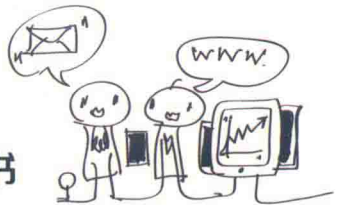
网络工程师教育丛书



Big Data Technologies

大数据技术

◎ 刘化君　吴海涛　毛其林　等编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

7

网络工程师教育丛书

大数据技术

Big Data Technologies

刘化君　吴海涛　毛其林　等编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书是《网络工程师教育丛书》的第 7 册，介绍和讨论大数据的基础知识、技术原理和应用。全书内容分为 6 章，包括绪论、大数据采集和预处理、大数据存储与管理、大数据分析与计算、大数据可视化和大数据应用。本书既介绍大数据技术基础知识，又将这些知识与具体应用有机结合起来，并借助可视化图表深入剖析大数据技术原理和洞见数据价值的方法。各章均配有练习、本章小结及小测验，以便理解掌握重要知识点。另外，考虑到大数据技术涉及许多新名词和专业性极强的词汇，书末以附录形式给出了相关术语的注释，以方便读者查阅。

本书可作为网络工程师培训和认证考试教材，或作为本科及职业技术教育相关课程的教材或参考书，也可供网络技术人员、管理人员以及有志于自学成为网络工程师的读者阅读。

本书的相关资源可从华信教育资源网（www.hxedu.com.cn）免费下载，或通过与本书责任编辑（zhangls@phei.com.cn）联系获取。

图书在版编目（CIP）数据

大数据技术 / 刘化君等编著. —北京：电子工业出版社，2019.7
（网络工程师教育丛书）
ISBN 978-7-121-36706-9

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①数据处理－基本知识 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 103256 号

责任编辑：张来盛（zhangls@phei.com.cn）
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×1092　1/16　印张：18　字数：460 千字
版　　次：2019 年 7 月第 1 版
印　　次：2019 年 7 月第 1 次印刷
定　　价：59.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254467；zhangls@phei.com.cn。

出 版 说 明

人类已进入互联网时代，以物联网、云计算、移动互联网和大数据为代表的新一轮信息技术革命，正在深刻地影响和改变经济社会各领域。随着信息技术的发展，网络已经融入社会生活的方方面面，与人们的日常生活密不可分。我国已成为网络大国，网民数量位居世界第一；但我国要成为网络强国，推进网络强国建设，迫切需要大量的网络工程师人才。然而据估计，我国每年网络工程师缺口约 20 万人，现有网络人才远远无法满足建设网络强国的需求。

为适应网络工程技术人才教育、培养的需要，电子工业出版社组织本领域专家学者和工作在一线的网络专家、工程师，按照网络工程师所应具备的知识、能力要求，参考新的网络工程师考试大纲（2018 年审定通过），共同修订、编撰了这套《网络工程师教育丛书》。

本丛书全面规划了网络工程师应该掌握的技术，架构了一个比较完整的网络工程技术知识体系。丛书的编写立足于计算机网络技术的最新发展，以先进性、系统性和实用性为目标：

- 先进性——全面地展示近年来计算机网络技术领域的新成果，做到知识内容的先进性。例如，对软件定义网络（SDN）、三网融合、IPv6、多协议标签交换（MPLS）、云计算、云存储、大数据、物联网、移动互联网等进行介绍。
- 系统性——加强学科基础，拓宽知识面，各册内容之间密切联系、有机衔接、合理分配、重点突出，按照“网络基础→局域网→城域网与广域网→TCP/IP 基础→网络互连与互联网→网络安全与管理→大数据技术→网络设计与应用”的进阶式顺序分为 8 册，形成系统的知识结构体系。
- 实用性——注重工程能力的培养和知识的应用。遵循“理论知识够用，为工程技术服务”的原则，突出网络系统分析、设计、实现、管理、运行维护和安全方面的实用技术；书中配有大量网络工程案例、配置实例和实验示例，以提高读者的实践能力；每章还安排有针对性的练习和近年网络工程师考试题，并对典型试题和练习给出解答提示，以帮助读者提高应试能力。

本丛书从一开始就搭建了一个真实的、接近网络工程实际的网络，丛书各册均基于这个实例网络的拓扑和 IP 地址进行介绍，逐步完成对路由器、交换机、客户端和服务器的配置、应用设计等，灵活、生动地展现各种网络技术。

