21世纪高等院校
云计算和大数据人才培养规划教材

大数据导论

思维、技术与应用

武志学 编著

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

INTRODUCTION TO BIG DATA

Thinking, Technology and Application

大数据导论

思维、技术与应用

武志学 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据导论 : 思维、技术与应用 / 武志学编著. -- 北京 : 人民邮电出版社, 2019.4
21世纪高等院校云计算和大数据人才培养规划教材
ISBN 978-7-115-50485-2

Ⅰ. ①大… Ⅱ. ①武… 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(2018)第289518号

内 容 提 要

本书将基本概念与实例相结合，由浅入深、循序渐进地对大数据思维、技术和应用做了全面系统的介绍。全书共 12 章，分为大数据基础篇、大数据存储篇、大数据处理篇、大数据挖掘篇和大数据应用篇。

大数据基础篇的内容涵盖了大数据思维理念、大数据的产生与作用、大数据基本概念、大数据采集工具 Flume 和 Scribe、大数据爬虫工具 Nutch 和 Scapy、大数据预处理工具 Kettle、大数据处理架构 Hadoop；大数据存储篇的内容包含分布式文件存储系统 HDFS、海量数据存储数据库系统 HBase 和海量数据仓库系统 Hive；大数据处理篇主要介绍了分布式并发计算批处理模式 MapReduce，基于内存的快速处理模式 Spark，以及基于实时数据流的实时处理模式 Spark Streaming；大数据挖掘篇主要对分类、预测、聚类和关联等各类大数据挖掘算法的原理和使用场景进行了描述，并使用 Spark MLlib 提供的机器学习算法进行了实例讲解；大数据应用篇分别从大数据场景应用的横向和纵向出发，介绍了大数据在各个功能领域的应用场景和在各个行业的应用场景。

本书可作为高校大数据相关专业和其他专业的大数据导论课程的教材，每个知识点都配有与理论学习内容相结合的案例介绍和代码实例，并在每章后面配有丰富的作业。本书也可以作为广大 IT 从业人员系统了解大数据技术和应用的参考书。

◆ 编　　著　武志学
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
北京鑫正大印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：15.75　　2019 年 4 月第 1 版
字数：414 千字　　2019 年 4 月北京第 1 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

前　言

随着信息技术的快速发展，海量的数据已经成为企业最具价值的财富。移动互联网和物联网技术使得信息传播极其迅速，大数据开始蔓延到社会的各行各业，从而影响着人们的学习、工作、生活，以及社会的发展。大数据技术的应用场景也越来越广泛，从市场营销到产品设计，从市场预测到决策支持，从效能提升到运营管理，并且大数据的应用已经从早期的互联网公司开始走向传统企业。

目前，大数据领域正面临全球性的“人才荒”，根据麦肯锡报告显示，2018 年，美国市场的大数据人才和高级分析专家的人才缺口将高达 19 万。此外，美国企业还需要 150 万位能够提出正确问题、运用大数据分析结果的大数据相关管理人才。目前，国内的大数据人才仅 46 万，未来 3 到 5 年内大数据人才的缺口将高达 156 万，并且随着时间的推移，人才缺口还会逐渐放大，在很长时间内企业将面临大数据人才严重紧缺状态。

为了满足社会需求，加快大数据人才的培养，2016 年教育部先后设置“数据科学与大数据技术”本科专业和“大数据技术与应用”高职专业。截至 2018 年，我国已有 283 所本科院校获批大数据相关的本科专业，212 所高职院校获批大数据相关的高职专业。尽管各方都意识到了大数据人才培养的重要性，但是到底如何培养好的大数据人才，还是个亟待解决的问题。作者基于 2012 年至今在高校从事大数据人才培养的实际经验，结合多年从事大数据技术研究和大数据应用开发的实践体会，编写了这本教材。本书围绕着大数据技术和应用问题，由浅入深、循序渐进，以基本概念与实例相结合的方法，对大数据思维、技术和应用做了系统的介绍，包括大数据获取、大数据预处理、大数据存储和管理、大数据批处理、大数据在线处理、大数据挖掘和大数据应用等各个技术环节。

