%左右,创新力强、附加值高、自主可控的现代化大数据产业体系基本形成。大数据产业规模快速增长需要大量人才,预计到2025年,大数据核心人才缺口将高达230万人。所以,不负青春,强国有我!学好大数据技术,为产业发展贡献自己的力量。

思考与练习

一、选择题

- 下面哪一项不是《大数据时代》中总结的大数据的特点?()
A. 海量 B. 多样 C. 新奇 D. 高速
- 目前按照数据类型可将大数据分为三类:结构化数据、半结构化数据和()。
A. 全结构化数据 B. 非结构化数据 C. 零结构化数据 D. 无结构化数据
- 首次提出“大数据”概念的是()公司。
A. 马斯克 B. 麦肯锡 C. 微软 D. Google

二、判断题

- 大数据是指数据量大于某个阈值,并且数据集超出了传统数据库处理能力。
()
- 2012年3月,中华人民共和国科学技术部发布的“十二五国家科技计划信息技术领域2013年度备选项目征集指南”把大数据研究列在首位。
()
- 2022年,中国大数据产业规模达1570亿元,同比增长18%,成为推动数字经济发展的力量。
()

三、实践题

大数据在日常生活中的应用现状调研

1. 实践目的:

了解大数据技术在日常生活中的实际应用情况,培养实践调研能力、数据分析能力和问题解决能力

2. 实践要求:

- (1) 调研问卷或访谈提纲设计要合理,能够全面反映大数据应用现状;
- (2) 调研过程中要保持客观、公正的态度,尊重被调研者的隐私;
- (3) 数据整理和分析要准确、严谨,确保调研结果的可靠性;
- (4) 调研报告要条理清晰、逻辑严密,能够清晰地表达调研结果和结论;
- (5) 汇报展示时要自信、大方,能够准确地传达调研成果和收获。