本丛书在编写时力求文字简洁，通俗易懂，图文并茂；在内容编排上既系统全面，又切合实际；在知识设计上层次分明、由浅入深，读者可根据自己的需要选择相应的图书进行学习，然后逐步进阶。

鉴于网络技术仍在不断地飞速发展，本丛书将根据需要和读者要求适时更新、完善。热忱欢迎广大读者多提宝贵意见和建议。联系方式：zhangls@phei.com.cn。

电子工业出版社

前　言

人类信息社会正在从 IT（Information Technology）时代快步进入 DT（Data Technology）时代，数据科学与大数据技术以惊人的速度迅猛发展。以大数据为核心研究对象，利用大数据的方法解决具体行业应用问题，成为 DT 时代的核心。数据科学与大数据技术属于新兴的交叉学科，它以统计学、数学、计算机为三大支撑性学科，以生物、医学、环境科学、经济学、社会学、管理学等为应用拓展性学科，由此构架自己的学科领域；其知识体系涵盖了数据采集、分析、处理、数据挖掘算法、计算机编程语言，以及相关软件开发应用等。大数据作为继云计算、物联网之后 IT 行业的又一颠覆性技术，备受人们的关注和青睐。

随着大数据时代的到来，社会急需大批数据科学与大数据技术专业人才。本书编写的目的，就是为培养运用大数据思维洞见数据价值的技术人才提供大数据技术入门指导，为读者架起一座通向“大数据知识空间”的桥梁。在内容取舍上，本书秉承凝练大数据技术的宗旨，着重阐释大数据的基本概念、原理、大数据分析计算及其应用技术，搭建起知识框架，以便形成对大数据知识体系及其应用领域的概括性认识，为读者在大数据领域的继续“深耕细作”奠定基础。

本书紧紧围绕“构建知识体系、阐释工作原理、引导实践应用”的指导思想，对大数据知识体系进行系统梳理，做到面向应用、有序组织，结构合理、层次清楚，并借助可视化图表深入剖析大数据技术及其应用。全书内容分为 6 章，包括绪论、大数据采集和预处理、大数据存储与管理、大数据分析与计算、大数据可视化和大数据应用，每章配有练习、本章小结及小测验，以便于理解掌握重要知识点。

本书可为数据科学与大数据技术、计算机网络和通信领域的教学、科研和工程设计提供参考，适用范围较广；既可以用作数据科学与大数据技术专业教材、网络工程师教育培训，或者作为本科和高职院校相关课程的教材或参考书，也可供大数据技术、网络技术人员和管理人员以及网络爱好者阅读参考。

本书由刘化君、吴海涛、毛其林、顾洪、刘枫编著。其中，刘化君执笔编写第 1 章并负责全书的修改定稿，吴海涛执笔编写第 4 章及第 5 章部分初稿并负责全书内容的审定，毛其林执笔编写第 2、3 章及第 6 章第 3 节初稿，顾洪编写第 5、6 章部分初稿，刘枫执笔编写其余章节有关内容。在编写过程中，得到了许多同志的支持和帮助，参考了大量国内外的教材、专著、论文以及互联网文献资料，在此一并表示衷心感谢！

由于数据科学与大数据技术、计算机网络技术发展很快，囿于作者理论水平和实践经验，书中可能存在不妥之处，恳请广大读者不吝赐教，以便再版时予以订正。

编著者

2019 年 1 月 8 日

目　录

第一章　绪　论

概　述

在信息技术蓬勃发展的时代，物联网方兴未艾，云计算风起云涌，移动互联网崭露头角，大数据初露锋芒，IT 已改变甚至颠覆了社会生活。为紧跟“云物大”（云计算、物联网、大数据）的发展应用，各行各业都在审时度势、精心谋划、系统部署大数据（Big Data）应用工作。毋庸置疑，大数据时代已经来临。然而，对于诸如大数据、云计算之类的热点技术，许多人往往趋之若鹜却又难以说个明白。如果问起“大数据是什么？”也许会随口而出“大数据就是数据大”，或者大谈“4 V”——Volume（体量）、Velocity（速度）、Variety（多样性）、Value（价值密度）来表达自己对大数据的专业理解，而很少有人能够准确说出个一二三来。究其原因，虽然人们对大数据这类新技术有着原始渴求，但真正能参与大数据体验的还比较少，无法勾勒出对大数据及其技术内涵的整体认识。