本书不仅可以作为高校大数据相关专业和其他专业的大数据导论课程的教材，也可以作为广大 IT 从业人员系统了解大数据技术和应用的参考书。作者力图使读者通过学习，能够基本理解各类大数据技术，能够初步使用大数据思维分析问题，能够掌握大数据技术解决实际问题的基本原理，并能够了解大数据技术在各个行业的应用场景。

作为高校的教材，本书在每一个环节都配有与理论学习内容相结合的案例介绍，还有使用 Java 和 Python 语言编写的应用实例，使读者能够在大数据平台上通过实践亲身体验大数据处理和分析的过程，从而加快和加深对大数据理论和技术的理解。为了使读者方便检验和复习巩固学习到的知识，本书每章后面都配有丰富的作业。

全书内容主要分为 5 部分，共 12 章。

第一部分是大数据基础篇（第 1～5 章），对大数据思维、大数据技术、大数据平台和大数据应用进行了基本介绍。第 1 章主要阐述了大数据的产生与作用及大数据思维；第 2 章对大数据技术和大数据应用进行了基本介绍；第 3 章介绍了大数据的采集方法，相应的日志采集系统 Flume 和 Scribe，以及网络爬虫工具 Nutch 和 Scapy；第 4 章介绍了大数据预处理技术，以及数据预处理工具 Kettle；第 5 章对大数据的技术基础进行了描述，并对著名的 Google 和 Hadoop 大数据处理系统进行了介绍。

第二部分是大数据存储篇（第 6～7 章），主要介绍了大数据存储和管理技术，分别讲解了分布式文件存储系统 HDFS，以及支持大规模、半结构化海量数据存储的数据库系

统 HBase。

第三部分是大数据处理篇（第 8～10 章），主要介绍了大数据处理技术，分别为分布式并发计算批处理模式 MapReduce，基于内存的快速处理模式 Spark，以及基于实时数据流的实时处理模式 Spark Streaming。

第四部分是大数据挖掘篇（第 11 章），主要对分类、预测、聚类和关联等各类大数据挖掘算法的原理和使用场景进行了描述，并使用 Spark MLlib 提供的机器学习算法进行了实例讲解。

第五部分是大数据应用篇（第 12 章），首先，从大数据应用场景横向角度出发，介绍了大数据在各个功能领域的应用场景，包括精准营销、个性化推荐和大数据预测；然后，从大数据应用场景纵向角度出发，介绍了各个行业的大数据应用场景，包括银行、证券、保险、互联网、电信和物流等行业。

采用本书作为教材时，授课教师可以参考下述教学安排。

章	理论课	实验课	合计
第 1 章　大数据思维	2 学时		2 学时
第 2 章　大数据技术概述	2 学时		2 学时
第 3 章　大数据采集	2 学时	2 学时	4 学时
第 4 章　大数据预处理	4 学时	4 学时	8 学时
第 5 章　大数据处理系统	4 学时	2 学时	6 学时
第 6 章　大数据文件系统 HDFS	4 学时	4 学时	8 学时
第 7 章　NoSQL 数据库 HBase	4 学时	4 学时	8 学时
第 8 章　大数据批处理 Hadoop MapReduce	4 学时	4 学时	8 学时
第 9 章　大数据快速处理 Spark	4 学时	4 学时	8 学时
第 10 章　大数据实时流计算 Spark Streaming	4 学时	4 学时	8 学时
第 11 章　大数据挖掘	12 学时	12 学时	24 学时
第 12 章　大数据应用	4 学时		4 学时
合计	50 学时	40 学时	90 学时

本书的编写得到了五舟汉云公司研发人员和电子科技大学成都学院教师们的大力支持和帮助，在这里表示感谢。特别感谢赵阳老师在本书编写过程中提出的宝贵建议，感谢五舟汉云大数据小组成员汪雪飞、龚晓宇和杨棋等对书中实例的验证，感谢五舟汉云教育小组成员邓依洁、屈太源、吕姗姗、杨莹和杨燕等为本书提供的校对和制图工作。

