生态环境大数据

ECOLOGICAL ENVIRONMENT

BIG DATA ON ECOLOGICAL ENVIRONMENT

汪先锋　编著

中国环境出版集团

生态环境大数据

汪先锋　编著

中国环境出版集团·北京

图书在版编目（CIP）数据

生态环境大数据/汪先锋编著. —北京：中国环境出版集团，2019.10
ISBN 978-7-5111-4118-7

Ⅰ. ①生… Ⅱ. ①汪… Ⅲ. ①生态环境—山东—教材
Ⅳ. ①X321.252

中国版本图书馆 CIP 数据核字（2019）第 221190 号

出 版 人 武德凯
责任编辑 韩 睿
责任校对 任 丽
封面设计 彭 杉

出版发行 中国环境出版集团
（100062 北京市东城区广渠门内大街 16 号）
网 址：http://www.cesp.com.cn
电子邮箱：bjgl@cesp.com.cn
联系电话：010-67112765（编辑管理部）
发行热线：010-67125803，010-67113405（传真）
印 刷 北京建宏印刷有限公司
经 销 各地新华书店
版 次 2019 年 10 月第 1 版
印 次 2019 年 10 月第 1 次印刷
开 本 787×1092 1/16
印 张 26.75
字 数 550 千字
定 价 80.00 元

前　言

大数据（Big Data）现在可以说是人尽皆知，国内外的政府和企业都高度重视大数据的建设应用，大数据的发展也进入了“快车道”。全球范围内，运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势，有关发达国家相继制定实施大数据战略性文件，大力推动大数据发展和应用。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，各地政府已相继启动大数据相关工作。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

大数据时代的来临，改变了数据与信息的传统处理方式，为生态环境监管和治理带来了前所未有的机遇。

当前，生态环境保护工作面临着许多新情况和新问题，要求我们必须创新工作思路，深入研究运用新方式和大数据思维破解难题。全面提升生态环境保护能力，就要走精细化、智慧化管理之路，而生态环境大数据就是这条道路的强力支撑！大数据、云计算、物联网、移动互联网、区块链等信息技术已成为推进环境治理体系和治理能力现代化的重要手段。2016 年 3 月，原环境保护部发布《生态环境大数据总体方案》，正式开启了全国生态环境大数据建设的大幕。大数据应用和发展趋势势不可挡，大数据必将对环境管理理念、管理方式产生巨大的影响，必将引领环境治理现代化和环境监管精细化、智慧化的新变革、新发展。开展生态环境大数据应用具有重要的现实意义和紧迫的需求。

近些年来，本书作者亲历并组织了物联网技术在环境管理中的广泛应用，通过构建全方位、多层次、全覆盖的环境自动监控网络，实现全天候 24 小时不

间断的对污染源排放企业和空气质量等进行自动监控，推动环境信息资源高效、精准、实时传递。通过构建海量环境信息资源中心和统一的环境信息服务支撑平台，支持各生态环境保护业务的全过程智能化，能够节省人力投入，提高环境监管能力，解决监管中的时效性、权威性等问题。

从2017年1月起，本书作者亲自组织实施了山东省生态环境大数据建设，历时10个月编制完成了《山东省生态环境大数据发展规划》和《山东省生态环境大数据建设方案》。通过科学的顶层设计，准确确定生态环境大数据建设应用的范围，把握建设重点，避免重复建设，详细规划建设任务与实施路径，并把顶层设计上升到决策高度，保证顶层设计落实。只有从顶层设计着手，统一规划、统一标准、统一建设、统一管理，全面优化现有系统，构建新增业务系统，才能建立好高效可用的“用数据管理、用数据决策、用数据服务”的生态环境大数据体系。

本书从理论与实践相结合的角度出发，对大数据技术在环境监管中的应用、发展方向及趋势进行了阐述，并结合建设成果对生态环境大数据建设过程进行了较为详细的介绍。作者希望通过创新理论方法来指导环境信息化工作实践，逐步总结探索出环境信息化对环境管理的支撑服务道路。

本书认为，现阶段生态环境监管工作急需通过引入大数据思维和理念，提升环境监管水平、实现管理创新，通过整合来优化资源管理，促使加快改善本地区的环境质量。落实在具体行动上，就是要通过生态环境大数据平台建设，将核心环境管理业务都串联起来，实现省、市、县三级生态环境业务一体化协同，向管理者提供统一、可靠、真实的数据，再通过大数据技术运用基于可视化方法的环境数据分析结果和治理模型，立体化展现本地区的环境变化趋势，使生态环境管理部门主动把握环境管理重点，有针对性的开展监管和治理工作，不断加强生态环境大数据综合应用和集成分析，为生态环境科学决策提供强有力支撑。

