“十三五”高等学校规划教材

大数据导论

DASHUJU DAOLUN

周鸣争　陶　皖　主编
杨　丹　李臣龙　万家山　参编

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

“十三五”高等学校规划教材

大数据导论

周鸣争　陶　皖　主编
杨　丹　李臣龙　万家山　参编

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书全面阐述了大数据的内涵与特征、体系架构以及所涉及关键技术。全书共分7章，内容包括大数据概论、大数据存储、大数据处理、大数据分析、大数据可视化、大数据应用和大数据发展趋势与展望，每章内容都与主流技术和典型案例紧密结合，以便读者对大数据及其关键技术有更好的了解和掌握。

本书适合作为高等院校数据科学与大数据技术、计算机、软件工程、电子信息等相关专业以及创新创业或素质教育的大数据课程教材，也可作为其他读者深入了解大数据技术的参考用书。

图书在版编目（CIP）数据

大数据导论 / 周鸣争，陶皖主编 . —北京：中国铁道出版社，2018.3

“十三五”高等学校规划教材

ISBN 978-7-113-24263-3

Ⅰ.①大… Ⅱ.①周… ②陶… 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020032号

书　　名：大数据导论

作　　者：周鸣争　陶　皖　主编

策　　划：翟玉峰　　**读者热线**：（010）63550836

责任编辑：翟玉峰　包　宁

封面设计：刘　颖

责任校对：张玉华

责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街8号）

网　　址：http://www.tdpress.com/51eds/

印　　刷：三河市航远印刷有限公司

版　　次：2018年3月第1版　2018年3月第1次印刷

开　　本：787 mm×1 092 mm　1/16　**印张**：11.5　**字数**：249千

印　　数：1～2 000册

书　　号：ISBN 978-7-113-24263-3

定　　价：32.00元

Preface
前 言

如今大数据已被提升为国家战略并写进政府工作报告，相信该战略的实施将对各行各业产生深远的影响，同时会触发社会思维的新变革。大数据技术的应用战略落地生根，除政府政策支持外更需要大量的人才资源作为后盾。面对新需求，高等院校作为人才培养主阵地，将义不容辞地为普及相关技术知识作出应有的贡献，本书正是出于此目的而编写。

本书在结构安排与内容撰写时遵循教学规律，考虑读者对象特点，紧紧围绕“大数据技术”这一中心，以浅显易懂的语言详细介绍了大数据的基本特征、体系结构、相关技术及其应用领域，做到由浅入深、环环紧扣。章节中结合案例与典型应用帮助读者增强对大数据技术的感性认识，了解大数据对未来学习、生活、工作与社会发展的重要性，理解构建大数据应用系统所需的技术、方法。

本书共分为7章，较全面地阐述分析了大数据的内涵、体系结构以及所涉及的相关支撑技术。第1章主要分析大数据提出的背景及内涵特征，并基于大数据的来源分析了大数据系统结构与主要相关技术；第2章介绍了大数据存储相关技术的概念与原理，包括传统大数据存储系统的3种架构、分布式文件系统（HDFS）、NoSQL数据库、分布式数据库（HBase）以及NewSQL数据库技术；第3章主要介绍了目前大数据处理主流技术和平台以及Hadoop MapReduce并行处理和编程技术；第4章主要介绍了大数据分析的特点、类型、流程及大数据分析的各种方法和主要应用领域；第5章主要阐述了大数据可视化技术的基本概念、可视化流程、可视化编码、可视化设计以及大数据可视化的一些软件和工具；第6章主要介绍了大数据在互联网行业、医疗、交通、自动问答等领域的具体应用；第7章主要介绍了大数据技术在安全与隐私保护、数据共享和数

据科学等方面存在的挑战与发展趋势。

本书由周鸣争、陶皖主编，杨丹、李臣龙、万家山参与编写。具体编写分工如下：周鸣争编写第1章，李臣龙编写第2、3章，陶皖编写第4、6章，杨丹编写第5章，万家山编写第7章。周鸣争、陶皖负责全书的统稿及定编工作。

