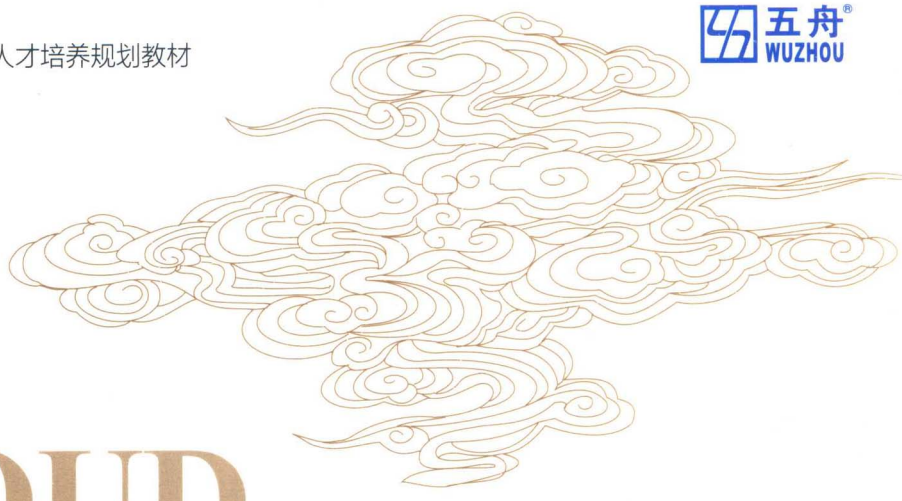
21 世纪高等院校
云计算和大数据人才培养规划教材

五舟 WUZHOU



CLOUD COMPUTING AND BIG DATA

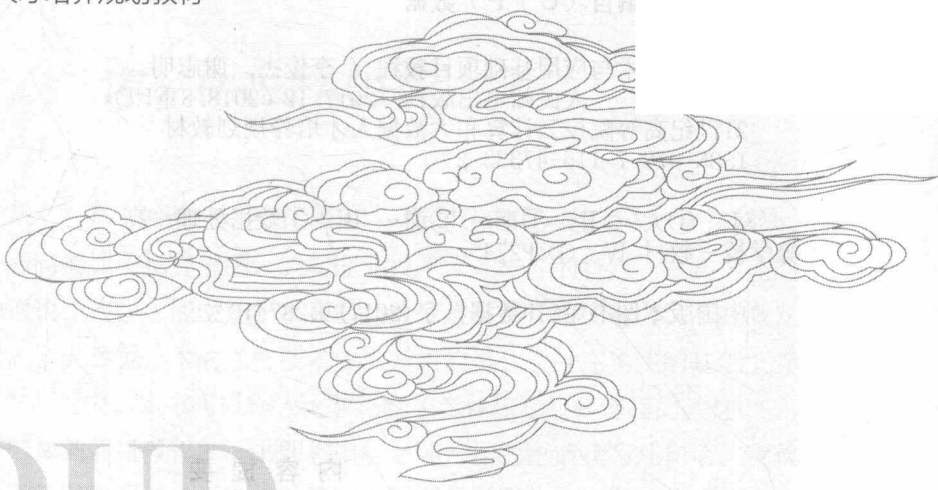
大数据技术与应用基础项目教程

李俊杰 谢志明 ◎ 主编
肖政宏 石慧 谢高辉 杨泽强 ◎ 副主编

- 内容基础、案例简单、实操性强、举一反三
- 将实验环节及实操内容融入到各个知识点与课程教学中
- 以项目实战为主线，循序渐进，逐步深入

中国工信出版集团
人民邮电出版社 POSTS & TELECOM PRESS

21 世纪高等院校
云计算和大数据人才培养规划教材



CLOUD
COMPUTING

大数据技术与应用基础项目教程

李俊杰 谢志明 ◎ 主编
肖政宏 石慧 谢高辉 杨泽强 ◎ 副主编

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

大数据技术与应用基础项目教程 / 李俊杰，谢志明主编. -- 北京 : 人民邮电出版社，2017.12（2018.8重印）
21世纪高等院校云计算和大数据人才培养规划教材
ISBN 978-7-115-47333-2

Ⅰ. ①大… Ⅱ. ①李… ②谢… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第287410号

内 容 提 要

全书共十个项目，除了项目一介绍大数据基础理论外，其余项目均以实战为主线，内容循序渐进，逐步深入，围绕大数据技术的应用层层展开。内容主要包括大数据的基本概念、Ubuntu 及服务安装配置、Hadoop 集群部署、MapReduce 编程、HBase 数据库部署与应用、Hive 数据仓库安装与应用、Pig 数据分析、Sqoop 数据迁移、Spark 部署及数据分析等知识，最后以大数据技术的具体应用介绍了 MapReduce 大数据编程、Mahout 的 K-Means 计算、决策树和随机森林的分类预测、频繁项集运算和关联分析等知识。本书秉承“实践为主、理论够用，注重实用”原则，将实验环节及实操内容融入各个知识点与课程教学中，以便读者能更好地学习和掌握大数据关键技术。