对于“大数据”，研究机构 Gartner 给出了这样的定义：“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，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来适应海量、高增长率和多样化的信息资产。大数据技术用来从各种各样类型的数据中快速获得有价值的信息；适用于大数据的技术，包括大规模并行处理（MPP）数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统等。

现代信息化社会是一个高速发展的社会，科技发达，信息畅通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的信息化产物，要想较为系统地认知大数据，掌握大数据技术，就必须全面而细致地分解它，深入地描述它。但是，如何认识大数据？如何从大数据中获得价值？分析处理大数据需要哪些技术？本章将针对这些问题勾画一个大数据知识图谱：首先介绍大数据的基本概念，包括大数据的定义、特征和作用；然后探讨大数据的分析计算及其基本技术体系和基础架构——云计算、云数据中心和大数据计算平台，包括典型的开源软件；最后给出构建 Hadoop 平台及其安装与运行的方法，为此后的学习应用奠定基础。

第一节　大数据的概念

随着信息技术和人类生产生活的交汇融合，互联网快速普及，使得全球数据呈现出爆发式增长和海量集聚的特点。大数据正以前所未有的速度颠覆着人们探索世界的方法，引起工业、商业、医学、军事等领域的深刻变革。因此，在当前大数据浪潮的猛烈冲击下，IT 领域迫切需要充实和完善已有的知识、技术结构，提升两种“能力”：一是大数据基本技术与应用能力，使大数据能够为我所用；二是能够挖掘数据之间隐藏的规律与关系，使大数据更好地服务于经济社会发展。

学习目标

- 掌握大数据的基本概念和核心特征；
- 熟悉大数据的主要来源；
- 了解大数据对科学研究、思维方式和社会发展的影响。

关键知识点

- 大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合。

何谓大数据

伴随着信息化社会的到来，日常使用最为频繁的一个术语就是“数据”。数据是什么？数据就是数值，它被看成是现实世界中自然现象和人类活动所留下的轨迹，即人们通过观察、实验或计算得出的结果。数据可以用于科学研究、设计、查证等。在计算机科学中，数据被定义为所有能输入到计算机并被计算机程序处理的符号集，是具有一定意义的数字、字母、符号和模拟量的统称。在计算机科学之外，可以更加抽象地定义数据，如人们通过观察世界中的自然现象、人类活动，都可以形成数据。实际上，数据的形式有很多种，最简单的是数字；数据也可以是文字、图形图像、音频和视频等。人类几千年的历史所产生的所有文明记录，包括历史、文学、艺术、哲学以及一切科学成就，都能够以数据的形式存储和保留。

随着信息科学的发展，在“数据”（Data）、“信息”（Information）、“知识”（Knowledge）和“价值”（Value）4 个词语的相互关联中，“数据”一词呈现的是一种过程、状态或结果的记录，这类记录被数字化后可以被计算机存储和处理。其实，设计计算机的最初目的就是用于数据处理。但计算机需要将数据表达成 0、1 的二进制形式，用一个或若干个字节（Byte，B）来表示。因此计算机对数据的处理，首先需要对数据进行表示和编码，从而衍生出不同的数据类型。对于数字，可以将它编码成二进制形式；对于文本数据，通常采用 ASCII 码将其编码为一个整数；有时候，可能还需要采用更加复杂的数据结构（如向量、矩阵）来表达一个复杂的状态，如表达地图上的位置信息需要用二维坐标。

显然，表达一个实体的不同方面，会用到不同的数据。例如，要描述一个员工，可能会包括姓名、性别、年龄、单位等多种属性，其中每种属性都需要相应类型的数据来表达。有时，如果需要观察一个实体在某一段时间内的状态变化，就可能得到一个时间序列的数据。例如，监测城市空气质量所含有的细颗粒物（PM2.5）时，传感器监测到的数据就会形成一个 PM2.5 随时间变化的数据序列。当信息科学处理的数据发展到 Facebook、Google、百度等的数据规模时，数据本身（类型、规模、属性、用途等）及相关的大规模数据的分析计算就形成了数据科学（Data Science）或数据工程（Data Engineering）这样一门新的学科（领域），进而迎来了大数据时代，使人类拥有了更多的机会和条件在各个领域更深入地全面获得、使用完整数据和系统数据，深入探索现实世界的规律。那么，大数据究竟是什么？