本书作者进入环保领域近二十年，见证并亲历了国内外信息技术、互联网、物联网、云计算、大数据和环境自动监控系统的发展与应用，组织参与了很多环境保护业务信息系统的开发、应用与管理，积累了丰富的环境信息化实践经

验。通过开展山东省生态环境大数据建设，总结了生态环境大数据应用经验，展示了建设成果，供生态环境管理者和环境信息化战线的同仁们借鉴。

本书共分十一章。第一章为大数据概述，主要介绍大数据的发展背景与历程、概念等。第二章介绍了大数据的关键技术，从数据采集到数据可视化的全过程技术应用。第三章对生态环境大数据进行了阐述，重点论述了生态环境大数据的特点、重要意义，对生态环境大数据面临的挑战和发展途径提出学术见解。第四章介绍了生态环境大数据平台的总体架构和技术架构。第五章对生态环境大数据支撑平台的基础支撑服务和业务应用支撑服务进行阐述。第六章重点论述了生态环境大数据资源中心的规划、建设和管理。第七章梳理了建设生态环境大数据平台需要的各类数据资源。第八章从数据采集、数据处理、数据挖掘、可视化和大数据共享服务等方面介绍生态环境大数据的应用开发实践。第九章介绍生态环境大数据的安全建设与管理，从安全体系建立、数据安全防护、备份与恢复等方面进行了论述。第十章重点阐述了生态环境大数据保障体系建设。第十一章介绍了山东省生态环境大数据平台建设的主要做法。附录为生态环境大数据总体建设方案全文。

本书可作为环境信息化、环境监管等人员和环境科学专业研究生及教职人员、相关IT人员的参考书。

在本书的编写过程中，作者参阅了国内外大量的文献资料，在此向这些文献资料的原作者表示衷心感谢。在参考文献中如有漏掉引用出处者，敬请谅解。

大数据在生态环境领域的应用还刚刚开始，而信息技术的发展又非常迅速，本书涉及的专业知识较为广泛，由于编著时间仓促和作者理论水平有限，难免有错误和不足之处，恳请广大读者批评指正。

本书在出版过程中，得到朱凤涛先生的大力支持和帮助，在此表示衷心感谢。最后，也感谢家人的大力支持和付出！

汪先锋

2019 年 10 月

目　录

第一章　大数据概述

第一节　大数据的发展背景与历程

大数据（Big Data）现在可以说是人尽皆知，国内外的政府和企业都高度重视大数据的发展应用，大数据的发展已进入了快车道。

一、大数据的发展背景

几年前，人们把大规模数据称为“海量数据”，但实际上，大数据这个概念早在2008年就已被提出。2008年，在Google成立10周年之际，著名的《自然》杂志出版了一期专刊，专门讨论未来与大数据处理相关的一系列技术问题和挑战，其中就提出了“Big Data”的概念。

由于大数据处理需求的迫切性和重要性，近年来大数据技术已经引起了全球学术界、工业界和各国政府的高度关注和重视，全球掀起了一场可与20世纪90年代的信息高速公路相提并论的研究热潮。美国和欧洲一些发达国家政府都从国家科技战略层面提出了一系列的大数据技术研发计划，以推动政府机构、重大行业、学术界和工业界对大数据技术的探索研究和应用。

早在2010年12月，美国总统办公室下属的科学技术顾问委员会（PCAST）和信息技术顾问委员会（PITAC）向奥巴马和国会提交了一份《规划数字化未来》的战略报告，把大数据收集和使用的工作提升到体现国家意志的战略高度。报告列举了5个贯穿各个科技领域的共同挑战，而第一个最重大的挑战就是“数据”问题。报告指出：“如何收集、保存、管理、分析、共享正在呈指数增长的数据是我们必须面对的一个重要挑战。”报告建议：“联邦政府的每一个机构和部门，都需要制定一个‘大数据’的战略。”2012年3月，美国总统奥巴马签署并发布了一个“大数据研究发展创新计划”（Big Data R&D Initiative），由美国国家自然基金会（NSF）、卫生健康总署（NIH）、能源部（DOE）、国防部（DOD）等六大部门联合，投资2亿美元启动大数据技术研发，这是美国政府继1993年宣布“信