本书可作为院校计算机专业的教材，也可作为培训机构云计算和大数据等相关课程的培训用书，还可作为相关技术人员的参考书。

◆ 主　　编　李俊杰　谢志明
副 主 编　肖政宏　石　慧　谢高辉　杨泽强
责任编辑　左仲海
责任印制　马振武

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：787×1092　1/16
印张：19　　2017 年 12 月第 1 版
字数：455 千字　　2018 年 8 月河北第 2 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

序 PREFACE

大数据产业的发展，人才培养是关键，特别是用得上、留得住的职业技术人才更是稀缺。本书的作者以面向应用、面向实战为指导思想，紧扣职业技术人才培养的特点，在知识点的讲解和实验中避免复杂的理论，使学生能快速上手体验、验证大数据系统的魅力，不断激发学生的学习兴趣。本书覆盖了大数据技术的主要技术要点，目的是让学生在实践的基础上养成在大数据系统环境下工作的习惯。本书可以作为大数据相关专业核心实践课程的教材。

在《大数据技术与应用基础项目教程》一书即将出版之际，作者邀请我写几句话，我欣然应允。本书的作者来自于汕尾职业技术学院云计算与大数据教学团队，由于工作的关系，我在汕尾职业技术学院挂职两年，见证了这个团队艰难成长的历程。汕尾职业技术学院地处广东省相对欠发达的粤东地区，发展条件和资源远远落后于珠三角地区的职业院校，但在云计算、大数据这一新的技术概念提出之后，汕尾职业技术学院的领导和老师敏锐地意识到这一新兴技术所带来的对相关职业技术人才需求的爆发，率先设立了云计算、大数据专业，并与广州五舟科技股份有限公司发起成立了广东省高等教育学会高职高专云计算与大数据专业委员会，并编写了国内第一本面向高职高专的云计算与大数据专业的国家级规划教材。在新专业建设的初期，师资问题是困扰专业发展的一个核心问题，本书的作者团队在最为困难的时候承担起了专业建设的重任，相继建设了多门云计算和大数据专业核心课程，所培养的学生在国内云计算和大数据竞赛中屡获佳绩，使汕尾职业技术学院在这一新专业建设上实现了弯道超车，成为广东省云计算、大数据专业建设的示范和标杆。

数据科学的春天已经到来，大数据、云计算、人工智能、高性能计算技术已呈现出相互交织和相互促进的局面，本书的出版正是响应了这一新时代的召唤，我想是时候我们共同去拥抱这一新技术时代的到来了。

广东省高等教育学会高职高专云计算与大数据专业委员会理事长
西南民族大学计算机科学与技术学院
并行计算实验室　王鹏

前言 FOREWORD

近年来，国家陆续出台了许多与大数据有关的各类专项支持政策，并把大数据列为国家重点支持和发展的战略新兴产业，大数据的迅速发展已然成为当今科技界、企业界甚至世界各国政府关注的热点。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为国家间和企业间新的争夺焦点。全球著名管理咨询公司麦肯锡率先提出“大数据时代”的到来，并声称：“数据已经渗透到当今各行各业的职能领域，成为重要的生产因素。”

本书的体系结构及知识点的分布按照学习思维逻辑由浅入深、循序渐进、以操代教的模式编排，编者建议初学者尽可能按照章节编排顺序学习和开展实训，这样有助于较为全面地了解大数据技术及其应用。为了快速提升学习者掌握大数据技术实战的能力，我们在每个项目或任务学习后安排多个精选的具有代表性的应用实例作为同步训练。全书注重实用性，图文并茂，知识点也以精练为主，读者在学习每一任务之时只需花上少许时间学习该任务的知识点即可直接进行实操或实训。

本书由广东省高等教育学会高职高专云计算与大数据专业委员会牵头，并组织专家、教师与企业共同参与编写，是全国高等院校计算机基础教育研究会2016年度高职科研规划纵向课题（课题编号：2016GHB02025、2016GHB02005）、广东省高等职业教育质量工程教育教学改革项目（课题编号：GDJG2015245、GDJG2015244）和广东省高职教育信息技术教指委教改项目（课题编号：XXJZW2015002）、广东省高等教育学会高职高专云计算与大数据专业委员会教育科研课题（课题编号：GDYJSKT16-01、GDYJSKT16-03、GDYJSKT16-05）、汕尾职业技术学院教学改革与科研重点立项课题（课题编号：SWKT15-002、SWKT16-002、swjy15-004、swjy16-012）的科研成果。本书编写期间得到了广东技术师范学院、广州五舟科技股份有限公司、汕尾市创新工业设计研究院的鼎力支持，同时还得到了汕尾职业技术学院各处系领导和老师的关怀和支持，正因为有了他们的支持和帮助，我们才能如期完成本书的撰写和编排工作。

大数据技术涉及面很广且更新较快，编者经过大量的重复实验与多年的工作经验积累，参考并引用了无数前辈学者的研究成果、论述，并在与之相关的科学研究上不断扩充和改进，编者在此向这些前辈学者深表敬意。大数据技术作为一门正在高速发展的新兴技术日新月异，新技术、新方法、新架构层出不穷，加之编者的经验和水平有限，本书的结构、内容肯定存在诸多疏漏和不妥之处，诚恳接受读者的批评与指正。如有任何问题和建议，可发送电子邮件至441814853@qq.com，以便我们及时修正完善。