由于编者水平有限，书中疏漏与不足之处在所难免，望读者提出意见和建议。

编　者

2017年11月

Contents

目 录

第1章 概 论

大数据作为继云计算、物联网之后IT领域又一次颠覆性技术，备受人们的关注。大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。对人类的社会生产和生活必将产生重大而深远的影响。

本章重点介绍大数据产生的背景、基本概念、关键技术、处理流程与发展机遇。

1.1 什么是大数据

1.1.1 大数据产生的背景

随着以博客、社交网络、基于位置服务（Location Based Service，LBS）为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，数据正以前所未有的速度在不断地增长和累积，大数据时代已经来到。

根据国际数据公司（IDC）做出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（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，预计到2020年，全球将拥有35 ZB的数据量，相较于2010年，数据量将增长近30倍。物联网、云计算、移动互联网、车联网、手机、平板电脑、PC以及遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式。 学术界、工业界以及政府机构都已经开始密切关注大数据问题，并对其产生浓厚兴趣。*Nature* 早在2008 年就推出了*Big Data*专刊。计算社区联盟（Computing Community Consortium）在2008 年发表了报告*Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society*，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术以及面临的一些挑战。*Science* 在2011 年2 月推出专刊*Dealing with Data*，主要围绕科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性。美国一些知名数据管理领域的专家学者则从专业的研究角度出发，联合发布了一份白皮书*Challenges and Opportunities with Big Data*。该白皮书从学术的角度出发，介绍了大数据的产生，分析了大数据的处理流程，并提出大数据所面临的若干挑战。全球知名的咨询公司麦

肯锡（McKinsey）于2016年6 月发布了一份关于大数据的详尽报告*Big data: The next frontier for innovation, competition, and productivity*，对大数据的影响、关键技术和应用领域等都进行了详尽分析。从2012 年以来，大数据的关注度与日俱增。2012年1月的达沃斯世界经济论坛上，大数据是主题之一，该次会议还特别针对大数据发布了报告*Big Data, Big Impact: New Possibilities for International Development*，探讨了新的数据产生方式下，如何更好地利用数据来产生良好的社会效益。该报告重点关注了个人产生的移动数据与其他数据的融合与利用。2012年3 月，美国政府发布了《大数据研究和发展倡议》（*Big Data Research and Development Initiative*），投资2 亿以上美元，正式启动“大数据发展计划”。计划在科学研究、环境、生物医学等领域利用大数据技术进行突破。美国政府的这一计划被视为美国政府继信息高速公路（Information Highway）计划之后在信息科学领域的又一重大举措。与此同时，联合国一个名为Global Pulse的倡议项目在2016年5 月发布报告*Big Data for Development: Challenges & Opportunities*，该报告主要阐述大数据时代各国特别是发展中国家在面临数据洪流（Data Deluge）的情况下所遇到的机遇与挑战，同时还对大数据的应用进行了初步解读。《纽约时报》的文章*The Age of Big Data*则通过主流媒体的宣传使普通民众开始意识到大数据的存在，以及大数据对于人们日常生活的影响。

人类历史上从未有哪个时代和今天一样产生如此海量的数据。数据的产生已经完全不受时间、地点的限制。从开始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以下3 个阶段。

（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，实际中数据库大都为运营系统所采用，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。人类社会数据量第一次大的飞跃正是建立在运营式系统广泛使用数据库开始。这个阶段最主要的特点是数据往往伴随着一定的运营活动而产生并记录在数据库中，比如超市每销售出一件产品就会在数据库中产生相应的一条销售记录。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 的重要标志就是用户原创内容（User Generated Content，UGC）。这类数据近几年一直呈现爆炸性增长，主要有两个方面的原因。 首先，以博客、微博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；其次，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。这个阶段数据的产生方式是主动的。

（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布置于社会的各个角落，通过这些设备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根本原因。这一阶段的大数据主要来源为：

