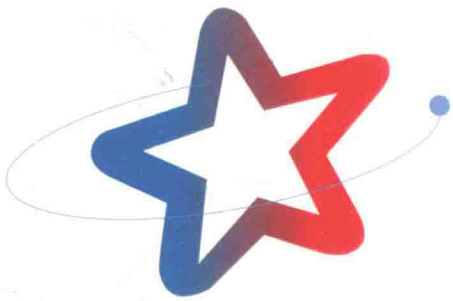
- 以云计算与大数据融合的视角阐述了云计算环境下Spark大数据处理与相应的算法实现
- 结合经典案例，详解云计算环境下Spark大数据处理生态圈，包括系统结构、大数据存储、批处理、流计算、交互式数据分析、并行机器学习架构与算法等技术
- 掌握云计算环境下Spark大数据处理的架构搭建和算法实现过程等关键技术，扩展大数据从业人员的理论与实践能力

大数据 技术



Lightning-fast Cluster Computing

云计算环境下
Spark大数据处理技术与实践

邓立国　佟强　著

清華大學出版社

云计算环境下
Spark大数据处理技术与实践

邓立国　佟强　著

清華大學出版社
北京

内 容 简 介

本书围绕互联网重大的技术革命：云计算、大数据进行阐述。云计算环境下大数据处理构建是国民经济发展的信息基础设施，发展自主的云计算核心技术，拥有自己的信息基础设施，当前正处于重要的机遇期。

本书重点在大数据与云计算的融合，给出了大数据与云计算的一些基本概念，并以 Spark 为开发工具，全面讲述云环境下的 Spark 大数据技术部署与典型案例算法实现，最后介绍了国内经典 Spark 大数据与云计算融合的架构与算法。

本书适合云计算环境下 Spark 大数据技术人员、Spark MLlib 机器学习技术人员，也适合高等院校和培训机构相关专业的师生教学参考。

图书在版编目（CIP）数据

云计算环境下 Spark 大数据处理技术与实践 / 邓立国，佟强著. — 北京：清华大学出版社，2017
ISBN 978-7-302-47971-0

I. ①云… II. ①邓… ②佟… III. ①数据处理软件 IV. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 207679 号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张**：22.25　　**字　　数**：570 千字
版　　次：2017 年 9 月第 1 版　　**印　　次**：2017 年 9 月第1 次印刷
印　　数：1～3500
定　　价：69.00 元

产品编号：075719-01

前　言

麦肯锡全球研究所给出的大数据定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

大数据技术的战略意义不在于获取了庞大的数据，而在于对这些特定领域的数据进行处理分析。换而言之，关键是把这些巨大的数据实现盈利式的加工，提供效率，具有增值的处理模式。

本书背景

大数据像飓风一样席卷而来，改变着信息时代的数据处理方式。产业经营方式经历着革命性的变革，大数据与云计算的融合改变着数据处理流程和模式，对互联网、信息经济发展提出了新的方向和扩展空间。应用驱动技术发展产生的数据越多，可供分析的数据越多，越能推动研发和出现更先进的用来分析数据的工具和方法。

国家对互联网、信息经济的发展提出了方向，明确说要拓展发展新的空间，实施网络强国战略，实施“互联网+”行动计划，发展分享经济，实施国家大数据战略，将网络强国战略作为新的一个创新的重要支撑。

本书内容

本书围绕互联网重大的技术革命：云计算、大数据（未来世界新一代信息技术的关键和核心）进行阐述。云计算环境下大数据处理构建是国民经济发展的信息基础设施，发展自主的云计算核心技术，拥有自己的信息基础设施，当前正处于重要的发展机遇期。本书重点在大数据与云计算的融合，给出了大数据与云计算的一些基本概念的同时，以 Spark 为开发工具，全面讲述云环境下的大数据技术部署与典型案例算法实现，最后介绍了国内经典 Spark 大数据与云计算融合的架构与算法。