为方便读者学习和教学需要，本教材配备了大量的电子资源，欢迎读者登录人民邮电出版社教育社区（http://www.ryjiaoyu.com）免费下载使用，同时欢迎相关课程的教师加入云计算大数据HPC教育QQ群（321168742）讨论交流。读者还可以通过发送邮件给编者以获得更多下载资源（资源配套邮箱：gdswyun@126.com）。

感谢您使用本书，期待本书能成为您的良师益友。

编　者

2017年6月于云计算与数据中心工程设计研究所

目 录 CONTENTS

PART 1

项目一 走进大数据

【项目介绍】

本项目的主要目标是理解大数据的内涵和定义，学会常用的大数据的计算模式和厘清大数据处理的基本流程，了解大数据的应用情况及对当今社会各界产生的影响，乐观看待大数据发展所带来的一系列问题，积极面对各种挑战等。

本项目分为以下 6 个任务：

- 任务 1　概述大数据的内涵
- 任务 2　关注大数据的影响
- 任务 3　认识常见的大数据计算模式
- 任务 4　厘清大数据处理的基本流程
- 任务 5　大数据应用大显神通
- 任务 6　大数据的发展及面临的挑战

【学习目标】

一、知识目标

- 了解大数据的基本概念、产生的原因及其特性。
- 理解科学研究的 4 种范式。
- 清楚了解大数据对社会发展及思维方式产生的影响。
- 学会常用的几种大数据计算模式。
- 掌握大数据处理的基本流程，如数据采集、数据清洗、数据分析和数据解释等过程。
- 了解大数据在各个领域的应用情况，熟悉大数据推荐系统及搜索引擎系统等。
- 了解大数据的发展历程及发展现状，积极面对各种挑战。

二、能力目标

- 能够从大数据角度理解大数据对生产力的影响。
- 学会按大数据处理流程挖掘分析有价值的信息。
- 能够理解使用大数据计算模式。
- 能够理解使用大数据推荐系统及搜索引擎系统。

- 能够理解大数据的发展及面临的挑战。

任务 1　概述大数据的内涵

【任务概述】

大数据已成为社会各界研究及关注的焦点。本任务着重介绍大数据的内在含义，其中包括大数据的多种定义表述、大数据产生的原因、大数据特性的演进以及在大数据时代才出现的一些数据计量单位。

【支撑知识】

近几年，大数据迅速发展成为科技界和企业界甚至世界各国政府关注的热点。人们对于大数据的挖掘和运用，预示着新一波生产力增长和消费盈余浪潮的到来。美国政府认为大数据是“未来的钻石矿和新石油”，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为国家间和企业间新的争夺焦点。全球著名管理咨询公司麦肯锡（McKinsey&Company）首先提出了“大数据时代”的到来并声称：“数据已经渗透到当今各行各业的职能领域，成为重要的生产因素。”

数据的产生方式由“人机”“机物”的二元世界向着融合社会资源信息系统及物理资源的三元世界转变，数据规模呈膨胀式发展，例如，互联网领域中，谷歌搜索引擎的每秒使用用户量达到 200 万；科研领域中，仅某大型强子对撞机在一年内积累的新数据量就达到 15PB 左右；电子商务领域中，eBay 的分析平台每天处理的数据量高达 100PB，超过了纳斯达克交易所每天的数据处理量；“双十一”大型商业活动中，淘宝商城屡创神话，销售额由 2010 年的 9 亿元一路攀升到现今的 1200 多亿元，支付宝平台平均每秒成功交易 12 万笔，交易覆盖 235 个国家和地区；航空航天领域中，仅一架双引擎波音 737 飞机在横贯大陆飞行的过程中，传感器网络便会产生近 240TB 的数据。综合各个领域，目前积累的数据量已经从 TB 量级上升至 PB、EB 甚至已经达到 ZB 量级，其数据规模已经远远超出了现有通用计算机所能够处理的量级。

根据全球著名咨询机构互联网数据中心（Internet Data Center，IDC）做出的估测，人类社会产生的数据一直都在以每年 50%的速度增长，也就是说，每两年数据量就会增加一倍，即已形成了“大数据摩尔定律”，这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量之和。据 IDC 统计，2011 年全球被创建和复制的数据总量为 1.8ZB，到 2020 年这一数据将攀升到 40ZB，是 2012 年的 12 倍。而我国的数据量到 2020 年将超过 8ZB，是 2012 年的 22 倍。其中 80%以上来自于个人（主要是图片、视频和音乐），远远超过人类有史以来所有印刷材料的数据总量（200PB）。目前，全球的数据量正以每 18 个月翻一番的速度呈膨胀式增长，数据量的飞速增长同时也带来了大数据技术和服务市场的繁荣发展。

一、大数据的定义

“大数据”一词由英文“Big Data”翻译而来，是近几年兴起的概念。往前追溯却发现由来已久，早在 1980 年就已由美国著名未来学家阿尔文·托夫勒在《第三次浪潮》一书中，将大数据赞颂为“第三次浪潮的华彩乐章”。