1．互联网数据源

互联网作为信息交换和网络服务的主要平台，集中大量特征形态各异的数据，成为产生大数据的重要源泉之一。在信息科技发展的历程中，互联网的出现可以比肩于人类社会“火”与“电”的发明，具有里程碑式意义。如今互联网从早期的几台机器联网已发展成全球机器联网模式，可以完全透明化地实施通信交流和资源共享。基于互联网平台的相关服务和应用已经深度融入社会，影响人们的生活工作方式，同时为社会带来巨大的经济效益。2009年阿里巴巴旗下淘宝商城“双十一”营销额为5 000万元；2011年同期达到34亿元；2012年的“双十一”为 191亿元；2013年淘宝商城销售总额达到350亿元；2014年“双十一”销售额571亿元；2015在线交易额912亿元；2016“双十一”购物狂欢节天猫交易额达1207亿；2017年双十一1小时0分49秒，成交额超过571亿元，与2014年“双十一”全天成交额持平，7点22分54秒成交额达912亿元，与2015年“双十一”全天成交额持平，9点0分4秒，成交额超1000亿元，10点40分48秒，无线成交额超过1000亿元，10点54分26秒，成交额超过1100亿元，无线占比91%，12点整，成交额1161亿元，13点9分49秒，成交额超1207亿元，与2016年“双十一”全天成交额持平，全天成交额达到1682亿元。这些新型网络服务的出现改变了传统的行为习惯，并触发新一轮的思维变革。

据2015年第36次中国互联网应用调查报告可知，中国网民的总体规模上升较快，互联网的普及率约为48.8%，网民人数达到6.7亿。大量网民聚集于网络平台，享受着互联网提供的各种优质资源，如网络新闻、搜索引擎、电子商务、即时通信/社交网络、博客微博、网络音视频和网络游戏等，主动或被动地留下大量网络使用“足迹”，汇聚成PB或EB数量级的网络数据。2014年小米云用户达到6 795.5万人，云端数据总存储量达到47 PB，而在2014年单日数据存储量最高达到380 TB。目前，国内个人云存储运营较好的是百度云，在2014年百度云整体的数据存储量超过5 EB，平均每个用户存储量约为26.84 GB。淘宝网会员约3.7亿，在线商品8.8亿，每天交易产生的数据约20 TB。根据我国互联网数据中心的《中国互联网市场洞见：互联网大数据技术创新研究2015》报告显示：截至2015 年年底，中国互联网行业持有的数据总量已达到7 900 EB，预计2020年数据持有量将增长到8 600 EB 以上。

在国外搜索巨头谷歌公司每天处理的数据量达到24 PB，换句话说，谷歌公司每天处理的数据量相当于美国国家图书馆所有纸质出版物所含数据量的上千倍。美国另一知名的社交网站Facebook，每天更新的照片数量超过1 000万张，每天网民在其网站点击按钮或写评论约30亿次。YouTube流媒体网站每月约有8亿人次的访问量，平均每秒就会有一段时长一小时以上的视频上传共享。

2．物联网数据源

2013年中国大数据专家委员会发表的《中国大数据技术与产业发展白皮书》中提出：物联

网作为当前信息科技发展中的热点，其应用所产生的数据成为大数据的重要来源之一。物联网究其本质是传感器技术进步的产物。当前各种传感监控网络无处不在，从大气监测、交通路况监测、桥梁矿井的安全监测等，到各种仪器设备状态监控和科学实验的监控传感网络，都长期不间断地返回各种数据，汇聚成大数据。当前在智慧城市建设浪潮中，几乎每个城市都在建立各种监控网络。在城市各个角落部署大量的高清监控摄像头，一个1080P的摄像头按照码流率为8 Mbit/s，在一天时间内将会产生86.4 GB的视频数据。飞机汽轮机压缩器叶片的监控数据约为588 GB/天，大约是Twitter每天产生数据的7倍左右。目前，形态各异的物联网平台不断自主产生数据，正成为大数据主要源泉之一，同时也为大数据的分析处理带来更多的挑战。