本书目的

3 年前就开始着手准备写关于大数据和云计算融合的相关技术方面的书，由于书中的算法需要模拟验证，所以交稿拖延了很长时间。目前这方面的书还不系统，还没有全面融合两

者技术的书出现，也是笔者想写这本书的初衷。随着岁月侵蚀，白发杂生，大数据技术发展也日新月异。

得益于国内 IT 企业的后发制人战略，目前国内的 IT 公司在大数据应用方面已经迎头赶上了国际巨头，在云大数据技术方面的研发和技术突破经历了大幅的跨越发展。当今世界迎来大数据时代，工欲善其事，必先利其器，在大数据和云计算的规则制定和新技术研发上还需努力，这方面还需要加大研发与突破。

致谢

感谢家人给我的全身心的支持与关爱，没有你们的宽容与支持即使是 10 年也没法完成这本书。由于撰写时间紧迫，夜晚孤灯，每晚多想陪着妻子月夜树影婆娑，多想在闺女的校门口等待闺女背着书包颠颠地跑来。最后感谢单位给予的大力支持与帮助。

著者

2017 年 8 月

目 录

第 1 章

◂ 大数据处理概述 ▸

大数据是当今一个最热门的话题，我们每一个人都无法置身其外。就像几年前出现的云计算一样，大数据已经引起市场的广泛关注；同样，企业迫切需要对大数据下定义。大数据缺少一个标准且普及性的定义，至少不像 NIST 对云的定义那样，能被人们广泛接受。调研公司 IDC 的定义可能比较容易被人们所接受。它对大数据的定义是：一种新一代的技术和架构，具备高效率的捕捉、发现和分析能力，能够经济地从类型繁杂、数量庞大的数据中挖掘出价值。

1.1 大数据处理技术概述

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。《Nature》和《Science》等相继出版专刊专门探讨大数据带来的机遇和挑战。著名管理咨询公司麦肯锡称："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。大数据已成为社会各界关注的新焦点，"大数据时代"已然来临[1]。

"大数据"是一个体量特别大、数据类别特别大的数据集，并且这样的数据集无法用传统数据库工具对其内容进行抓取、管理和处理。

百度知道大数据（bigdata）的定义，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理，并整理成为帮助企业经营决策更积极目的的资讯。大数据的 5V 特点：Volume、Velocity、Variety、Veracity、Value。

1.1.1 什么是大数据

"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的海量、高增长率和多样化的信息资产。从数据的类别上看，"大数据"指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统

处理方法的数据集。亚马逊网络服务（AWS）大数据科学家 JohnRauser 提到一个简单的定义：大数据就是任何超过了一台计算机处理能力的庞大数据量。其研发小组对大数据的定义："大数据是最大的、最时髦的技术，当这种现象出现时，定义就变得很混乱。"学者 Kelly 说："大数据是可能不包含所有的信息，但我觉得大部分是正确的。对大数据的一部分认知在于，它是如此之大，分析它需要多个工作负载，这是 AWS 的定义。当你的技术达到极限时，也就是数据的极限"。大数据不是关于如何定义，最重要的是如何使用。最大的挑战在于哪些技术能更好地使用数据以及大数据的应用情况如何。这与传统的数据库相比，开源的大数据分析工具如 Hadoop 的崛起，这些非结构化的数据服务的价值在哪里。

相较于传统的数据，人们将大数据的特征总结为 5 个 V，即体量大（Volume）、速度快（Velocity）、模态多（Variety）、难辨识（Veracity）和价值大（Value）。"大数据"首先是指数据体量(volumes)大，指代大型数据集，一般在 10TB 规模左右，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了 PB 级的数据量；其次是指数据类别（Variety）多，数据来自多种数据源，数据种类和格式日渐丰富，已冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据；接着是数据处理速度（Velocity）快，在数据量非常庞大的情况下，也能够做到数据的实时处理；还有一个特点是指数据真实性（Veracity）高，随着社交数据、企业内容、交易与应用数据等新数据源的兴趣，传统数据源的局限被打破，企业愈发需要有效的信息之力以确保其真实性及安全性。但大数据的主要难点并不在于数据量大，因为通过对计算机系统的扩展可以在一定程度上缓解数据量大带来的挑战。其实，大数据真正难以对付的挑战来自于数据类型多样（Variety）、要求及时响应（Velocity）和数据的不确定性（Veracity）。因为数据类型多样使得一个应用往往既要处理结构化数据，同时还要处理文本、视频、语音等非结构化数据，这对现有数据库系统来说难以应付；在快速响应方面，在许多应用中时间就是利益；在不确定性方面，数据真伪难辨是大数据应用的最大挑战。追求高数据质量是对大数据的一项重要要求，最好的数据清理方法也难以消除某些数据固有的不可预测性。