“大数据”并不等同于“大规模数据”，那么何谓“大数据”呢？迄今并没有公认的定义，由于大数据是相对概念，因此，目前的定义都是对大数据的定性描述，并未明确定量指标。维基（Wiki）百科从处理方法角度给出的大数据定义，即大数据是指利用常用软件工具捕获管理和处理数据所耗时间超过可容忍时间限制的数据集。麦肯锡公司认为将数据规模超出传统数据库管理软件的获取存储管理，以及分析能力的数据集称为大数据；高德纳咨询公司（Gartner）则将大数据归纳为需要新处理模式才能增强决策力、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；徐宗本院士在第 462 次香山科学会议上的报告中，将大数据定义为不能够集中存储并且难以在可接受时间内分析处理，其中个体或部分数据呈现低价值性而数据整体呈现高价值的海量复杂数据集。虽说这些关于大数据定义的定义方式角度及侧重点不同，但是所传递的信息基本一致，即大数据归根结底是一种数据集，其特性是通过与传统的数据管理及处理技术对比来凸显，并且在不同需求下，其要求的时间处理范围具有差异性，最重要的一点是大数据的价值并非来自数据本身，而是来自由大数据所反映的“大决策”“大知识”“大问题”等。

从宏观世界角度来看，大数据则是融合物理世界、信息空间和人类社会三元世界的纽带，因为物理世界通过互联网、物联网等技术有了在信息空间中的大数据反映，而人类社会则借助人机界面、脑机界面、移动互联等手段在信息空间中产生自己的大数据映像。从信息产业角度来讲，大数据还是新一代信息技术产业的强劲推动力。所谓新一代信息技术产业，本质上是构建在第三代平台上的信息产业，主要是指云计算、大数据、物联网、移动互联网（社交网络）等。

二、大数据产生的原因

“大数据”并不是一个凭空出现的概念，其出现对应了数据产生方式的变革，生产力决定生产关系的道理对于技术领域仍然是有效的，正是由于技术发展到了一定的阶段才导致海量数据被源源不断地生产出来，并使当前的技术面临重大挑战。归纳起来大数据出现的原因有以下几点。

（1）数据生产方式变得自动化

数据的生产方式经历了从结绳计数到现在的完全自动化，人类的数据生产能力已不可同日而语。物联网技术、智能城市、工业控制技术的广泛应用使数据的生产完全实现了自动化，自动数据生产必然会产生大量的数据。甚至当前人们所使用的绝大多数数字设备都可以被认为是一个自动化的数据生产设备：我们的手机会不断与数据中心进行联系，通话记录、位置记录、费用记录都会被服务器记录下来；我们用计算机访问网页时访问历史、访问习惯也会被服务器记录并分析；我们生活的城市、小区遍布的传感器、摄像头会不断产生数据并保证我们的安全；天上的卫星、地面的雷达、空中的飞机也都在不断地自动产生着数据。

（2）数据生产融入每个人的日常生活

在计算机出现的早期，数据的生产往往只是由专业的人员来完成的，能够有机会使用计算机的人员通常都是因为工作的需要，物理学家、数学家是最早一批使用计算机的人员。随着计算机技术的高速发展，计算机得到迅速普及，特别是手机和移动互联网的出现使数据的生产和每个人的日常生活结合起来，每个人都成为数据的生产者：当你发出一条微博时，你

在生产数据；当你拍出一张照片时，你在生产数据；当你使用手中的市民卡和银行卡时，你在生产数据；当你在 QQ 上聊天时，你在生产数据；当你在用微信发朋友圈或聊天时，你在生产数据；当你在玩游戏时，你在生产数据。数据的生产已完全融入人们的生活：在地铁上，你在生产数据；在工作单位，你在生产数据；在家里，你也在生产数据。个人数据的生产呈现出随时、随地、移动化的趋势，我们的生活已经是数字化的生活，如图 1-1 所示。

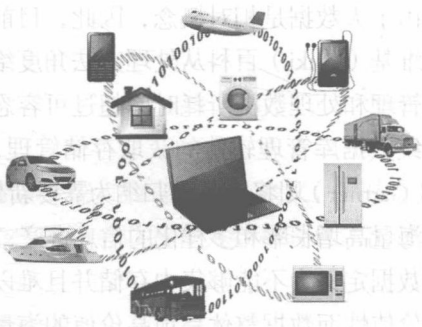


图 1-1　数据生产融入人们的生活

（3）图像和音视频数据所占比例越来越大

人类在过去几千年主要靠文字记录信息，而随着技术的发展，人类越来越多地采用视频、图像和音频这类占用空间更大、更形象的手段来记录和传播信息。从前聊天我们用文字，现在用微信和视频，人们越来越习惯利用多媒体方式进行交流，城市中的摄像头每天都会产生大量视频数据，而且由于技术的进步，图像和视频的分辨率变得越来越高，数据变得越来越大。

（4）网络技术的发展为数据的生产提供了极大的方便

前面说到的几个大数据产生原因中还缺乏一个重要的引子：网络。网络技术的高速发展是大数据出现的重要催化剂：没有网络的发展就没有移动互联网，我们就不能随时随地实现数据生产；没有网络的发展就不可能实现大数据视频数据的传输和存储；没有网络的发展就不会有现在大量数据的自动化生产和传输。网络的发展催生了云计算等网络化应用的出现，使数据的生产触角延伸到网络的各个终端，使任何终端所产生的数据能快速有效地被传输并存储。很难想象在一个网络条件很差的环境下能出现大数据，所以，可以这么认为：大数据的出现依赖于集成电路技术和网络技术的发展，集成电路为大数据的生产和处理提供了计算能力的基础，网络技术为大数据的传输提供了可能。