3．智能终端数据源

近年来智能终端的大量普及和带宽使用成本的逐步下降，基于通信网络平台所设计的各种服务吸引大量用户，人们通过智能终端享受网络服务已成为潮流和趋势。这一新的应用方式，对大数据的产生更是起到推波助澜的作用。截至2016年12月，我国手机网民规模达6.95亿人，与2015年相比增加7 550万人。网民中使用手机上网人群的占比提升至95.1%，网民手机上网比例进一步攀升。中国移动凭借其在移动领域内的优势，与全国大量企业和政府机构展开合作，形成一系列基于移动网络的服务，如电话会议、视频会议、集团V 网、移动办公、企业一卡通、M2M应用、视频监控、车务通等生产控制类型服务。2011年中国移动数据流量达5.77亿GB，2013年底翻番达到14.33亿GB，这种快速增长趋势将在近段时期内得到保持。据GSMA估计，至2018年全球移动数据流量将比2012年增加12倍。高速的流量增长必然导致大量的数据产生，通过对数据的挖掘分析将产生高额的经济效益回报，如在2015年电信行业的大数据应用产生的市场价值达到18.3亿元。

正如Google 的首席经济学家Hal Varian 所说，数据是广泛可用的，所缺乏的是从中提取出知识的能力。数据收集的根本目的是根据需求从数据中提取有用的知识，并将其应用到具体的领域之中。不同领域的大数据应用有不同的特点，表1-1 列举了若干具有代表性的大数据应用及其特征。

表1-1 若干具有代表性的大数据应用及其特征

应用领域	示例	用户群	响应时间	数据规模	可靠度	准确度
科学	生物信息学	小	慢	TB	中等	很高
计算						
金融	高频交易	大	很快	GB	很高	很高
社交网络	Facebook	庞大	快	PB	高	高
移动数据	移动电话	庞大	快	TB	高	高
物联网	传感器网络	大	快	TB	高	高
网络数据	新闻网站	庞大	快	PB	高	高
多媒体	视频网站	庞大	快	PB	高	中等

正是由于大数据的广泛存在，才使得大数据问题的解决很具挑战性。而它的广泛应用，则促使越来越多的人开始关注和研究大数据问题。

1.1.2 大数据的概念及特征

1. 大数据的内涵

大数据（Big Data）术语早在20世纪80年代就被提出，直到2008年科学家在*Nature*杂志上撰写文章*Big Data*: *Science in the Petabyte Era*，大数据概念逐渐被人们所熟知。2011年*Science*杂志推出专刊*Dealing with Data*，围绕科学研究中的大数据问题展开讨论，说明大数据的重要性。进入2012年大数据的研究热潮开始，全球的许多学术会议均围绕大数据议题展开。虽然大数据的研究与应用获得全球各个国家的高度重视，并取得令人惊叹的成绩，促进了社会经济的快速发展，但是大数据的定义至今未有统一的描述形式，各大研究机构和科研院所，从大数据的各个角度进行阐述得到各自相应的定义形式。

全球著名的管理咨询公司麦肯锡，也是大数据研究先驱者之一，在其研究报告*Big data*: *the next frontier for innovation*, *competition*, *and productivity*（《大数据：创新、竞争和生产力的下一个前沿领域》）给出大数据的定义：大数据是指无法通过传统的存储管理和分析处理软件进行采集、存储、管理和分析的数据对象集合。同时该报告还强调，大数据不一定要求数据量一定要到TB级别。

国际数据公司（IDC）从4个方面来描述大数据，即数据规模量大、数据快速动态可变、类型丰富和巨大的数据价值，具有这些特征的数据集合称为大数据。

研究机构Gartner提出：大数据是指超出正常处理范围，迫使用户寻求新的处理模式才能够较好地解决数据分析问题，使其具备更强的决策能力和洞察发现力，获取更多的信息资产。

维基百科关于大数据的定义是指在合理的时间内，无法通过现有软、硬件体系结构对数据资料进行收集、存储和处理，并帮助决策者进行决策服务。

全球最大的电子商务公司亚马逊公司关于大数据的定义更为简单直接，大数据就是指超越一台计算机处理能力的数据量。

综合以上几个代表性的定义可知，大数据概念较为宽泛，具备“仁者见仁、智者见智”的特点。大数据除具备数据量大外、还具备数据的多样性，关键是利用现有技术水平和处理模式，无法在一个合理的时间范围内得到所需要的信息资产。这也说明在大数据时代，我们要关心大数据本身的特点，更要关心大数据所具备的功能特性，即能够帮助人们做什么。