息高速公路”计划后的又一次重大科技发展部署。美国白宫科技政策办公室还专门支持建立了一个大数据技术论坛，鼓励企业和组织机构间的大数据技术交流与合作。

2012 年 7 月，联合国在纽约发布了一本关于大数据政务的白皮书《大数据促发展：挑战与机遇》，全球大数据的研究和发展进入了前所未有的高潮。这本白皮书总结了各国政府如何利用大数据响应社会需求，指导经济运行，更好地为人民服务，并建议成员国建立“脉搏实验室”（Pulse Labs），挖掘大数据的潜在价值。

由于大数据技术的特点和重要性，目前国内外已经出现了“数据科学”的概念，即数据处理技术将成为一个与计算科学并列的新的科学领域。已故著名图灵奖获得者 Jim Gray 在 2007 年的一次演讲中提出，“数据密集型科学发现”（Data-Intensive Scientific Discovery）将成为科学研究的第四范式，科学研究将从实验科学、理论科学、计算科学，发展到目前兴起的数据科学。

为了紧跟全球大数据技术发展的浪潮，我国政府、学术界和工业界对大数据也给予了高度关注。央视于 2013 年 4 月 14 日和 21 日邀请了《大数据时代——生活、工作与思维的大变革》作者维克托·迈尔·舍恩伯格，以及美国大数据存储技术公司 LSI 总裁阿比分别做客《对话》节目，做了两期大数据专题谈话节目《谁在引爆大数据》《谁在掘金大数据》，国家央视媒体对大数据的关注和宣传体现了大数据技术已经成为国家和社会普遍关注的焦点。

而国内的学术界和工业界也都迅速行动，广泛开展大数据技术的研究和开发。2012 年 7 月出版的《大数据》（涂子沛著）是中国大数据领域第一本著作，引领了中国社会对大数据战略、数据治国和开放数据的讨论。2013 年以来，国家自然科学基金、“973 计划”、核高基、“863 计划”等重大研究计划都已经把大数据研究列为重大的研究课题。为了推动我国大数据技术的研究发展，2012 年中国计算机学会（CCF）发起组织了 CCF 大数据专家委员会，CCF 专家委员会还特别成立了一个“大数据技术发展战略报告”撰写组，并已撰写发布了《2013 年中国大数据技术与产业发展白皮书》。

大数据在带来巨大技术挑战的同时，也带来了巨大的技术创新与商业机遇。不断积累的大数据包含着很多在小数据量时不具备的深度知识和价值，大数据分析挖掘将能为行业/企业带来巨大的商业价值，实现各种高附加值的增值服务，进一步提升行业/企业的经济效益和社会效益。由于大数据隐含着巨大的深度价值，美国政府认为大数据是“未来的新石油”，对未来的科技与经济发展将带来深远影响。因此，在未来，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、控制和运用也将成为国家间和企业间新的争夺焦点。

大数据的研究和分析应用具有十分重大的意义和价值。被誉为“大数据时代预言家”的维克托·迈尔·舍恩伯格在其《大数据时代——生活、工作与思维的大变革》一书中列举了大量翔实的大数据应用案例，并分析预测了大数据的发展现状和未来趋势，提出了很

多重要的观点和发展思路。他认为“大数据开启了一次重大的时代转型”，指出大数据将带来巨大的变革，改变我们的生活、工作和思维方式，改变我们的商业模式，影响我们的经济、政治、科技和社会等各个层面。

由于大数据行业应用需求日益增长，未来越来越多的研究和应用领域将需要使用大数据并行计算技术，大数据技术将渗透到每个涉及大规模数据和复杂计算的应用领域。不仅如此，以大数据处理为中心的计算技术将对传统计算技术产生革命性的影响，广泛影响计算机体系结构、操作系统、数据库、编译技术、程序设计技术和方法、软件工程技术、多媒体信息处理技术、人工智能以及其他计算机应用技术，并与传统计算技术相互结合产生很多新的研究热点和课题。

大数据给传统的计算技术带来了很多新的挑战。大数据使得很多在小数据集上有效的传统的串行化算法在面对大数据处理时难以在可接受的时间内完成计算；同时大数据含有较多噪声、样本稀疏、样本不平衡等特点，使得现有的很多机器学习算法有效性降低。因此，微软全球副总裁陆奇博士在 2012 年全国第一届“中国云/移动互联网创新大奖赛”颁奖大会主题报告中指出：“大数据使得绝大多数现有的串行化机器学习算法都需要重写。”