（5）云计算概念的出现进一步促进了大数据的发展

云计算这一概念是在 2008 年左右进入我国的，而最早可以追溯到 1960 年人工智能之父麦卡锡所预言的“今后计算机将会作为公共设施提供给公众”。2012 年 3 月在国务院政府工作报告中云计算被作为附录给出了一个政府官方的解释，表达了政府对云计算产业的重视，在政府工作报告中云计算的定义是这样的：“云计算：是基于互联网的服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。是传统计算机和网络技术发展融合的产物，它意味着计算能力也可作为一种商品通过互联网进行流通。”云计算的出现使计算和服务都可以通过网络向用户交付，而用户的数据也可以方便地利用网络传递，云计算这一模式网络的作用被进一步凸显出来，数据的生产、处理和传输可以利用网络快速地进行，改变传统的数据生产模式，这一变化大大加快了数据的产生速度，对大数据的出现起到了至关重要的作用。

三、大数据特性

在大数据的定义中，已经包含了大数据的特性，即数据量大、处理速度要求快、价值密度低等，目前对于大数据的特性认可度较高的是 3V 特性：数据的规模性（Volume）、高速性（Velocity）及数据结构多样性（Variety），而在此基础上已经有不同的公司及研究机构对其进

行了扩展，大数据特性描述的演化如表 1-1 所示。

表 1-1 大数据特性描述的演化情况

特　点	提出时间	作者或机构	内　涵
规模性（Volume）	2001 年	Doug Laney（高德纳咨询公司）	体量大，数据量级可达 TB、PB 乃至 EB 以上
高速性（Velocity）			数据分析和处理速度快，俗称“秒级定律”
多样性（Variety）			数据类型多样
价值性（Value）	2012 年	咨询机构 IDC	价值稀疏性，即具有高价值低密度的特点
真实性（Veracity）	2012 年	IBM（国际商业机器公司）	数据反映客观事实
易变性（Variability）	2012 年	Brian Hopkins&Boris Evelson（弗雷斯特研究公司）	大数据具有多层结构

由表 1-1 可以看出，随着时间的演化，业界对于大数据的认识也更深入、全面。除以上对大数据特性的通用性描述之外，不同应用领域的大数据的具体特性也存在差异性。如互联网领域需要实时处理和分析用户购买行为，以便及时制定推送方案，返回推荐结果来迎合和激发用户的消费行为，精度及可靠性要求较高；医疗领域需要根据用户病例及影像等信息判断病人的病情，由于其与人们的健康息息相关，所以，其精度及可靠性要求非常高。表 1-2 列举了不同领域大数据的具体特点及应用案例。

表 1-2 不同领域大数据的具体特点及应用案例

领　域	用户数目	响应时间	数据规模	可靠性要求	精度要求	应用案例
科学计算	小	慢	TB	一般	非常高	大型强子对撞机数据分析
金融	大	非常快	GB	非常高	非常高	信用卡营销
医疗领域	大	快	EB	非常高	非常高	病历、影像分析
物联网	大	快	TB	高	高	迈阿密戴德县的智慧城
互联网	非常大	快	PB	高	高	网络点击流入侵检测
社交网络	非常大	快	PB	高	高	Facebook、QQ 等结构挖掘
移动设备	非常大	快	TB	高	高	可穿戴设备数据分析
多媒体	非常大	快	PB	高	一般	史上首部大数据制作的电视剧《纸牌屋》

由表 1-2 可以看出，不同应用领域的数据规模、用户数目及精度要求等均存在较大的差异，例如，互联网领域与人的正常活动息息相关，其数据量达 PB 级别，用户数目非常大，而且以用户实时性请求为主。与此不同，在科研领域中，其用户数目相对较少，产生的数据量级别在 TB 级。因此，对大数据后续的分析及处理必须因地制宜，才能实现大数据价值的最大化。

四、数据的计量

大数据出现后人们对数据的计量单位也逐步变化，常用的 KB、MB 和 GB 已不能有效地描述大数据。在大数据研究和应用时我们经常会接触到数据存储的计量单位。下面对数据存储的计量单位进行介绍。

计算机学科中一般采用 0、1 这样的二进制数来表示数据信息，信息的最小单位是 bit（比特），一个 0 或 1 就是一个比特，而 8bit 就是一字节（Byte），如 10010111 就是一 Byte。习惯上人们用大写的 B 表示 Byte。信息的计量一般以 2^{10} 为一个进制，如 1024Byte=1KB（KiloByte，千字节），更多常用的数据单位换算关系如表 1-3 所示。