在信息科技发展道路上，与大数据相近的另一个术语是海量数据（Vast Data），它们都是数据化时代出现的一种现象。它们具有的共同特点是数量大，但两者之间也存在某些显著差异。Informatica中国区首席产品顾问但彬认为：大数据包含海量数据，但在形式多样性、内容复杂性方面远远超越海量数据，因此在理解大数据时可以认为是由海量数据+复杂类型的数据

构成。正是两者之间存在差异，导致在进行大数据应用时仍然存在许多技术障碍，无法把海量数据处理技术直接迁移至大数据分析环境中。

2．基本特征

目前在描述大数据特征时，一般均是按照国际数据公司IDC所提的“4V”模型来刻画，即体量大（Volume）、多样性（Variety）、速度快（Velocity）和价值（Value）。

1）体量大

当前数据正以前所未有的速度快速聚集和增长，大数据时代已经到来。在电商、社交网络、能源、制造业和服务业等领域都已积累了TB级、PB级甚至EB级的数据量。全球著名连锁超市沃尔玛每小时处理100多万条用户记录信息，维护着超过2.5 PB的客户关系数据库；在科学实验方面，如2008年投入使用的大型强子对撞机每年产生25 PB的数据；社交网络Facebook存储的照片已超过500亿张。在大数据时代，数据存储单位逐渐被PB、EB、ZB、YB所替代。

近年来，数据快速增长趋势一直持续。根据国际数据公司（IDC）的《数据宇宙》报告显示，2008年全球数据量仅为0.5 ZB，2010年就达到1.2 ZB，人类社会正式进入ZB时代。根据报告所列举的统计数据可知，2020年以前全球数据量将保持40%的速度快速增长，2020年全球数据量将达到40 ZB，此现象被人们称为“大数据爆炸定律”。2020年前全球累积的数据量变化预测趋势如图1-1所示。

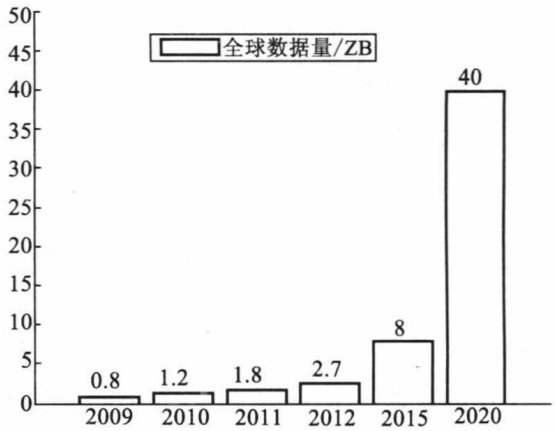

图1-1 全球数据量预测

2）多样性

大数据除了体量大外，另一个最重要的特征就是数据类型的多样化，即数据存在形式包括结构化数据、半结构化数据和非结构化数据。在早期，数据类型主要是以结构化数据为主，这一类型数据存储方便、处理简单、相关的技术非常成熟。在该阶段数据存储主要以关系数据库为主，如Oracle、SQL Server等；结构化查询语言（Structure Query Language, SQL）作为访问中间件嵌入各种开发环境中。随着互联网应用的深入，特别是社交网络、电子商务、流媒体应用环境中所出现的文本数据、交互数据、图像、视频和音频等，这些非结构化数据大量涌现加剧大数据环境中数据存储、检索和分析的难度。在2012年非结构化数据占有量占整个互联网数据量的75%以上。有统计表明，全球结构化数据增长率大约是32%，而非结构化数据增长率达到63%。相信在今后数据存储方面仍然以非结构化数据为主，因此，针对非结构化数据的处理技术和模型研究将是大数据时代数据分析的重点。