编者

2018 年 6 月

目 录 CONTENTS

第一部分 大数据基础篇

第 1 章 大数据思维 2

第 2 章 大数据技术概述 15

第 3 章 大数据采集 22

第 4 章　大数据预处理　37

第 5 章　大数据处理系统　57

第二部分　大数据存储篇

第 6 章　大数据文件系统 HDFS　66

第 7 章　NoSQL 数据库 HBase　80

第三部分 大数据处理篇

第 8 章 大数据批处理 Hadoop MapReduce 103

第 9 章 大数据快速处理 Spark 124

第 10 章 大数据实时流计算 Spark Streaming 143

第一部分

大数据基础篇

Chapter 1 第 1 章 大数据思维

大数据开启了一次重大的时代转型。大数据技术在短短的数年之内，从少数科学家的主张，转变为全球领军公司的战略实践，继而上升为大国的竞争战略，形成一股无法忽视、无法回避的历史潮流。互联网、物联网、云计算、智慧城市正在使数据沿着“摩尔定律”飞速增长，一个与物理空间平行的数字空间正在形成。在新的数字世界当中，数据成为最宝贵的生产要素，顺应趋势、积极谋变的国家和企业将乘势崛起，成为新的领军者；无动于衷、墨守成规的组织将逐渐被边缘化，失去竞争的活力和动力。毫无疑问，大数据正在开启一个崭新时代。

1.1 什么是大数据

大数据本身是一个抽象的概念。从一般意义上讲，大数据是指无法在有限时间内用常规软件工具对其进行获取、存储、管理和处理的数据集合。目前，业界对大数据还没有一个统一的定义，但是大家普遍认为，大数据具备 Volume、Velocity、Variety 和 Value 四个特征，简称“4V”，即数据体量巨大、数据速度快、数据类型繁多和数据价值密度低，如图 1-1 所示。下面分别对每个特征作简要描述。

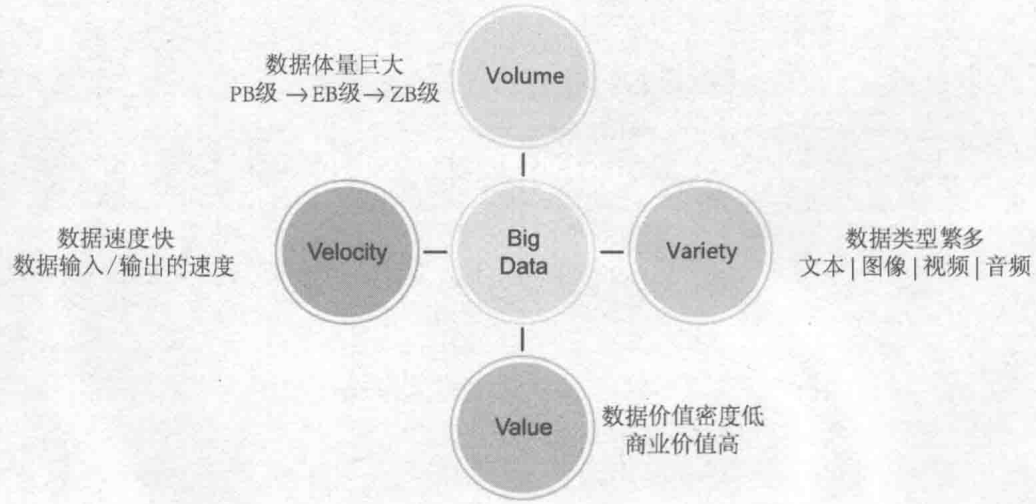

图 1-1 大数据特征

（1）Volume：表示大数据的数据体量巨大。数据集合的规模不断扩大，已经从 GB 级增加到 TB 级再增加到 PB 级，近年来，数据量甚至开始以 EB 和 ZB 来计数。例如，一个中型城市的视频监控信息一天就能达到几十 TB 的数据量。百度首页导航每天需要提供的数据超过