表 1-3　数据存储单位之间的换算关系

单位名称	换算关系
Byte（字节）	1 Byte=8 bit
KB（KiloByte，千字节）	1 KB=1024 Byte
MB（MegaByte，兆字节）	1 MB=1024 KB
GB（GigaByte，吉字节）	1 GB=1024 MB
TB（TeraByte，太字节）	1 TB=1024 GB
PB（PetaByte，拍字节）	1 PB=1024 TB
EB（ExaByte，艾字节）	1 EB=1024 PB
ZB（ZettaByte，泽字节）	1 ZB=1024 EB
YB（YottaByte，尧字节）	1 YB=1024 ZB
BB（Brontobyte，珀字节）	1 BB=1024 YB
NB（NonaByte，诺字节）	1 NB=1024 BB
DB（DoggaByte，刀字节）	1 DB=1024 NB

目前市面上主流的硬盘容量大都为 TB 级，典型的大数据一般都会用到 PB、EB 和 ZB 这 3 种单位。

任务 2　关注大数据的影响

【任务概述】

大数据对科学研究、思维方式和社会发展都具有重要而深远的影响。本任务除了重点介绍曾为大数据做出卓越贡献的科学家之外，还着重介绍了大数据所带来的影响，其中影响较深的有大数据对科学研究的影响及大数据对社会发展的影响，主要体现在大数据改变了科学研究的思维方式、大数据改变了人们的生存方式、大数据改变了人类的生产方式。

【支撑知识】

大数据对科学研究、思维方式和社会发展都具有重要而深远的影响。在科学研究方面，大数据使得人类科学研究在经历了实验、理论、计算 3 种范式之后，迎来了第四种范式——数据；在思维方式方面，大数据具有“全样而非抽样、效率而非精确、相关而非因果”三大显著特征，完全颠覆了传统的思维方式；在社会发展方面，大数据决策逐渐成为一种新的决

策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现；在就业市场方面，大数据的兴起使得数据科学家成为热门职业；在人才培养方面，大数据的兴起，将在很大程度上改变我国高校计算机信息技术相关专业的现有教学和科研体制。

一、大数据之父——吉姆·格雷（Jim Gray）

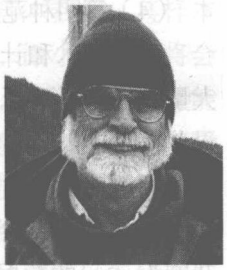


图 1-2 吉姆·格雷

云计算和大数据是密不可分的两个概念，云计算时代网络的高度发展，使每个人都成为数据产生者，物联网的发展更是使数据的产生呈现出随时、随地、自动化、海量化的特征，大数据不可避免地出现在了云计算时代。吉姆·格雷（见图 1-2）生于 1944 年，在著名的加州大学伯克利分校计算机科学系获得博士学位，是声誉卓著的数据库专家、1998 年度的图灵奖获得者。2007 年 1 月 11 日在美国国家研究理事会计算机科学与通信分会上吉姆·格雷明确地阐述了科学研究第四范式——“数据密集型科学”，认为依靠对数据分析挖掘也能发现新的知识，其实质是科学研究将从以计算为中心向以数据为中心转变，即数据思维的到来。这一认识吹响了大数据前进的号角，计算应用于数据的观点在当前的云计算大数据系统中得到了大量的体现。在发表这一演讲后的十几天，2007 年 1 月 28 日格雷独自驾船出海就再也没有了音信，虽然经多方努力搜寻，也没有发现他的一丝信息，人们再也没能见到这位伟大的天才科学家。

二、大数据对科学研究的影响

第四范式的命名是与之前的 3 种科学范式“实验科学”“理论科学”“计算科学”相呼应和一脉相承的，是人类在科学研究领域上新的发现与突破。这 4 种范式在不同时代或时期都给人类社会带了巨大的财富与文明，是人类发现世界、探索世界的利器，下面将分别对这 4 种范式进行简明扼要的表述，如图 1-3 所示。

（1）第一种范式：观测与实验科学

出于好奇的天性，人类一直都在不断地认识自己所生活的世界。最早人类通过自己的观察来认知这个世界，发现了火能烤熟食物、石头能够凿开坚果，发现月亮有阴晴圆缺。随着知识的不断积累，人类开始将之前通过观察和实验得到的感性认识总结为理论。伽利略在比萨斜塔进行“两个大小不同的铁球同时落地”的实验推翻了持续 1900 年之久的亚里士多德“物体下落速度和重量成比例”的学说，这就是人类的认识由感性经验上升到理性理论的重要实验之一。

（2）第二种范式：理论科学

随着科学的进步，人类开始采用各种数学、几何、物理等理论，有了理论以后人类可以用理论来分析世界、预测世界和寻求科学的解决方案。我们有了历法，能够预言一年四季，能够指导春耕秋收，能预测尚未被发现的行星，譬如，海王星、冥王星的发现就不是通过观测而是通过理论计算而得到的。我们还能运用各种理论，如牛顿三大定律、麦克斯韦方程组、相对论等去认识世界和改造世界。

（3）第三种范式：计算与仿真科学

随着世界上第一台通用计算机 ENIAC 在美国宾夕法尼亚大学的诞生，人类社会开始步入

计算机时代，科学研究也进入了一个以“计算”为中心的全新时期。理论的逐步完善使人类仅仅通过计算和仿真就能发现和认识新的规律，目前在材料科学研究中物质大量的特性正是利用“第一性原理”，通过软件的仿真来完成的，在全面禁止核爆条款下，原子弹的研究也完全依赖计算模拟核爆炸来进行。人类认识世界的方法就这样走过了实验科学、理论科学和计算科学三大阶段。