3）速度快

大数据环境中速度快有两层含义：一是数据产生快；二是要求分析处理速度快。随着各种高

性能存储设备的出现，人们对于数据产生后的高效处理有了物质基础。据统计，每秒人们通过互联网平台发送电子邮件290封；亚马逊公司每秒需要处理72.9笔客户订单。另外，在日常生活中各种监控网络每时每刻均在产生大量的数据信息，如道路交通监控网络、智慧城市等。大量的数据快速产生，信息价值稍纵即逝。因此要想从高速、体量大的大数据中获取有效信息，要求相应的大数据分析处理模型具有较高的处理速度，以满足实时性需求。针对各种应用分析实时性要求，后文把大数据分析分为在线分析（Online Analysis）和离线分析（Offline Analysis）。

4）价值

大数据拥有大量有价值信息，通过提炼的信息，能够在更高的层面和视角，将在更大的范围帮助用户提高决策力，洞察未来创造出更大的价值和商机，对社会、经济和科学研究等方面具有重要的战略意义。2010年，医疗科技公司CardioDX通过对1亿个基因样本的分析，得出能够预测冠心病的23个主要基因信息；通过对社交网络和微博上的舆情监控分析，及时跟踪社会动态，实现对突发事件进行预警和疏导。电子商务网站通过对顾客在网络上的点击和停留时间等行为分析，实现商品的精准推荐等。

通常情况下，大数据背后的价值信息分布毫无规律，隐藏较深。发现大数据价值势必为大数据的分析预测环节带来挑战，并要求预测分析系统具备高性能、实时性、可扩展性等特征。纵观大数据特征和分析环境可知，要想实现大数据价值的有效分析需具备三大要素，即大分析（Big Analytic）、大带宽（Big Bandwidth）、大内容（Big Content）。大分析是指通过新的方法实现对大数据快速、高效、实时的分析计算，旨在得出数据之间的隐含规律，帮助用户掌握事件背后的机理、预测发展趋势，得到更大的价值；大带宽是指提供良好的通信设施基础，以便能够在更大的范围、较复杂的环境中，使各节点之间的数据传输高效安全，为大分析奠定基础；大内容是指价值信息隐匿较深，需要足够多、足够大的数据才能更加有效地挖掘出其具有的规律。因此，大分析是技术实现途径，大带宽是物质保障，大内容是获取大价值的前提条件。

1.2 大数据带来的变革

大数据时代已经到来，认同这一判断的人越来越多。那么大数据意味着什么，它到底会改变什么？仅仅从技术角度回答，已不足以解惑。我们需要把大数据放在人的背景中加以透视，理解它作为时代变革力量的所以然。

1. 对价值的变革

未来十年，决定一个国家是不是有大智慧的核心是国民幸福。一体现在民生上，二体现在生态上，通过大数据让有意义的事变得明晰，看我们在人与人的关系上做得是否比以前更有意义。总之，让我们从前十年的意义混沌时代，进入未来十年的意义明晰时代。

2．对经济的变革

生产者是有价值的，消费者是价值的意义所在。有意义的才有价值，消费者不认同的，就卖不出去，就实现不了价值；只有消费者认同的，才卖得出去，才能实现价值。大数据帮助我们从消费者这个源头识别意义，从而帮助生产者实现价值。这就是启动内需的原理。

3．对组织的变革

随着具有语义网特征的数据基础设施和数据资源发展起来，组织的变革就越来越显得不可避免。大数据将推动网络结构产生无组织的组织力量。最先反映这种结构特点的，是各种各样去中心化的 Web 2.0 应用，如 RSS、维基、博客等。 大数据之所以成为时代变革力量，在于它通过追随意义而获得智慧。

4．对思维的变革

在舍恩伯格的《大数据时代：生活、工作与思维的大变革》一书中指出，大数据时代对社会的最大影响就是对人们思维方式的3种转变，即：

（1）全样而非抽样；在过去，由于缺乏获取全体样本的手段，人们发明了“随机调研数据”的方法。理论上，抽取样本越随机，就越能代表整体样本。但问题是获取一个随机样本代价极高，而且很费时。人口调查就是典型一例，一个稍大一点的国家甚至做不到每年都发布一次人口调查，因为随机调研实在是太耗时耗力了。 但有了云计算和数据库以后，获取足够大的样本数据乃至全体数据，就变得非常容易了。谷歌可以提供谷歌流感趋势的原因就在于它几乎覆盖了7成以上的北美搜索市场，这些数据完全没有必要抽样调查，所有记录都在那里等待人们挖掘和分析。

