大数据与智慧社会

数据驱动变革、构建未来世界

张克平
陈曙东 主编

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

大数据与智慧社会

数据驱动变革、构建未来世界

张克平
陈曙东 ◎主编

人民邮电出版社
北　京

图书在版编目（CIP）数据

大数据与智慧社会 ： 数据驱动变革、构建未来世界 / 张克平，陈曙东主编. -- 北京 ： 人民邮电出版社，2017.6
ISBN 978-7-115-45624-3

Ⅰ. ①大… Ⅱ. ①张… ②陈… Ⅲ. ①数据管理 Ⅳ. ①F279.23

中国版本图书馆CIP数据核字(2017)第074300号

内容提要

大数据正在改变人们的生活、社会的运行方式以及各行业的竞争生态，是提升政府治理水平和企业竞争力的核心要素。然而，政府和企业如何才能抓住大数据带来的宝贵机遇，改善公共服务、激发商业创新？推进大数据应用的进程对现有技术框架、管理机制、评价体系又有哪些新的要求？

针对这一系列问题，《大数据与智慧社会》一书做出了系统的回答。本书从全局出发，对大数据的基本内涵进行了系统描述，概括了大数据的前世今生，揭示了其哲学本质；以技术为主线，深刻剖析了大数据的技术框架，预测了大数据的技术发展趋势；理论与实践相结合，形成大数据系统评价标准；选取大数据在生活、政务、交通、医疗、金融领域落地应用的实战案例，进行深入分析和解读，以期为我国的政府治理、经济发展、企业创新提供有效的指导和帮助。

本书适合政府决策者、企业管理者、IT实施者（CTO、CDO、技术人员等）以及高等院校相关专业的师生阅读。

◆主　　编　张克平　陈曙东
责任编辑　张国才
责任印制　焦志炜
◆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
三河市中晟雅豪印务有限公司印刷
◆开本：700×1000　1/16
印张：16.5　　2017 年 6 月第 1 版
字数：180 千字　　2017 年 6 月河北第 1 次印刷

定　价：59.00 元

读者服务热线：（010）81055656　印装质量热线：（010）81055316
反盗版热线：（010）81055315
广告经营许可证：京东工商广字第 8052 号

本书编委会

主　编　张克平　陈曙东

编　委（按姓氏笔画排序）

孔聪聪　张　亮　杜　蓉　李伟炜　倪　民

推荐序一

“大数据”的概念从问世到现在仅几年时间，却在全球引起了一次又一次热潮。这其中有两个重要因素在起推动作用。第一个是人类社会在发展过程中对信息的渴求。但是为什么直到5年前才“突然”出现大数据的概念？这就引出了第二个因素——以传感技术、互联网、移动智能终端为代表的一系列新信息技术，使信息的获取、利用和集聚在数量、作用和影响力等方面发生了突飞猛进的变化，成为推动历史进入新阶段的根本原因之一。这一切，深深植根于大数据的内在含义中。

认识大数据的本质，就是认识信息资源的本质。信息资源在人类发展的全部历程中扮演着极其重要的角色。语言这种特定的信息形式使人类摆脱了相互交流的障碍，个体的发现和能力可以在一个群体扩散利用，加快了人类进化的步伐。从结绳记事、岩画到文字的诞生，这些方式的出现使人类信息的交流摆脱了口口相传的时空约束，使经验和知识有了客观载体，可以跨越时空，显著加速了人类文明的进化。活字印刷、机器印刷的发明提升了信息生产和传播的效率及质量，推动了农业文明的快速发展。各类书籍、报刊、杂志的出版发行，各种藏书楼、图书馆的产生，为人类知识的汇集和利用提供了新的平台，加速了科技和文明的发展，推动了农业社会向工业社会迈进。