（4）第四种范式：数据密集型科学

网络技术和计算机技术的发展使人类在近期获得了一种新的认识世界的手段，就是利用大量数据来发现新的规律，这种认识世界的方法被称为“第四范式”，是美国著名的科学家图灵奖得主吉姆·格雷在 2007 年提出的。这标志着数据正式成为大家公认的认识世界的方法。大数据出现后人类认识世界的方法就达到 4 种：实验、理论、计算和数据，如图 1-3 所示。现在人类在一年内所产生的数据可能已经超过人类过去几千年产生的数据的总和，即使是复杂度为 $O(n)$的数据处理方法，在面对庞大的 n 时也显得力不从心，人类逐步进入大数据的时代。在大数据环境下，一切都将以数据为中心，从数据中发现问题、解决问题，真正体现出数据的宝贵价值。第四范式的出现正说明了可以利用海量数据加上高速计算发现新知识，计算和数据的关系在大数据时代将会变得更加紧密。

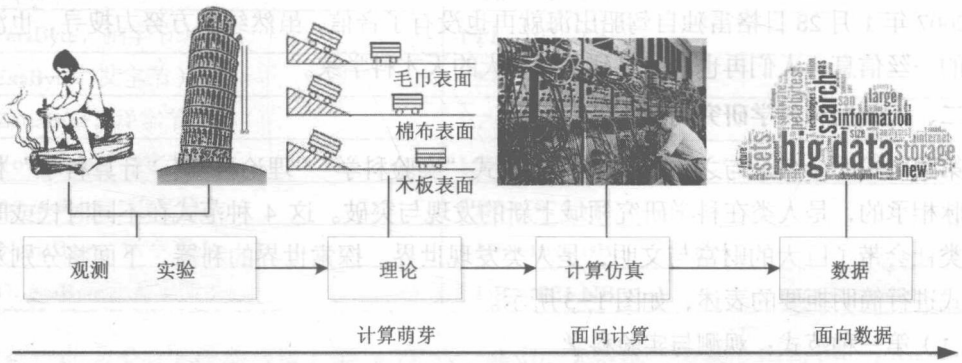


图 1-3　人类认识世界的四种手段

三、大数据对社会发展的影响

大数据的发展不仅改变了科学思维，也必然会引起企业、政府及个人的思维方式的变革，维克托·迈尔·舍恩伯格在《大数据时代：生活、工作与思维的大变革》一书中指出，对于大数据时代，应放弃对因果关心的渴求，而更关注相关关系，正如其在福布斯·静安南京路论坛上的演讲所述：“在大数据时代，人们每天醒来，要想的事情就是面对如此庞大复杂的数据可以用来做什么，其价值可以体现在哪些方面，是否可以找到一个别人从未涉及的事情使得思路及想法成为重要的资产。”由此可见，大数据时代必然会引起思维的转变，而且思维的转变越快，越能在如今竞争激烈的社会中抢占先机。

（1）大数据改变科学研究的思维方式

① 要全体不要抽样。在以往的科学分析中，由于数据存储和处理能力的限制，通常采用抽样的方法，即从全体数据集中抽取一部分样本数据，通过对样本数据的分析，来推断全体数据集的总体特征。通常，样本数据规模要比全集数据小得多，因此，可以在可控的代价内实现数据分析的目的。在大数据时代，其核心技术就是对海量数据进行处理和存储，分布式

文件系统和分布式数据库技术提供了理论上近乎无限的数据存储能力，分布式并行编程框架MapReduce 提供了强大的海量数据并行处理能力。因此，有了大数据技术的支持，科学分析完全可以直接针对全集数据并可以在短时间内快速得到分析结果，例如，Google 公司的 Dremel 可以在 2～3 秒完成 PB 级的数据查询，其速度之快，超乎想象。

② 要效率不要绝对精确。抽样分析方法是科学研究人员常用的一种科学实验分析方法，一般来说，把采集到的数据进行抽样，并以精确性的分析方法分析样本数据，其样本分析结果通常来说较为精准，但是如将其分析结果应用到全体数据集后，微小误差也将会被放大许多，这就意味着抽样分析的微小误差，被放大到全体数据集后，其误差也有可能会随之放大很多。正是由于这个原因，传统的数据分析方法往往更加注重提高算法的精确性，其次才是提高算法效率。现在，大数据时代采用的是全体数据集分析而非抽样数据分析，其分析结果就不存在误差被放大的问题，因此，算法的高精确性已经不是现在所要追求的首要目标，相反，大数据时代具有“秒级响应”的特征，要求在几秒内就迅速给出针对海量数据的实时分析结果，否则，就会丧失数据的价值，因此，数据分析的效率将会是大数据时代关注的焦点。