大数据的数据源

近年来，随着信息技术的发展，人们开始越来越频繁地使用“大数据”一词，用以描述和定义信息爆炸时代所产生的海量数据。根据著名国际数据公司（International Data Corporation，

IDC）做出的预测，人类社会产生的数据一直在以平均每年 50%的速度增长，也就是说，每两年就要增加约 1 倍，这被称为“大数据摩尔定律”。估计到 2020 年，全球将总共拥有 35 ZB 的数据量。数据存储单位之间的换算关系如表 1.1 所示。

表 1.1 数据存储单位之间的换算关系

单 位	换 算 关 系
B（Byte，字节）	1 B = 8 bit
KB（Kilobyte，千字节）	1 KB=1 024 B=10^3 B
MB（Megabyte，兆字节）	1 MB=1 024 KB=10^6 B
GB（Gigabyte，吉字节）	1 GB=1 024 MB=10^9 B
TB（Trillionbyte，太字节）	1 TB=1 024 GB=10^{12} B
PB（Petabyte，拍字节）	1 PB=1 024 TB=10^{15} B
EB（Exabyte，艾字节）	1 EB=1 024 PB=10^{18} B
ZB（Zettabyte，泽字节）	1 ZB=1 024 EB=10^{21} B

大数据的来源众多，科学研究、企业应用和 Web 应用等都在源源不断地生成新的数据。生物大数据、交通大数据、医疗大数据、电信大数据、金融大数据等都呈现出“井喷式”增长，大数据的类型丰富多彩。在讨论大数据的定义之前，先了解一下我国海量数据的主要来源和分布领域。

1. 以 BAT 为代表的互联网公司

我国以百度公司（Baidu）、阿里巴巴集团（Alibaba）、腾讯公司（Tencent）三大互联网公司（以其首字母合称为“BAT”）为代表的互联网公司，是产生海量数据的主要来源。

- 百度公司（Baidu）：2013 年的数据总量已接近 1 000 PB，主要来自中文网、百度推广、百度日志、用户原创内容（User Generated Content，UGC）。由于它占有 70%以上的搜索市场份额，因而坐拥庞大的搜索数据。
- 阿里巴巴集团（Alibaba）：目前保存的数据量近 100 PB，其中 90%以上为电商数据、交易数据、用户浏览和点击网页数据、购物数据。
- 腾讯公司（Tencent）：存储数据经压缩处理后总量为 100 PB 左右，数据量月增 10%，主要是大量社交、游戏等领域积累的文本、音频、视频和关系类数据。

2. 电信、“金融与保险”、“电力与石化”系统

- 电信系统：包括用户上网记录、通话、信息、地理位置等，运营商拥有的数据量都在 10 PB 以上，年度用户数据增长数十 PB。
- 金融与保险系统：包括开户信息数据、银行网点和在线交易数据、自身运营的数据等，金融系统每年产生数据达数十 PB，保险系统数据量也接近 PB 级别。
- 电力与石化系统：仅国家电网采集获得的数据总量就达到 10 PB 级别，石化行业、智能水表等每年产生和保存下来的数据量也达数十 PB。

3. 公共安全、医疗、交通领域

- 公共安全领域：在北京就有 50 万个监控摄像头，每天采集视频数量约为 3 PB，整个视频监控每年保存下来的数据在数百 PB 以上。

- 医疗卫生领域：据了解，整个医疗卫生行业一年能够保存下来的数据就可达数百 PB。
- 交通领域：航班往返一次就能产生 TB 级别的海量数据；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十 PB。

4. “气象与地理”“政务与教育”等领域

- 气象与地理领域：中国幅员辽阔，气象局保存的数据为 4～5 PB，每年增加数百 TB，各种地图和地理位置信息每年增加数十 TB。
- 政务与教育领域：各地政务数据资源网涵盖旅游、教育、交通、医疗等门类。据估计，一个市级政务数据资源网每年的上线公告也要达数百个数据包。网络在线教育（如爱课程网的视频课程）的数据规模呈快速上升的发展态势。

5. 其他行业

其他行业（包括线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业）的数据量，还处于积累期，目前整个数据规模还不算大，多则为 PB 级别，少则为几百 TB 或者数十 TB 级别，但增速很快。

以上这些数量巨大、与微观情境相结合的运行记录信息就是大数据吗？显然，运行记录信息不是大数据的全部，只能说是大数据的主体。目前看得到的金融、电信、航空、电商、教育等领域中的大数据，多数都是运行记录信息。