1.5PB，如果将这些数据打印出来，会超过 5 000 亿张 A4 纸。图 1-2 展示了每分钟互联网产生的各类数据的量。

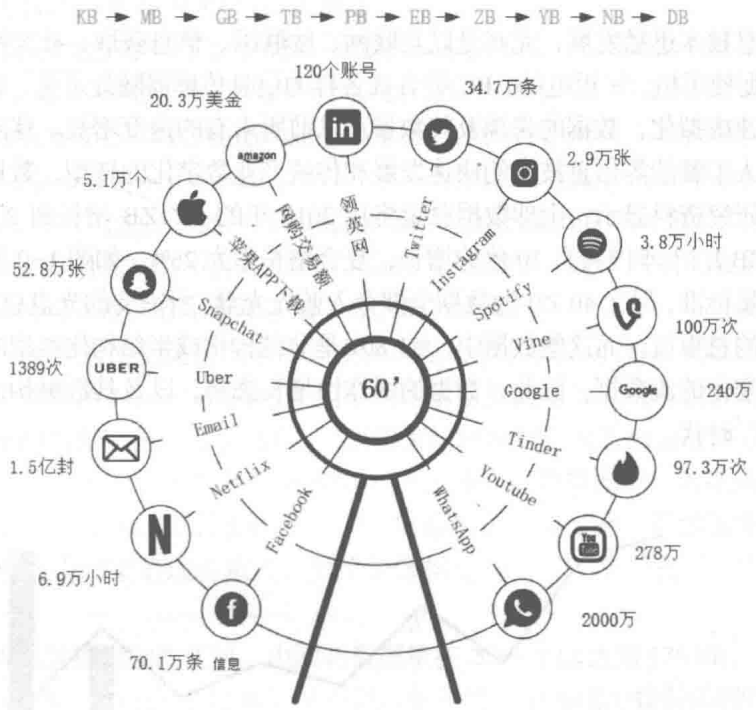

图 1-2 互联网每分钟产生的数据

（2）Velocity：表示大数据的数据产生、处理和分析的速度在持续加快。加速的原因是数据创建的实时性特点，以及将流数据结合到业务流程和决策过程中的需求。数据处理速度快，处理模式已经开始从批处理转向流处理。业界对大数据的处理能力有一个称谓——“1 秒定律”，也就是说，可以从各种类型的数据中快速获得高价值的信息。大数据的快速处理能力充分体现出它与传统的数据处理技术的本质区别。

（3）Variety：表示大数据的数据类型繁多。传统 IT 产业产生和处理的数据类型较为单一，大部分是结构化数据。随着传感器、智能设备、社交网络、物联网、移动计算、在线广告等新的渠道和技术不断涌现，产生的数据类型无以计数。现在的数据类型不再只是格式化数据，更多的是半结构化或者非结构化数据，如 XML、邮件、博客、即时消息、视频、照片、点击流、日志文件等。企业需要整合、存储和分析来自复杂的传统和非传统信息源的数据，包括企业内部和外部的数据。

（4）Value：表示大数据的数据价值密度低。大数据由于体量不断加大，单位数据的价值密度在不断降低，然而数据的整体价值在提高。以监控视频为例，在一小时的视频中，有用的数据可能仅仅只有一两秒，但是却会非常重要。现在许多专家已经将大数据等同于黄金和石油，这表示大数据当中蕴含了无限的商业价值。

根据中商产业研究院发布的《2018—2023 年中国大数据产业市场前景及投资机会研究报告》显示，2017 年中国大数据产业规模达到 4 700 亿元，同比增长 30%。随着大数据在各行业的融合应用不断深化，预计 2018 年中国大数据市场产值将突破 6 000 亿元，达到 6 200 亿元。

通过对大数据进行处理，找出其中潜在的商业价值，将会产生巨大的商业利润。

1.2 从 IT 时代到大数据时代

近年来，信息技术迅猛发展，尤其是以互联网、物联网、信息获取、社交网络等为代表的技术日新月异，促使手机、平板电脑、PC 等各式各样的信息传感器随处可见，虚拟网络快速发展，现实世界快速虚拟化，数据的来源及其数量正以前所未有的速度增长。伴随着云计算、大数据、物联网、人工智能等信息技术的快速发展和传统产业数字化的转型，数据量呈现几何级增长，根据市场研究资料显示，全球数据总量将从 2016 年的 16.1ZB 增长到 2025 年的 163ZB（约合 180 万亿 GB），十年内将有 10 倍的增长，复合增长率为 26%，如图 1-3 所示。若以现有的蓝光光盘为计量标准，那么 40 ZB 的数据全部存入蓝光光盘，所需要的光盘总重量将达到 424 艘尼米兹号航母的总重量。而这些数据中，约 80%是非结构化或半结构化类型的数据，甚至更有一部分是不断变化的流数据。因此，数据的爆炸性增长态势，以及其数据构成特点使得人们进入了“大数据”时代。