1970年，哈佛大学的奥汀格教授和他的研究队伍提出了信息、材料、能源是推动人类社会进步的三种基本资源的论断。40年后，全球经济发展的实

践证明了这一论断的正确性。把握大数据本质，就是要深刻理解“信息资源是推动人类社会进步的一个基础资源”这一观点。

传感技术、互联网、虚拟现实、大数据等一系列新信息技术的诞生和发展，使人类对信息的处理、传输、利用能力得到全面的提升，信息资源在社会发展中的作用日趋重要，推动着工业社会向信息社会迈进。2008 年金融危机后，一些欧洲国家又相继发生主权债务危机，与贸易保护主义、恐怖主义、南北不平衡等重大全球性问题纠缠在一起，全世界的理论家、战略家、政治家都在思考同一个问题，如何使人类社会摆脱危机，走向新的发展阶段。从 2011 年开始，新工业革命、第三次工业革命、互联网能源、工业革命 4.0、CPS、两化融合和两化深度融合、第二个机器时代等概念不断产生和发展。麻省理工学院（MIT）的埃里克·布林约尔松（Erik Brynjolfsson）和安德鲁·麦卡菲（Andrew McAfee）合著的《第二个机器时代》（*The Second Machine Age*）提出，我们将经历人类历史上两个最神奇的事件：创造真正的机器智能，以及全体人类通过一个共同的数字网络互联互通、从根本上改变地球经济的格局。第二个机器时代与第一个机器时代的不同之处在于智能化。第一个机器时代的机器取代并倍增了人类和动物的体力劳动，第二个机器时代的机器将取代并倍增我们的智慧。

这些现象和趋势的共同指向就是经济社会正在发生重大变革，这个变革的核心是信息技术体系和工业技术体系的融合，信息资源与能源、材料的协同，人类社会的经济和社会活动将以赛博 – 物理空间为依托。大卫·兰德斯指出：“工业革命是指生产方式上的深刻变革。即通过用机器代替人工、用非生物力代替人力和畜力，实现从手工工业向机器大生产的转变。”由于工业技术体系本身已经不足以从根本上继续提供推动历史转型的技术能力，人类需要构建赛博 – 物理空间，将信息、材料、能源三种资源利用综合起来，提升到一个新的水平。信息、材料、能源是推动经济社会发展的三驾马车，工业革命形成的生产力和信息革命形成的生产力推动人类社会进入一个新的

历史阶段，已成为历史发展的必然要求。

面向未来，抓住大数据技术带来的机遇，主动推进社会发展变革，要特别重视技术、产业和应用。从技术的角度来看，主要有两大问题：一是大数据每隔几年就会提升一个数量级，从这个角度看，如今的计算机处理体系不符合大数据处理的需求，所以要从芯片开始重构适合大数据发展的处理系统，要有新的芯片和新的处理结构，这是技术问题的一个制高点；第二个制高点是大数据的语义处理能力，也是智能技术的核心部分，这一技术将成为今后一个阶段信息技术创新的核心内容。从产业角度看，大数据产业大概可以分为两类：一类是“技术变成产业”，就像当年数据库管理系统变成了数据库公司，当真正的大数据处理芯片和计算架构形成时，还将会形成新的产业；另一类是各个企业、机构甚至个人（以后我们很多人）都可以变成大数据的拥有者和大数据产业的从业者。从应用的角度看，大数据最重要的意义在于，所有企业、机构和个人如何将大数据变成自身提升能力、提升竞争力、提升生活质量的来源，“以信息化培育新动力、以新动力推动新发展”，用“信息流引领技术流、资金流、人才流、物质流”，使其成为资源配置优化、全要素生产率提升、经济社会发展转型、经济结构调整的新动能。

《大数据与智慧社会》一书系统地介绍了什么是大数据和大数据技术框架，详细分析了以 Hadoop 和 Spark 为代表的典型技术，介绍了大数据在生活、政务、交通、医疗和金融领域的应用，为我们认识大数据、利用大数据提供了又一份精神食粮。更要指出的是，本书出自几位基层信息岗位的主管，着实难能可贵。