大数据的定义

大数据在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业已存有时日，近年来因互联网和信息行业的发展而引起人们的关注。“大数据”一词开始越来越多地被提及，并用于描述和定义信息爆炸时代所产生的海量数据。“Big Data”（大数据）已经上过《纽约时报》《华尔街日报》的专栏封面，进入了美国白宫的官网新闻。目前，这一专业术语不但现身于国内互联网研究领域，而且被列为加快建设数字中国的国家大数据战略。

最早提出“大数据”时代到来的是全球知名咨询公司——麦肯锡。麦肯锡在《Big Data: The next frontier for innovation, competition and productivity》报告中指出：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。麦青锡对大数据给出的定义是：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、多样的数据类型和低价值密度四大特征。但它同时强调，并不是说一定要超过特定太字节（TB）值的数据集才能算是大数据。

在维克托·迈尔–舍恩伯格、肯尼斯·库克耶编写的《大数据时代》中，大数据不用随机分析法（抽样调查）这样的捷径分析处理，而采用所有数据进行分析处理。全球最具权威的 IT 研究与顾问咨询公司——高德纳咨询公司（Gartner）于 2012 年将大数据的定义修改为：“大数据是大量、高速和（或）多变的信息资产，它需要新型的处理方式去促成更强的决策能力、洞察力与最优化处理。”

亚马逊公司（全球最大的电子商务公司）的大数据科学家 John Rauser 给出了一个简单的定义：大数据是任何超过了一台计算机处理能力的数据量。

维基百科中只有短短的一句话：巨量资料（或称大数据），指的是所涉及的资料量规模巨

大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。在百度百科中是这样定义的：大数据是指无法在可承受的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。

可见，对大数据的定义尚未达成共识，有多少人就有可能有多少个定义。可以从分类的角度将其定义为，“大数据=交易+互动+观察”；也可以根据用途定义，即“大数据=用于实时预测的数据”；当然还可以从技术的角度、业务的角度、数据本身的角度以及娱乐的角度描绘大数据。但究竟如何比较深入、全面地理解大数据呢？其实，大数据并不是一开始就流行起来的，而是在新技术的支持下，尤其是各类先进的开源存储系统或处理工具迅速发展以后，Big Data的概念才得以展现出来。因此，大数据的概念难免不断变化，任何定义都有一定的时间和技术局限性，应该用发展的观点解释大数据，从不同的角度给出多个概念。在此，针对大数据的基本特征描述如下：

- 大数据由巨型数据集（Data Set）组成，这些数据集大小常常超出人类在可接受时间内的收集（Data Acquisition）、废用（Data Curation）、管理和处理能力。大数据必须借助计算机对数据进行统计、比对、解析方能得出客观结果，通过数据挖掘可以获得有价值的信息。这也是“Big Data”一词较为贴切的含义。
- 大数据的大小是相对的，并没有明确的界限。例如，单一数据集的大小从数 TB 不断增至数十 PB 不等。在今天的不同行业中，大数据的范围可以从几 TB 到几 PB，但在 20 年前 1 GB 的数据已然是大数据了。可见，随着计算机软硬件技术的发展，符合大数据标准的数据集容量也会增长。
- 大数据不只是大，它还包含了数据集规模已经超过了传统数据库软件获取、存储、分析和管理能力的意思。

大数据的核心特征

麦塔集团（META Group，现为高德纳）分析员道格·莱尼（Doug Laney）早在 2001 年就在其一份研究报告与相关的演讲中指出，数据增长的挑战和机遇有三个方向：体量（Volume，数据大小）、速度（Velocity，数据输入输出的速度）与多样性（Variety，数据类型多样性），合称“3V”或“3Vs”。高德纳与现在大部分大数据产业中的公司，都还继续使用“3V”来描述大数据。这个“3V”特征从数量、类型、速度 3 个维度描述了大数据的本质构建，如图 1.1 所示。