大数据技术的发展将给我们研究计算机技术的专业人员带来新的挑战和机遇。目前，国内外 IT 企业对大数据技术人才的需求正快速增长，未来 5～10 年内业界将需要大量的掌握大数据处理技术的人才。IDC 研究报告指出：“下一个 10 年里，世界范围的服务器数量将增长 10 倍，而企业数据中心管理的数据信息将增长 50 倍，企业数据中心需要处理的数据文件数量将至少增长 75 倍，而世界范围内 IT 专业技术人才的数量仅能增长 1.5 倍。”因此，未来 10 年内大数据处理和应用需求与能提供的技术人才数量之间将存在一个巨大的差距。目前，由于国内外高校开展大数据技术人才培养的时间不长，技术市场上掌握大数据处理和应用开发技术的人才十分短缺，因而这方面的技术人才十分抢手，供不应求。国内几乎所有著名的 IT 企业，如百度、腾讯、阿里巴巴、奇虎 360 等，都需要大量的大数据技术人才。

2015 年 10 月，党的十八届五中全会公报提出要实施“国家大数据战略”，这是大数据第一次写入党的全会决议，标志着大数据战略正式上升为国家战略，五中全会开启了我国大数据建设的新篇章。中共中央政治局 2017 年 12 月 8 日下午就实施国家大数据战略进行第二次集体学习。习近平总书记强调推动实施国家大数据战略，加快完善数字基础设施，推进数据资源整合和开放共享，保障数据安全，加快建设数字中国，更好地服务我国经济社会发展和人民生活改善。

二、大数据的发展历程

大数据的发展历程总体上可以划分为三个重要阶段，初始期、发展期、大规模应用期。

第一阶段——初始期：20 世纪 90 年代到 21 世纪初。随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。

第二阶段——发展期：21 世纪前 10 年。Web2.0 应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌 GFS 和 MapReduce 等大数据技术受到追捧，Hadoop 平台开始大行其道。

第三阶段——大规模应用期：2010 年以后。大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅度提高。

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——在考克斯和埃尔斯沃斯案例中，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，甚至远超磁盘的承载能力。”他们称之为“大数据问题”。

2002 年，在“9·11”恐怖袭击后，美国政府为阻止恐怖主义已经涉足大规模数据挖掘，国家安全前顾问约翰·波因德克斯特领导国防部整合现有政府的数据集，组建一个用于筛选通信、犯罪、教育、金融、医疗和旅行等记录来识别可疑人的大数据库。一年后国会因担忧公民自由权而停止了这一项目。

2004 年，“9·11”委员会呼吁反恐机构应统一组建“一个基于网络的信息共享系统”，以便能快速处理应接不暇的数据。在 2010 年时，美国国家安全局的 30 000 名员工将拦截和存储 17 亿/年的电子邮件、电话和其他通信日报。与此同时，零售商积累了关于客户购物和个人习惯的大量数据，沃尔玛声称已拥有一个容量为 460 字节的缓存器——比当时互联网上的数据量还要多 1 倍。

2007—2008 年，随着社交网络的激增，技术博客和专业人士为“大数据”概念注入新的生机。“当前世界范围内已有的一些其他工具将被大量数据和应用算法取代。”《连线》的克里斯·安德森认为当时处于一个“理论终结时代”。一些政府机构和美国的顶尖计算机科学家声称：“应该深入参与大数据计算的开发和部署工作，因为它将直接有利于许多任务的实现。”

2008 年，《自然》杂志推出大数据专刊；计算社区联盟（Computing Community Consortium）发表了报告《大数据计算：在商业、科学和社会领域的革命性突破》，阐述了大数据技术及其面临的一些挑战。

2009 年 1 月，印度政府建立印度唯一的身份识别管理局，对 12 亿人的指纹、照片和虹膜进行扫描，并为每人分配 12 位的数字 ID 号码，将数据汇集到世界最大的生物识别数据库中。

2009 年 5 月，美国总统奥巴马政府推出 data.gov 网站作为政府开放数据计划的部分举

措。该网站的超过4.45万量数据集被用于保证一些网站和智能手机应用程序来跟踪从航班到产品召回再到特定区域内失业率的信息，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2009年7月，为应对全球金融危机，联合国秘书长潘基文承诺创建警报系统，抓住“实时数据带给贫穷国家经济危机的影响”。联合国全球脉冲项目已研究了对如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病暴发之类的问题。

2011年2月，扫描2亿年的页面信息，或4兆字节磁盘存储，只需几秒即可完成。IBM的沃森计算机系统在智力竞赛节目《危险边缘》中打败了两名人类挑战者。后来《纽约时报》称这一刻为“大数据计算的胜利”。