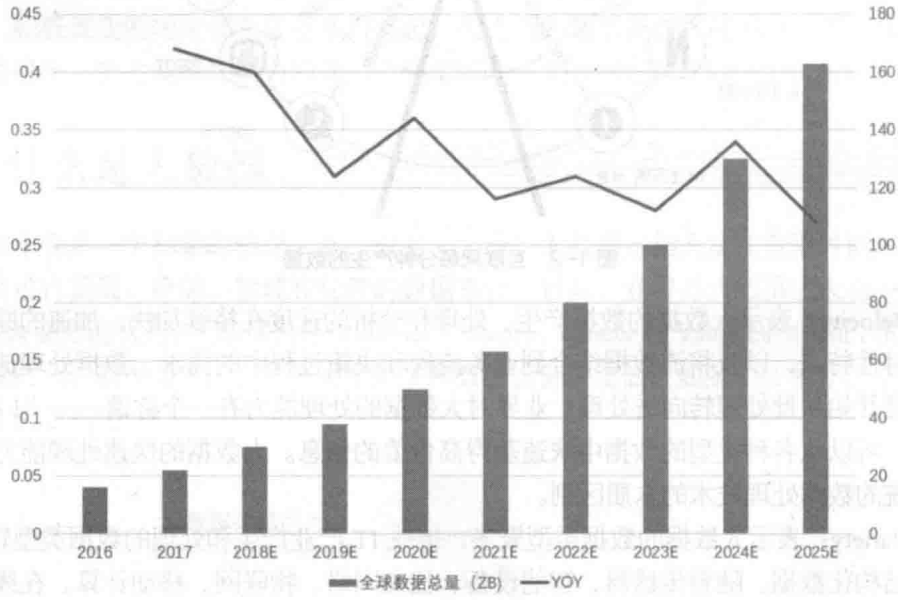

图 1-3 2016—2025 年全球数据产量及预测

如今，大数据已经被赋予多重战略含义。在资源的角度，数据被视为“未来的石油”，被作为战略性资产进行管理；在国家治理角度，大数据被用来提升治理效率，重构治理模式，破解治理难题，它将掀起一场国家治理革命；在经济增长角度，大数据是全球经济低迷环境下的产业亮点，是战略新兴产业的最活跃部分；在国家安全角度，全球数据空间没有国界边疆，大数据能力成为大国之间博弈和较量的利器。总之，国家竞争焦点将从资本、土地、人口、资源转向数据空间，全球竞争版图将分成新的两大阵营：数据强国与数据弱国。

从宏观上看，由于大数据革命的系统性影响和深远意义，主要大国快速做出战略响应，将大数据置于非常核心的位置，推出国家级创新战略计划。美国 2012 年发布了《大数据研究和发展计划》，并成立“大数据高级指导小组”，2013 年又推出“数据—知识—行动”计划，2014 年进一步发布《大数据：把握机遇，维护价值》政策报告，启动“公开数据行动”，陆续公开 50 个门类的政府数据，鼓励商业部门进行开发和创新。欧盟正在力推《数据价值链战略计划》；英

国发布了《英国数据能力发展战略规划》；日本发布了《创建最尖端 IT 国家宣言》；韩国提出了“大数据中心战略”。中国多个省市发布了大数据发展战略，国家层面的《关于促进大数据发展的行动纲要》也于 2015 年 8 月 19 日正式通过。

从微观上看，大数据重塑了企业的发展战略和转型方向。美国的企业以 GE 提出的“工业互联网”为代表，提出智能机器、智能生产系统、智能决策系统，将逐渐取代原有的生产体系，构成一个“以数据为核心”的智能化产业生态系统。德国的企业以“工业 4.0”为代表，要通过信息物理系统（Cyber Physical System，CPS）把一切机器、物品、人、服务、建筑统统连接起来，形成一个高度整合的生产系统。中国的企业以阿里巴巴提出的“DT 时代”（Data Technology）为代表，认为未来驱动发展的不再是石油、钢铁，而是数据。这 3 种新的发展理念可谓异曲同工、如出一辙，共同宣告“数据驱动发展”成为时代主题。