（2）效率而非精确；过去使用抽样的方法，就需要在具体运算上非常精确，因为所谓“差之毫厘便失之千里”。设想一下，一个总样本为1亿人随机抽取1 000人，如果在抽取的1 000人中运算出现错误的话，那么放大到1亿人会有多大的偏差。但全样本时，有多少偏差就是多少偏差而不会被放大。谷歌人工智能专家诺维格，在他的论文中写道：大数据基础上的简单算法比小数据基础上的复杂算法更加有效。数据分析的目的并非仅仅就是数据分析，而是有其他用途，故而时效性也非常重要。精确的计算是以时间消耗为代价的，但在小数据时代，追求精确是为了避免放大的偏差而不得已为之。但在样本=总体的大数据时代，“快速获得一个大概的轮廓和发展脉络，就要比严格的精确性重要得多”。

（3）相关而非因果。相关性表明变量A和变量B有关，或者说A变量的变化和B变量的变化之间存在一定的正比（或反比）关系。但相关性并不一定是因果关系（A未必是B的因）。亚马逊的推荐算法非常有名，它能够根据消费记录告诉用户可能会喜欢什么，这些消费记录有可能是别人的，也有可能是该用户历史上的。但它不能说出你为什么会喜欢的原因。难道大家都喜欢购买A和B，就一定等于你买了A之后就是买B吗？未必，但的确需要承认，相关性很高——或者说，概率很大。 舍恩伯格认为，大数据时代只需要知道是什么，而无须知道为什么，就像亚马逊推荐算法一样，知道喜欢A的人很可能喜欢B但却不知道其中的原因。

1.3 大数据的价值与挑战

1.3.1 大数据的价值

随着物联网、云计算、Web 2.0和移动互联等技术的快速发展，各种应用产生的数据正快速增长，人类社会已经步入大数据时代。不容置疑，大数据背后隐藏大信息，挖掘出其中的大信息将创造出大价值，同时能够帮助决策者进行趋势预测、洞察未来。国际数据公司（IDC）报告指出，2017年大数据应用所带来的市场价值达到324亿美元。目前，全球各国都高度重视大数据发展和应用，均制定出相应的政策措施予以支持大数据技术的研制和开发，同时各大IT公司均利用其技术优势或者数据资源成立专门研究机构进行大数据研究分析。

基于大数据挖掘分析成功预测金融危机的发生，一直成为数据挖掘分析中的经典案例。2008年受美国次贷危机的影响，席卷全球的金融危机悄无声息的发生。许多国家、企业对此毫无察觉，在这轮金融危机中纷纷中招，蒙受巨大损失。在中国，马云通过整合旗下电子商务网站中询盘数据和订单数据等信息，发现海外企业近期的询盘数量和采购量在急剧下滑。基于这些海量数据的分析结论，马云提前六个月的时间准确预测出世界外贸经济走势，得出将爆发金融危机的结论。这一预测结论提醒企业做好准备、抵御金融危机所带来的影响，成功度过经济发展的冬天。

2009年全球出现一种新型流感病毒H1N1，该病毒具有禽流感和猪流感的特点，传播速度极快。短短几周时间迅速漫延，全球陷入恐慌。按传统处理流程，在出现新型未知病毒时，政府机构要求一线医生及时把病例上传至疾病控制中心，再由疾病控制中心汇集信息进行预测。这一处理流程中，将出现两个滞后时间节点：一是病人就医滞后；二是信息汇集需要时间导致滞后。这种滞后直接导致医疗机构在疫情爆发关键时期无所事事，而在大面积传播后无所适从的困境。互联网巨头Google保存历年来人们的网上搜索记录的相关词条，如治疗咳嗽、发热等，然后依据特定的检索记录频率和时间、空间建立分析预测系统。2009年当甲型H1N1流感爆发，Google利用其数据汇聚的优势，凭借分析预测系统，准确及时发出预警信息。