2011年2月，《科学》杂志推出专刊《处理数据》，讨论了科学研究中的大数据问题。

2011年，维克托·迈尔·舍恩伯格出版著作《大数据时代——生活、工作与思维的大变革》引起轰动。

2011年5月，麦肯锡全球研究院发布《大数据：下一个具有创新力、竞争力与生产力的前沿领域》，提出“大数据”时代到来。

2012年3月，美国政府报告要求每个联邦机构都要有一个“大数据”的策略，作为回应，奥巴马政府宣布一项耗资2亿美元的大数据研究与发展项目，发布了《大数据研究和发展倡议》。

2012年，涂子沛的《大数据》出版，是中国大数据领域第一本著作，引领了中国社会对大数据战略、数据治国和开放数据的讨论。

2013年12月，中国计算机学会发布《中国大数据技术与产业发展白皮书》，系统总结了大数据的核心科学与技术问题，推动了我国大数据学科的建设与发展，并为政府部门提供了战略性的意见与建议。

2014年5月，美国政府发布2014年全球“大数据”白皮书《大数据：抓住机遇、守护价值》，报告鼓励使用数据来推动社会进步。

2015年8月，国务院印发《促进大数据发展行动纲要》，全面推进我国大数据发展和应用，加快建设数据强国。

2015年10月，党的十八届五中全会提出“实施国家大数据战略”，这是大数据第一次写入党的全会决议，标志着大数据战略正式上升为国家战略，五中全会开启了我国大数据建设的新篇章。

2016年12月，工信部发布《大数据产业发展规划（2016—2020年）》，有力地推进了我国大数据技术创新和产业发展。

2017年，国家大数据（贵州）综合试验区首批107家重点企业名单公布。

2017年12月8日，中共中央政治局就实施国家大数据战略进行第二次集体学习，中共中央总书记习近平在主持学习时强调，实施国家大数据战略，加快建设数字中国。

2017 年，全球的数据总量为 21.6ZB（1 个 ZB 等于十万亿亿字节），目前全球数据的增长速度在每年 40%左右。

2018 年，达沃斯世界经济论坛等全球性重要会议都把“大数据”作为重要议题，进行讨论和展望。

第二节　大数据的概念

最早提出大数据时代到来的是美国的麦肯锡，他指出：“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。”

当今“大数据”一词的重点其实已经不仅在于数据规模的定义，它更代表着信息技术发展进入了一个新的时代，代表着爆炸性的数据信息给传统的计算技术和信息技术带来的技术挑战和困难，代表着大数据处理所需的新的技术和方法，也代表着大数据分析和应用所带来的新发明、新服务和新的发展机遇。

一、大数据的概念

然而，到底什么是大数据？它的概念和外延包括哪些？由于大数据是最近新衍生出来的概念，它的内涵和外延也在不断地拓展和变化着，目前还没有一个被业界广泛采纳的明确定义。

随着大数据概念的普及，人们常常会问，多大的数据才叫大数据？其实，关于大数据，难以有一个非常定量的定义。维基百科给出了一个定性的描述：大数据是指无法使用传统和常用的软件技术与工具在一定时间内完成获取、管理和处理的数据集。

在维克托·迈尔·舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中，大数据是指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。

对于“大数据”，研究机构 Gartner 给出了这样的定义：“大数据”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来适应海量、高增长率和多样化的信息资产。

麦肯锡全球研究所曾经给大数据做了一个定义：超出传统的数据库软件工具处理能力的超大规模的数据集。但是大数据带来的技术方面的挑战，远远不止于处理工具，事实上对传统的网络结构、计算模型、安全体系，提出了全方位的课题。其主要包括以下几个方面：

第一，网络承载能力要满足“数据摩尔定律”的需要。数据摩尔定律，是指数据在未来 18 个月内，数据量将增加一倍。

第二，需要建立自主可控的安全防护体系、身份识别体系。必须在网络空间实现“4W”的机制，即“Who”“Where”“When”“What”。在网络空间中，安全能力必须能够对任何一个单体，掌握“在任何时间、任何地点的状态”的数据。

第三，需要参考仿生学，建立起“社会计算”的模型，应对日益增长的海量数据。

国务院2015年发布的《促进大数据发展行动纲要》（国发〔2015〕50号）对大数据做出这样的定义：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，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。