是以为序。

杨学山

工业和信息化部原副部长

北京大学教授

推荐序二

随着信息技术的飞速发展，“大数据”已被认为是继互联网、云计算、物联网之后又一大颠覆性的技术革命。通过对海量、动态、高增长、多元化数据的高速处理，大数据正在引发全球范围内的经济和商业变革，其应用涉及金融、交通、教育、医疗、制造、环保、零售、文化、娱乐等各行各业。大数据更带来了一场政府治理方式的变革，在提高公共决策能力的同时，改变着国家治理的架构和模式。可以毫不夸张地说：“大数据时代没有旁观者。”

研究发现，即使在缺乏精准的数据分析模型和算法的情况下，只要拥有足够多的数据，也能揭示事物的内在联系，引出重要的结论。这给计算衍生学科带来了里程碑式的启示：大数据本身可以保证数据分析结果的有效性。因此，大数据被誉为新的生产力。在大数据时代，大数据生产力将会推动生产关系和社会的发展，创造无穷无尽的价值，给人类思维的发展带来变革。

大数据使我们至少拥有了四个方面的核心能力。

首先是海纳百川的数据融合能力。通过将各种数量庞大、分布广泛、形式多样、变化迅速的异构数据资源汇聚、融合在一起，资源数量和质量的巨大提升引发资源价值的巨大提升，使大数据成为现代社会的巨大财富。

其次是基于大数据的科学研究能力。基于这种能力的科研范式有别于传统的实验归纳、模型推演、仿真模拟等范式，被称为数据密集型科学发现，

即第四范式。应当指出，运用这种能力，当数据达到一定量时，传统计算架构已经不再适用，云计算应运而生。实际上，大数据与云计算相辅相成，两者之间互相推动与促进。没有云计算能力，大数据的价值就无法被挖掘出来；没有大数据，云计算也就没有用武之地。

再次是明察秋毫的洞见力。透过数据发现隐藏在事物表面下的本质规律，发现事物之间的关联，揭示事物发展的规律，便于人类发现新的原理或者产生新的科学创造。这种运用第四范式获得的科学发现，既不像理论和模拟那样在一定程度上告诉我们“为什么”，也不像实验那样明确地告诉我们“是什么”，只能告诉我们“与什么相关”。第四范式强调了以大数据为基础的数据密集型研讨方法，这种方法将会在越来越多领域的研讨中发挥至关重要的甚至是决定性的作用。

最后是高瞻远瞩的预测和决策能力。在过去的商业决策中，管理者凭借自身的经验和对行业的敏感来决定企业的发展方向和方式，这种决策有时候仅仅参考一些模糊的数据和建议。而大数据和大数据分析工具的出现，让人们找到了一条新的科学决策之路。以数据为依据，立足事实，既观全局，又见未来。

大数据正是因为能赋予我们这四种核心能力，才受到越来越多的关注。我国已经将大数据提升到国家战略高度，在“十三五”规划纲要中指出：实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新；深化大数据在各行业的创新应用，探索与传统产业协同发展新业态、新模式，加快完善大数据产业链；加快海量数据采集、存储、清洗、分析发掘、可视化、安全与隐私保护等领域关键技术攻关。习近平总书记强调，机会稍纵即逝，抓住了就是机遇，抓不住就是挑战。我国发展大数据有非常好的机遇，同时我们也应该清醒地认识到，我国大数据产业刚刚起步，从技术上、观念上、法律上等多个层面都需要变革，才能满足大数据的发展需要。

由张克平局长、陈曙东研究员主编的《大数据与智慧社会》一书，顺应时势，系统地从大数据起源、大数据哲学本质、大数据技术框架、大数据应用案例等不同的角度为读者展示了一幅大数据技术图谱。该书首先概述了大数据的哲学本质、技术现状和发展趋势，然后详述了大数据的技术框架、大数据存储和大数据处理技术。“科学家要多做实践中的研究”，当前，大数据应用处于起步期，产业生态处于酝酿期，必须在实际应用中发挥大数据技术的作用，推动大数据产业的发展。因此，作者们又详述了大数据在生活、政务、交通、医疗和金融等相关领域的应用实战，为读者使用大数据指出了一条探索之路。