与此同时，大数据也是促进国家治理变革的基础性力量。正如《大数据时代》的作者舍恩伯格在定义中所强调的：“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。”在国家治理领域，大数据为解决以往的“顽疾”和“痛点”，提供了强大支撑，如建设阳光政府、责任政府、智慧政府；大数据使以往无法实现的环节变得简单、可操作，如精准医疗、个性化教育、社会监管、舆情监测预警；大数据也使一些新的主题成为国家治理的重点，如维护数据主权、开放数据资产、保持在数字空间的国家竞争力等。

中国具备成为数据强国的优势。中国的数据量在 2013 年已达到 576 EB，到 2020 年这个数字将会达到 8.06 ZB，增长超过 12 倍。从全球占比来看，中国成为数据强国的潜力极为突出，2010 年中国数据占全球数据的比例为 10%，2013 年占比为 13%，2020 年占比将达到 18%，如图 1-4 所示。届时，中国的数据规模将超过美国位居世界第一。中国成为数据大国并不奇怪，因为中国是人口大国、制造业大国、互联网大国、物联网大国，这都是最活跃的数据生产主体，未来几年，中国成为数据大国也是逻辑上的必然结果。

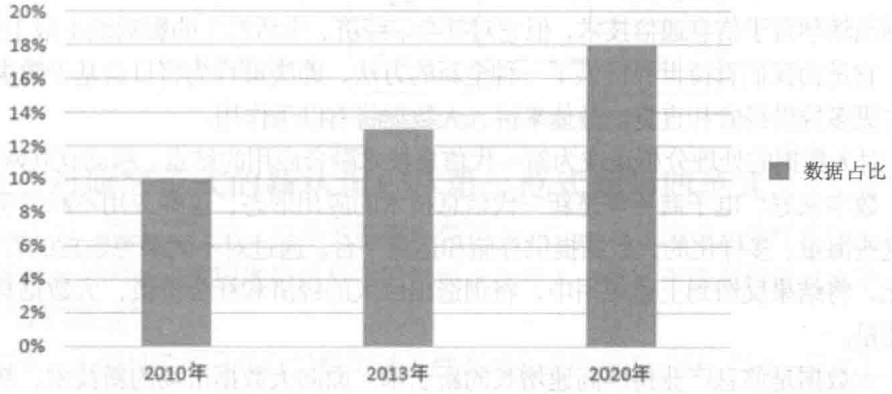

图 1-4　2010—2020 年中国数据的全球占比

1.3　大数据的产生与作用

大数据是信息通信技术发展积累至今，按照自身技术发展逻辑，从提高生产效率向更高级智能阶段的自然生长。无处不在的信息感知和采集终端为我们采集了海量的数据，而以云计算为代表的计算技术的不断进步，为我们提供了强大的计算能力。

1.3.1 大数据的产生

从采用数据库作为数据管理的主要方式开始，人类社会的数据产生方式大致经历了3个阶段，而正是数据产生方式的巨大变化才最终导致大数据的产生。

（1）运营式系统阶段。数据库的出现使得数据管理的复杂度大大降低，在实际使用中，数据库大多为运营系统所采用，作为运营系统的数据管理子系统，如超市的销售记录系统、银行的交易记录系统、医院病人的医疗记录等。人类社会数据量的第一次大的飞跃正是在运营式系统开始广泛使用数据库时开始的。这个阶段的最主要特点是，数据的产生往往伴随着一定的运营活动；而且数据是记录在数据库中的，例如，商店每售出一件产品就会在数据库中产生一条相应的销售记录。这种数据的产生方式是被动的。

（2）用户原创内容阶段。互联网的诞生促使人类社会数据量出现第二次大的飞跃，但是真正的数据爆发产生于Web 2.0时代，而Web 2.0的最重要标志就是用户原创内容。这类数据近几年一直呈现爆炸性的增长，主要有两个方面的原因。一是以博客、微博和微信为代表的新型社交网络的出现和快速发展，使得用户产生数据的意愿更加强烈。二是以智能手机、平板电脑为代表的新型移动设备的出现，这些易携带、全天候接入网络的移动设备使得人们在网上发表自己意见的途径更为便捷。这个阶段的数据产生方式是主动的。