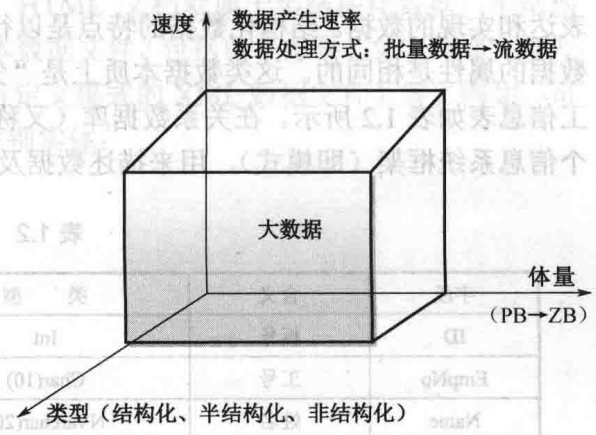


图 1.1 大数据的本质构建

国际数据公司（IDC）提出了大数据的“4V”特征，即海量的数据规模（Volume）、多样的数据类型（Variety）、快速的数据流转和动态的数据体系（Velocity）和数据的低价值密度（Value）。

- Volume（体量）：数据量大，包括采集、存储和计算的量都非常大。大数据的起始计量单位至少是 PB（10^3 TB）、EB（10^6 TB）或 ZB（10^9 TB）。

- Variety（多样性）：数据来自多种数据源，数据种类和格式多样化，囊括了结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等。多类型的数据对数据的处理能力提出了更高的要求。
- Velocity（速度）：数据增长速度快，处理速度也快，时效性要求高。对于实时的数据输入、处理与丢弃，分析结果立竿见影而非事后见效。例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。
- Value（价值密度）：大数据中含有大量不相关信息，价值密度相对较低，但可由其进行预测分析。随着互联网以及物联网的广泛应用，信息感知无处不在，数据呈海量但价值密度较低，可以结合业务逻辑并通过强大的深度复杂分析（机器学习、人工智能等）来挖掘数据价值。

综上所述，大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合。

大数据结构类型

大数据是互联网发展到现今阶段的一种表象或特征。当今企业存储的数据不仅内容多，而且其结构已发生了极大改变，不再仅仅以二维表的规范结构存储。根据数据所刻画的过程、状态和结果等特点，可以将数据划分为不同的类型。按照数据是否有强的结构模式，可以将其划分为结构化数据、半结构化数据和非结构化数据。

结构化数据

结构化数据是指数据经过分析后可分解成多个互相关联的组成部分，各组成部分间有明确的层次结构，其使用和维护通过数据库进行管理，并有一定的操作规范。通常我们所接触的数据，包括生产、业务、交易、客户信息等方面的记录，都属于结构化数据。

简单来说，结构化数据就是存储在结构化数据库里的数据，可以用二维表结构来进行逻辑表达和实现的数据。结构化数据的特点是以行为单位，一行数据表示一个实体的信息，每一行数据的属性是相同的。这类数据本质上是“先有结构，后有数据”。例如，记录员工信息的职工信息表如表 1.2 所示。在关系数据库（又称关系型数据库）或面向对象数据库中，都存在一个信息系统框架（即模式），用来描述数据及其相互关系，其特点是模式与数据完全分离。

表 1.2 职工信息表

字段	含义	类　型	是否可为空	备　注
ID	标号	Int	否	关键字
EmpNo	工号	Char(10)	否	
Name	姓名	NVarchar(20)	否	
Age	年龄	Int	否	
Sex	性别	Int	否	1——男，2——女
BirthDay	生日	NVarchar(20)	是	

多年来，结构化数据一直主导着信息技术（Information Technology，IT）和产业的应用，是联机事务处理过程（On-line Transaction Processing，OLTP）系统业务所依赖的信息。结构化数据还可对结构化数据库信息进行排序和查询。

另外，还有一种准结构化数据格式，是指具有不规则数据格式的文本数据，使用工具可以使之格式化。例如，包含不一致的数据值和格式化的网站点击数据。

非结构化数据

非结构化数据是相对于结构化数据而言的，就是没有固定结构的数据，指不方便用数据库二维逻辑表来表现的数据，包括所有格式的办公文档、文本、图片，标准通用标记语言下的子集 XML、HTML，各类报表、图像和音/视频信息等。本质上可认为，非结构化数据主要是位映射数据。据 IDC 的调查报告显示：企业中 80%的数据都是非结构化数据，这些数据每年都增长 60%。非结构化数据越来越成为数据的主要部分。

存储和处理非结构化数据通常要用到专用逻辑，一般直接整体进行存储，而且一般存储为二进制的数据格式。非结构化数据没有固定的数据模型，因此不能被直接处理或者用 SQL 语句查询。如果需要把它们存储在关系数据库中，就需要以二进制大型对象（BLOB）形式存储在表中。因此，需要非结构化数据库来承担存储任务。