我相信本书将受到关注大数据的“政产学研用”各界的欢迎，为大数据在中国的发展助一臂之力。

倪光南

中国工程院院士

目 录

第 1 章 大数据概述 / 1

1.1 什么是大数据 / 2

1.1.1 大数据的定义和特征 / 2

1.1.2 大数据的发展历程 / 6

1.1.3 大数据的来源 / 11

1.2 大数据的哲学本质 / 12

1.2.1 大数据与世界观 / 13

1.2.2 大数据与认识论 / 14

1.2.3 大数据与方法论 / 15

1.2.4 大数据与价值观 / 18

1.3 大数据技术框架 / 19

1.3.1 大数据处理系统综述 / 19

1.3.2 大数据平台基础 / 20

1.3.3 大数据存储系统 / 22

1.3.4 大数据计算模型 / 23

1.4 大数据发展趋势 / 26

1.4.1 大数据的技术发展趋势 / 26

1.4.2 大数据的应用发展趋势 / 30

第 2 章　大数据的云计算基础 /　33

2.1　虚拟化技术 /　34

2.1.1　虚拟化的概念 /　34

2.1.2　虚拟化技术分类 /　35

2.1.3　虚拟化解决方案 /　36

2.1.4　虚拟化技术与大数据 /　39

2.2　OpenStack 技术 /　40

2.2.1　OpenStack 概述 /　40

2.2.2　OpenStack 历史 /　41

2.2.3　OpenStack 系统架构 /　41

2.2.4　OpenStack 的优势和劣势 /　44

2.2.5　虚拟化与 OpenStack 技术比较 /　46

2.3　IaaS 平台建设 /　47

2.3.1　IaaS 平台介绍 /　47

2.3.2　IaaS 云平台的种类 /　49

2.3.3　IaaS 平台设计 /　51

2.3.4　IaaS 平台解决方案 /　53

2.3.5　IaaS 平台搭建 /　55

第 3 章　Hadoop 基础组件 /　57

3.1　Hadoop 概述 /　58

3.1.1　Hadoop 简介 /　58

3.1.2　Hadoop 系统架构 /　59

3.1.3　Hadoop 的优势与不足 /　60

3.1.4　Hadoop 的适用场景 /　63

3.1.5　Hadoop 的商业模式 /　64

3.2 Hadoop 分布式文件系统 HDFS/ 65
3.2.1 HDFS 的设计目标/ 65
3.2.2 HDFS 的基本架构/ 67
3.2.3 HDFS 的特点/ 68
3.2.4 HDFS 的优势与缺点/ 70
3.3 Hadoop 分布式计算框架 MapReduce/ 72
3.3.1 MapReduce 简介/ 72
3.3.2 MapReduce 的运行流程/ 73
3.3.3 MapReduce 与 DataFlow 比较/ 75
3.4 Hadoop 统一资源管理框架 YARN/ 76
3.4.1 YARN 架构简介/ 76
3.4.2 YARN 架构框架/ 77
3.4.3 YARN 与旧 MapReduce 框架对比/ 79
3.4.4 YARN 与 Mesos 框架对比/ 79
3.5 Hadoop 分布式集群管理系统 ZooKeeper/ 81
3.5.1 ZooKeeper 简介/ 81
3.5.2 ZooKeeper 总体架构/ 82
3.5.3 ZooKeeper 的运行模式/ 84
3.5.4 ZooKeeper 的设计要点/ 85
3.5.5 ZooKeeper 的使用/ 87