人们对大数据概念理解的不一致和认识上的分歧实际上反映了现有的大数据概念与现实需求的脱节，特别是与政府需求的脱节。作者认为，从推进国家信息化发展的角度看，对大数据进行严格定义或许并不重要，能够利用大数据提升全民数据意识、发展数据文化、释放数据红利、打造数据优势才是硬道理。大数据热强化了社会的数据意识，这对于中国才是至关重要的。

作者认为，大数据不是一项专门的技术，而是一系列信息技术的综合应用，《促进大数据发展行动纲要》给出的定义比较符合当前大数据发展和应用状况。

二、大数据的特征

IDC（International Data Corporation）在它编制的年度数据宇宙研究报告《从混沌中提取价值》（Extracting Value from Chaos）中给大数据下了一个定义：大数据技术是新一代技术与架构，它被设计用于在成本可承受的条件下，通过非常快速的采集、发现和分析，从大体量、多类别的数据中提取价值。

IDC的定义描述了大数据时代的四大特征，即俗称的“4V”，而这“4V”（Volumes、Variety、Velocity、Value）也被广泛地认为是大数据的最基本的内涵。

1. 海量化（Volumes）

数据体量巨大是大数据的首要特征，也是大家最容易发现的特征。全球数据正在以前所未有的速度增长着，每天都有数以百万兆字节的数据在互联网上产生。全球的数据储量仅在2011年就达到1.8ZB（或1.8万亿GB），相当于每个美国人每分钟写3条Twitter信息，总共写2.697 6万年。2015年全球大数据储量达到8.61ZB。预计到2020年全球数据总量将达到40ZB。40ZB相当于整个世界人口（到2017年为76亿人）全年每天观看14.5小时的高清视频流所产生的数据量。2016年微信月活跃用户达到8.893亿，正式超越QQ的8.685亿。在用户数和数据量上，微信超过QQ，成为名副其实的腾讯第一大平台和底层基础。数据量的快速增长已经远远超越单个计算机的存储和处理能力，数据中心处理能力变得日益重要，同时也驱动着数据中心网络不断向大带宽低时延方向演进。

2. 多样化（Variety）

数量类型的日趋繁多是大数据的另外一个显著特征。海量数据有不同格式，第一种是结构化数据，我们常见的数据大部分以二维表的形式存储在数据库中。第二种是非结构化数据，随着互联网多媒体应用的发展和兴起，图片、视频、音频等数据大量出现，这些数据的处理方式比较复杂，数据类型非常繁多。非结构化数据的超大规模和增长，占总数据量的 80%～90%，比结构化数据增长快 10～50 倍，是传统数据仓库的 10～50 倍。如何有效地处理非结构化数据，并挖掘出其中蕴含的商业价值和经济社会价值，是大数据技术要解决的问题。

3. 快速化（Velocity）

快速处理是大数据必须满足的基本要求。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。经济全球化形势下，企业面临的竞争环境越来越严酷。在此情况下，如何及时把握市场动态，深入洞察行业、市场、消费者的需求，并快速、合理地制定经营策略，就成为企业生死存亡的关键。而对大数据的快速处理分析，是实现这一目标的前提。

4. 价值比（Value）

大数据蕴含的整体价值是非常巨大的。大量的各类数据和信息，不经过处理则价值较低，属于价值密度低的数据。挖掘大数据的有用价值并加以利用，是数据拥有者的自然目标。以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒。海量数据分析非常复杂，使得过去单纯依靠数据库 BI 已经不是太适合了。市场形势瞬息万变，因此，如何在海量的、多样化的、低价值密度的数据中快速挖掘出其蕴含的有用价值，是大数据技术的革命。

三、大数据的现实价值

大量数据正在成为一种资源，一种生产要素，渗透至各个领域，而拥有大数据能力，即善于聚合信息并有效利用数据，将会带来层出不穷的创新，从某种意义上说它代表着一种生产力，麦肯锡认为，“人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来”。

大数据将带来此起彼伏的 IT 技术革命。为解决日益增长的海量数据、数据多样性、数据处理时效性等问题，一定会在存储器、数据仓库、系统架构、人工智能、数据挖掘分析以及信息通信等方面不断涌现突破性技术。当今世界 IT 巨头、IT 敏锐的创新者们正努力耕耘在大数据技术领域，大数据将成为 IT 的主战场。

大数据将在各行各业引发各类创新模式。随着大数据的发展，行业渐进融合，以前认为不相关的行业通过大数据技术有了相通的渠道，沃尔玛通过数据挖掘将风马牛不相及的