非结构化数据库的变长记录由若干不可重复和可重复的字段组成，而每个字段又可由若干不可重复和可重复的子字段组成。利用非结构化数据库，不仅可以处理结构化数据（如数字、符号等信息）而且更适合处理非结构化数据（全文文本、图像、声音、影视、超媒体等信息）。简单地说，非结构化数据库就是字段可变的数据库。NoSQL 数据库就是一个非结构化数据库，它是非关系数据库，能够用来同时存储结构化和非结构化数据。

半结构化数据

半结构化数据是指介于结构化数据（如关系数据库、面向对象数据库中的数据）和非结构化数据（如声音、图像文件等）之间的实时数据，HTML 文档就属于半结构化数据。半结构化数据是具有可识别的模式并可以进行解析的文本数据文件，包括电子邮件、文字处理文件及大量保存和发布在网络上的信息（即自描述和具有定义模式的 XML 数据文件）等。例如，可以用如下 XML 语言来描述序号、姓名、年龄和性别实体：

```
<person>
    <id>1</id>
    <name>张菲</name>
    <age>18</age>
    <gender>女</ gender >
</person>
```

在这个示例中，属性的顺序是不重要的，如果有的实体部分信息缺失，如年龄信息缺失或者性别信息缺失，数据集中也可以不包含这一属性。XML 是一个典型的用树形结构组织信息的方式。也就是说，半结构化数据一般是自描述的，数据的结构和内容混在一起，没有明显的区分。半结构化的数据模型通常表现为树、图结构。

目前，大量的数据已不仅仅是结构化数据，而是兼有半结构化数据或非结构化数据，如办

公文档、文本、图片、XML、HTML、各类报表、图片、音频和视频等，并且这些数据在企业的所有数据中是大量的且迅速增长的。例如，在互联网上出现的海量信息，通常包含有结构化、半结构化和非结构化三种类型的信息。

- 结构化信息（如电子商务信息），其性质、量值出现的位置是固定的；
- 半结构化信息（如专业网站上的细分频道），其标题和正文的语法相当规范，关键词的范围相当有限；
- 非结构化信息，如博客（BLOG）和网上社区论坛，所有内容都是不可预知的。

此外，还有一种重要的数据类型，称为元数据。元数据主要由机器产生，例如 XML 文件中提供作者和创建日期信息的标签，数码照片中提供文件大小和分辨率的属性文件等，并且能够添加到数据集中。搜寻元数据对大数据存储、处理和分析都很重要，因为它提供了数据系谱信息，以及数据处理的起源。

大数据的作用和影响

大数据作为一种重要的战略资产，已经不同程度地渗透到各行各业。目前，大数据对科学研究、人们的思维方式和社会发展都产生了重要而深远的作用及影响。

促进科学研究方法、手段发生新变化

图灵奖获得者、著名数据库专家吉姆·格雷（Jim Gray）博士观察并总结认为：人类自古以来在科学研究上先后经历了实验科学、理论科学、计算科学和数据密集型科学四种范式。随着数据的不断积累，其宝贵价值日益得到体现。在大数据环境下，一切将以数据为中心，从数据中发现问题、解决问题。

大数据时代科学研究的方法和手段将发生重大改变。例如，抽样调查是社会科学的基本研究方法。在大数据时代，可通过实时监测、跟踪研究对象在互联网上产生的海量行为数据来进行挖掘分析，揭示出规律性的东西，提出研究结论和对策。

大数据也将会催生新的学科和行业。数据科学将成为一门专门的学科，被越来越多的人所认知。越来越多的高等院校已经开设了与大数据相关的学科专业及相应的课程，为市场和企业培养数据科学专业技术人才。

启动信息产业发展新引擎

大数据将成为信息产业持续高速增长的新引擎。面向大数据市场的新技术、新产品、新服务、新业态会不断涌现。

- 在硬件与集成设备领域，大数据将对芯片、存储产业产生重要影响，还将催生一体化数据存储处理服务器、内存计算等市场；
- 在软件与服务领域，大数据将引发数据快速处理分析、数据挖掘技术和软件产品的发展。

创造经济和社会发展高效益

大数据将会对社会发展产生深远的影响。对大数据的分析处理正成为新一代信息技术融合