大数据的价值在商业领域的作用更是不可估量，如商业广告精准推送。作为中国传媒业领军的中央电视台，在2013年前广告收入独占鳌头、遥遥领先。凭借网民数量的上升，基于历史数据进行大数据分析，广告推送不再是盲目、被动的过程，实现智能化推送，提高广告宣传效益。2013年百度广告收入达到319.44亿元，远超过中央电视台的运营收入。成为新传媒技术完胜传统传媒手段中的里程碑式事件，具有划时代的意义。

上述示例仅是冰山一角，依赖大数据分析所带来的巨额回报的案例数不胜数。它们的共同特点是通过汇聚历史数据构建出大数据模型，在此基础上进行预测分析，挖掘出数据背后所隐藏的大信息，指导商务决策或者趋势分析。

1.3.2 大数据时代面临的新挑战

大数据时代的数据存在如下几个特点：多源异构、分布广泛、动态增长、先有数据后有模式。正是这些与传统数据管理迥然不同的特点，使得大数据时代的数据管理面临新的挑战。

1．数据集成的挑战

数据的广泛存在性使得数据越来越多地散布于不同的数据管理系统中，为了便于进行数据分析需要进行数据的集成。数据集成看起来并不是一个新的问题，但是大数据时代的数据集成却有了新的需求，因此也面临新的挑战。

（1）广泛的异构性。传统的数据集成中也会面对数据异构的问题，但是在大数据时代这种异构性出现了新的变化。主要体现在：①数据类型从以结构化数据为主转向结构化、半结构化、非结构化三者的融合。②数据产生方式的多样性带来的数据源变化。传统的数据主要产生于服务器或者是个人计算机，这些设备位置相对固定。随着移动终端的快速发展，手机、平板电脑、GPS 等产生的数据量呈现爆炸式增长，且产生的数据带有很明显的时空特性。③数据存储方式的变化。传统数据主要存储在关系数据库中，但越来越多的数据开始采用新的数据存储方式来应对数据爆炸，比如存储在Hadoop 的HDFS 中。这就必然要求在集成的过程中进行数据转换，而这种转换的过程是非常复杂和难以管理的。

（2）数据质量。数据量大不一定就代表信息量或者数据价值的增大，相反很多时候意味着信息垃圾的泛滥。一方面单个系统很难容纳下从不同数据源集成的海量数据；另一方面如果在集成的过程中仅仅简单地将所有数据聚集在一起而不做任何数据清洗，会使得过多的无用数据干扰后续的数据分析过程。大数据时代的数据清洗过程必须更加谨慎，因为相对细微的有用信息混杂在庞大的数据量中。如果信息清洗的粒度过细，很容易将有用的信息过滤掉。清洗粒度过粗，又无法达到真正的清洗效果，因此在质与量之间需要进行仔细的考量和权衡。

2．数据分析的挑战

传统意义上的数据分析（Analysis）主要针对结构化数据展开，且已经形成了一整套行之有效的分析体系。首先利用数据库来存储结构化数据，在此基础上构建数据仓库，根据需要构建数据立方体进行联机分析处理（Online Analytical Processing，OLAP），可以进行多个维度的下钻（Drill-down）或上卷（Roll-up）操作。对于从数据中提炼更深层次的知识的需求促使产生了数据挖掘技术，并发明了聚类、关联分析等一系列在实践中行之有效的方法。这一整套处理流程在处理相对较少的结构化数据时极为高效。但是，随着大数据时代的到来，半结构化和非结构化数据量的迅猛增长，给传统的分析技术带来了巨大的冲击和挑战，主要体现在：

（1）数据处理的实时性（Timeliness）。随着时间的流逝，数据中所蕴含的知识价值往往也在衰减，因此很多领域对于数据的实时处理有需求。随着大数据时代的到来，更多应用场景的数据分析从离线（Offline）转向了在线（Online），开始出现实时处理的需求，比如KDD 2012最佳论文所探讨的实时广告竞价问题。大数据时代的数据实时处理也面临着新的挑战，主要体现在数