（3）感知式系统阶段。人类社会数据量第三次大的飞跃最终导致了大数据的产生，今天我们正处于这个阶段。这次飞跃的根本原因在于感知式系统的广泛使用。随着技术的发展，人们已经有能力制造极其微小的带有处理功能的传感器，并开始将这些设备广泛地布置于社会的各个角落，通过这些设备来对整个社会的运转进行监控。这些设备会源源不断地产生新数据，这种数据的产生方式是自动的。

简单来说，数据产生经历了被动、主动和自动三个阶段。这些被动、主动和自动的数据共同构成了大数据的数据来源，但其中自动式的数据才是大数据产生的最根本原因。

1.3.2 大数据的作用

大数据虽然孕育于信息通信技术，但它对社会、经济、生活产生的影响绝不限于技术层面。更本质上，它是为我们看待世界提供了一种全新的方法，即决策行为将日益基于数据分析，而不是像过去更多凭借经验和直觉。具体来讲，大数据将有以下作用。

第一，对大数据的处理分析正成为新一代信息技术融合应用的结点。移动互联网、物联网、社交网络、数字家庭、电子商务等是新一代信息技术的应用形态，这些应用不断产生大数据。云计算为这些海量、多样化的大数据提供存储和运算平台。通过对不同来源数据的管理、处理、分析与优化，将结果反馈到上述应用中，将创造出巨大的经济和社会价值，大数据具有催生社会变革的能量。

第二，大数据是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，还将催生出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技术、数据挖掘技术和软件产品的发展。

第三，大数据利用将成为提高核心竞争力的关键因素。各行各业的决策正在从“业务驱动”向“数据驱动”转变。在商业领域，对大数据的分析可以使零售商实时掌握市场动态并迅速做出应对，可以为商家制定更加精准有效的营销策略提供决策支持，可以帮助企业为消费者提供更加及时和个性化的服务；在医疗领域，可提高诊断准确性和药物有效性；在公共事业领域，

大数据也开始发挥促进经济发展、维护社会稳定等方面的重要作用。

第四，大数据时代，科学研究的方法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，研究人员可通过实时监测、跟踪研究对象在互联网上产生的海量行为数据，进行挖掘分析，揭示出规律性的东西，提出研究结论和对策。

1.4 大数据时代的新理念

大数据时代的到来改变了人们的生活方式、思维模式和研究范式，我们可以总结出 10 个重大变化，如图 1-5 所示。

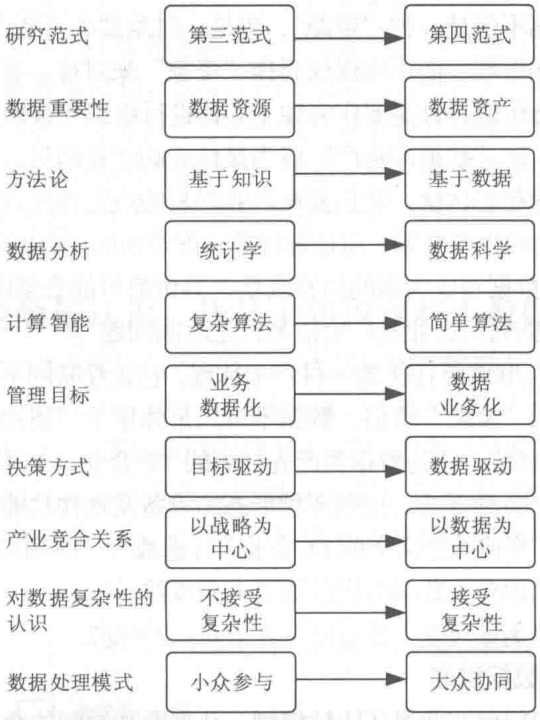

图 1-5 大数据时代的 10 个重大变化

1.4.1 对研究范式的新认识：从第三范式到第四范式

2007 年 1 月，图灵奖得主、关系型数据库鼻祖 Jim Gray 发表演讲，他凭着自己对于人类科学发展特征的深刻洞察，敏锐地指出科学的发展正在进入“数据密集型科学发现范式”——科学史上的“第四范式”。