第 4 章 Hadoop 其他常用组件/ 89

4.1 Hadoop 数据仓库工具 Hive/ 90
4.1.1 Hive 简介/ 90
4.1.2 Hive 架构设计/ 91
4.1.3 Hive 部署模式/ 92

4.1.4 Hive 与关系型数据库比较 / 94

4.2 Hadoop 分布式数据库 HBase/ 97

4.2.1 HBase 简介 / 97

4.2.2 HBase 体系架构 / 97

4.2.3 HBase 性能分析 / 99

4.2.4 HBase 容错机制 / 101

4.3 Hadoop 实时流式处理引擎 Storm-YARN/ 102

4.3.1 流式处理概述 / 102

4.3.2 Storm 简介 / 103

4.3.3 Storm 架构 / 105

4.3.4 Storm 与 Spark Streaming 比较 / 106

4.4 Hadoop 交互式查询引擎 Impala/ 108

4.4.1 Impala 简介 / 108

4.4.2 Impala 架构分析 / 109

4.4.3 Impala 与 Hive 比较 / 110

第 5 章 Spark 内存计算框架 / 113

5.1 内存计算与 Spark/ 114

5.1.1 内存计算概念 / 114

5.1.2 内存计算分类 / 116

5.1.3 Spark 与内存数据处理系统 / 118

5.2 Spark 概述 / 119

5.2.1 Spark 架构 / 119

5.2.2 Spark 的 RDD 模型 / 121

5.2.3 Spark 与 Hadoop 的性能对比 / 121

5.3 Spark 核心组件介绍 / 122
5.3.1 Spark SQL/ 122
5.3.2 Spark MLlib/ 123
5.3.3 Spark GraphX/ 123
5.3.4 Spark Streaming/ 124
5.4 Spark 集群管理 / 125
5.4.1 Spark 部署方式 / 125
5.4.2 Spark 资源调度 / 126
5.4.3 Spark 任务调度 / 127

第 6 章 大数据可视化技术 / 129

6.1 数据可视化的基本概念 / 131
6.2 数据可视化的发展趋势 / 132
6.3 数据可视化应用与设计 / 135

第 7 章 数据挖掘技术 / 139

7.1 什么是数据挖掘 / 140
7.2 数据挖掘的流程 / 142
7.3 数据挖掘典型算法 / 143
7.4 数据挖掘与大数据 / 151

第 8 章 大数据系统评价标准 / 153

8.1 大数据系统评价概述 / 154
8.1.1 信息时代的“云大物移” / 154
8.1.2 大数据项目失败的常见原因 / 155

8.2 评价指标选取原则 / 157
8.3 大数据系统评价标准 / 159
8.3.1 通用评价要素 / 159
8.3.2 专有评价要素 / 160
8.4 大数据系统定位 / 161
8.4.1 与企业战略相匹配 / 161
8.4.2 与企业架构相匹配 / 162
8.4.3 与企业需求相匹配 / 162
8.5 大数据价值评估模型 / 164
8.6 大数据质量评价 / 165
8.6.1 数据流程视角 / 165
8.6.2 数据技术视角 / 167
8.6.3 数据管理视角 / 168
8.7 大数据安全评价 / 169

第 9 章 大数据在生活中的应用 / 173

9.1 食：食品安全 / 174
案例：阿里巴巴大数据协助食品安全风险控制 / 175
9.2 住：智能家居 / 176
案例：无锡市智能家居 / 177
9.3 行：智能交通 / 177
案例：深圳市智能综合交通运行指挥中心 / 180
9.4 游：智慧旅游 / 180
案例：无锡市智慧旅游立体化营销体系 / 181

9.5 购：电商营销 / 183

案例：京东大数据营销 / 185

第 10 章 大数据在政务领域的应用 / 187

10.1 条块分割拖累政务发展 / 188

10.2 数据统筹助力决策参考 / 190

案例：佛山市南海区数据统筹 / 192

10.3 服务整合创新社会管理 / 193

案例：无锡市智慧城管系统 / 194

10.4 资源整合强化公共服务 / 196

案例：无锡市政务服务平台 / 197

10.5 数据公开辅助政府监督 / 200

案例：上海市利用大数据实现市场监管 / 201

第 11 章 大数据在交通领域的应用 / 203

11.1 频繁拥堵造就城市顽疾 / 204

11.2 客流分析改进公交线路设计 / 205

案例：北京市大数据路线优化 / 206

11.3 多源数据辅助交通调查 / 206

案例：上海市综合交通特征分析 / 208

11.4 整合信息优化资源配置 / 210

案例：无锡市智慧交通信息工程 / 211

11.5 智能数据释难最后一公里 / 213

案例：共享单车便捷出行 / 214