③ 要相关不要因果。数据分析的目的主要体现在如下两方面：一方面是解释事物背后的发展规律或机理，例如，找出某地区大型企业某类产品销售业绩大幅下滑的原因，并分析问题所在；另一方面是用于预测未来可能发生的事件，比如，通过实时分析各大媒体最新报道及微博等数据，当发现人们对雾霾的讨论明显增加时，就可建议销售部增加口罩的进货量，因为人们关注雾霾的一个直接结果是在这样的一个环境下如何保障身体损害最小，而简单方便的口罩将会是一个不错的产品。其实，不管是何目的，反映的都是“因果关系”。而在大数据时代，因果关系将不再那么重要，人们转而追求“相关性”而非“因果性”。譬如，当我们在淘宝网购买了一个键盘后，淘宝网还会自动提示你，与你购买相同产品的其他客户还购买了鼠标，也就是说，淘宝网只会告诉你“键盘”和“鼠标”之间存在相关性，但是，它不会告诉你其他客户为什么会在购买键盘之后还会继续购买鼠标。

（2）大数据改变人们的生存方式

在 21 世纪，信息技术突飞猛进的今天，物联网、嵌入式技术、传感技术等的发展，为人类更全面地感知客观存在的物理世界提供了基础；而互联网、云计算等信息技术的发展更是改变了人类通信与管理信息的方式。随着技术的发展及工具的更新换代，人类也提出了更高的生存需求，美国国家科学基金委员会在 2006 年提出了信息物理系统（Cyber Physical Systems，CPS）的概念；2007 年，不同机构及研究学者对其进行了定义，包括 LEE、Baheli、Sastry 及 Krogh 等，强调计算元素及物理元素，实体与虚拟网络的关系，并注重通信计算及控制能力，尽管不同定义的描述不同，但是都明确了 CPS 的内涵：Cyber 与 Physical 的深度融合后形成的智能系统。CPS 的运行状况如图 1-4 所示。

在图 1-4 中，最外面一层是物理实体，其代表我们生活的物理世界；中间一层为感知层，包括传感器等具有采集功能的设备；第三层为计算机等具有计算功能的设备，其负责实现对采集数据的分析及可视化呈现；最里面一层为决策层（具有决策能力的人或者其他事物），其通过感知及分析结果做出决策，并作用于物理实体。CPS 的运行图体现了在感知基础上，人、机、物的深层融合。CPS 的有效工作将改变人类的生存方式，如其可应用于无人

机、自主导航的汽车等以实现物理实体的自主工作，医疗领域中可应用于自动手术，物联网领域中可实现生活中的智能家居及智慧城市等。上述 CPS 的成功实现，最重要的基础就是系统中收集的大量数据的有效分析及处理，其是决策支持的重要来源。即如果没有大数据的积累及分析，那么 CPS 也就无从谈起。由此可见，大数据的产生及有效分析是 CPS 的重要资源和基础，结合其他技术的发展，将为改变人类生存方式提供重要动力。

图 1-4　CPS 的运行状况

（3）大数据改变人类的生产方式

目前已经先后经历了三次工业革命，包括 1760—1840 年因为蒸汽动力的发明产生了生产制造的机械化，开创了“蒸汽时代”；1840—1950 年因为电的发明开创了“电气时代”，使得生产得以批量化；1950 年至今，电子技术和计算机等信息技术的发展开创了“信息时代”，使得产品更为丰富，功能性更强；而随着科技的进一步发展，科技的进步也必定引起生产方式的变革。为此，德国提出了“工业 4.0”，即第四次工业革命，以智能制造为主导实现生产制造人机一体化，“工业 4.0”的提出预示着革命性的生产方式的诞生，而实现“工业 4.0”的基础就是大数据的分析及 CPS 的推广，其标志着生产制造业必须转向以数据分析为中心。由此可见，大数据的发展将在生产方式改变中起到关键作用。四次工业革命演化如图 1-5 所示。

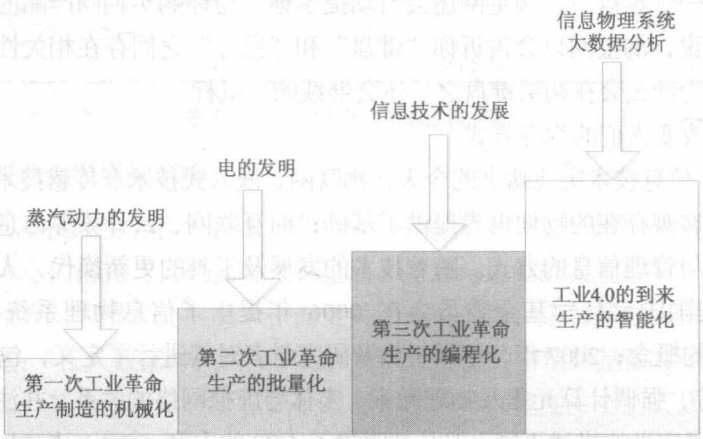


图 1-5　四次工业革命演化

“工业 4.0”是在 2013 年举办的汉诺威（Hannover）工业博览会上正式提出的，虽然第四次工业革命是否到来还存在很大争议，但是目前很多国家已经投入了大量资金及精力来推进“工业 4.0”的进程。成功的典范是特斯拉及西门子，特斯拉将自己的核心定位于大型可移动的智能终端，通过互联网将汽车设计为包含软件、硬件，以及内容和服务的体验工具，将互联网思维引入制造业；西门子的电子车间更是将“工业 4.0”付诸实践的典型代表，其建立了