在他看来，人类科学研究活动已经历过三种不同范式的演变过程。“第一范式”是指原始社会的“实验科学范式”。18 世纪以前的科学进步均属于此列，其核心特征是对有限的客观对象进行观察、总结、提炼，用归纳法找出其中的科学规律，如伽利略提出的物理学定律。“第二范式”是指 19 世纪以来的理论科学阶段，以模型和归纳为特征的“理论科学范式”。其核心特征是以演绎法为主，凭借科学家的智慧构建理论大厦，如爱因斯坦提出的相对论、麦克斯方程组、量子理论和概率论等。“第三范式”是指 20 世纪中期以来的计算科学阶段的“计算科学范式”。面对大量过于复杂的现象，归纳法和演绎法都难以满足科学研究的需求，人类开始借助计算机的高级运算能力对复杂现象进行建模和预测，如天气、地震、核试验、原子的运动等。

然而，随着近年来人类采集数据量的爆炸性增长，传统的计算科学范式已经越来越无力驾驭海量的科研数据了。例如，欧洲的大型粒子对撞机、天文领域的 Pan-STARRS 望远镜每天产生的数据多达几千万亿字节（PB）。很明显，这些数据已经突破了“第三范式”的处理极限，无法被科学家有效利用。

正因为如此，目前正在从“计算科学范式”转向“数据密集型科学发现范式”。“第四范式”的主要特点是科学研究人员只需要从大数据中查找和挖掘所需要的信息和知识，无须直接面对所研究的物理对象。例如，在大数据时代，天文学家的研究方式发生了新的变化，其主要研究任务变为从海量数据库中发现所需的物体或现象的照片，而不再需要亲自进行太空拍照。

1.4.2 对数据重要性的新认识：从数据资源到数据资产

在大数据时代，数据不仅是一种“资源”，更是一种重要的“资产”。因此，数据科学应把数据当作一种“资产”来管理，而不能仅仅当作“资源”来对待。也就是说，与其他类型的资产相似，数据也具有财务价值，且需要作为独立实体进行组织与管理。

大数据时代的到来，让“数据即资产”成为最核心的产业趋势。在这个“数据为王”的时代，回首信息产业发展的起起伏伏，我们发现产业兴衰的决定性因素，已不是土地、人力、技术、资本这些传统意义上的生产要素，而是曾经被一度忽视的“数据资产”。世界经济论坛报告曾经预测称，“未来的大数据将成为新的财富高地，其价值可能会堪比石油”，而大数据之父维克托也乐观地表示，“数据列入企业资产负债表只是时间问题”。

“数据成为资产”是互联网泛在化的一种资本体现，它让互联网不仅具有应用和服务本身的价值，而且具有了内在的“金融”价值。数据不再只是体现于“使用价值”方面的产品，而成为实实在在的“价值”。目前，作为数据资产先行者的 IT 企业，如苹果、谷歌、IBM、阿里、腾讯、百度等，无不想尽各种方式，挖掘多种形态的设备及软件功能，收集各种类型的数据，发挥大数据的商业价值，将传统意义上的 IT 企业，打造成为“终端+应用+平台+数据”四位一体的泛互联网化企业，以期在大数据时代获取更大的收益。

大数据资产的价值的衡量尺度主要有以下 3 个方面的标准。

1．独立拥有及控制数据资产

目前，数据的所有权问题在业界还比较模糊。从拥有和控制的角度来看，数据可以分为Ⅰ型数据、Ⅱ型数据和Ⅲ型数据。

Ⅰ型数据主要是指数据的生产者自己生产出来的各种数据，例如，百度对使用其搜索引擎的用户的各种行为进行收集、整理和分析，这类数据虽然由用户产生，但产权却属于生产者，并最大限度地发挥其商业价值。

Ⅱ型数据又称为入口数据，例如，各种电子商务营销公司通过将自身的工具或插件植入电商平台，来为其提供统计分析服务，并从中获取各类经营数据。虽然这些数据的所有权并不属于这些公司，在使用时也有一些规则限制，但是它们却有着对数据实际的控制权。

相比于前两类数据，Ⅲ型数据的产权情况比较复杂，它们主要依靠网络爬虫，甚至是黑客手段获取数据。与Ⅰ型和Ⅱ型数据不同的是，这些公司流出的内部数据放在网上供人付费下载。这种数据在当前阶段，还不能和资产完全画等号。

2．计量规则与货币资本类似

大数据要实现真正的资产化，用货币对海量数据进行计量是一个大问题。尽管很多企业都意识到数据作为资产的可能性，但除了极少数专门以数据交易为主营业务的公司外，大多数公