1.1.2 大数据来源

当今世界，大数据无处不在，它影响到了我们的工作、生活和学习，并将继续施加更大的影响。大数据用于描述这样的数据组，其规模超出了日常软件在可容忍期限内获取、管理和加工数据的能力。一些网络技术领先的公司持续地投资于昂贵的大数据技术，成效显著。大数据使得创新型公司变成了经营新方法的率先接受者，经营更为成功。通过大数据的分析挖掘，公司可以发现新的经营模式，对工艺加以改进。例如，在获悉消费者行为后，可以将发现用于某些改变，如降低成本或增加销售，就会产生价值。在任意大的数据组中应用统计方法可以发现有用信息，将这些信息商业化即可获益。

当今大数据的来源除了专业研究机构产生大量的数据外（CERN 的离子对撞机每秒运行产生的数据高达 40TB），相对于企业，大数据的数据来源主要有两个部分：一部分来自于企业内部自身信息系统中产生的运营数据，这些数据大多是标准化、结构化的；传统的商业智能系统中所用到的数据基本上属于这部分；另一部分则来自于外部，包括广泛存在于社交网络、物联网、电子商务等之中的非结构化数据。

与企业经营相关的大数据可以划分为 4 个来源：

- 越来越多的机器配备了连续测量和报告运行情况的装置。几年前，跟踪遥测发动机运行仅限于价值数百万美元的航天飞机。现在，汽车生产商在车辆中配置了监视器，连续提供车辆机械系统整体运行情况。一旦数据可得，公司将千方百计从中渔利。这些机器传感数据属于大数据的范围。
- 计算机产生的数据可能包含着关于因特网和其他使用者行动和行为的有趣信息，从而提供了对他们的愿望和需求潜在的有用认识。
- 使用者自身产生的数据/信息。人们通过电邮、短信、微博等产生的文本信息。
- 至今最大的数据是音频、视频和符号数据。这些数据结构松散、数量巨大，很难从中挖掘有意义的结论和有用的信息。

由于来源、类型不同的数据透视的是同一个事物的不同方面，以消费客户为例，消费记录信息能透视客户的消费能力、消费频率、消费兴趣点等，渠道信息能透视客户的渠道偏好，以及消费支付信息与支付渠道的关联情况等。

大型以 Internet 为核心的公司，如 Amazon、Google、eBay、Twitter 和 Facebook 正使用后三类海量信息认识消费行为、预测特定需求和整体趋势。第一类数据可能产生较少的业务，但可以推动某些经营模式实质变革。例如，汽车传感数据用于评价司机行为会推动汽车保险业的深刻变革。因此，大数据分析意味着企业能够从不同来源的数据中获得新的洞察力，并将其与企业业务体系的各个细节相结合，以便助力企业在市场拓展和产品创新上突破。

1.1.3　大数据应用价值

“大数据”的概念远不止大量的数据（TB）和处理大量数据的技术，而是涵盖了人们在大规模数据的基础上可以做的事情，而这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力。

根据麦肯锡全球研究所的分析，利用大数据在各行各业能产生显著的财务价值。美国健康护理利用大数据每年产出 3000 亿美元，年劳动生产率提高 0.7%；欧洲公共管理每年价值 2500 亿欧元，年劳动生产率提高 0.5%；全球个人定位数据服务提供商收益 1000 多亿美元，为终端用户提供高达 7000 亿美元的价值；美国零售业净收益可增长 6%，年劳动生产率提高 0.5~1%；制造业可节省 50%的产品开发和装配成本，营运资本下降 7%。

大数据改变了所有行业全部公司的经营方式。从对市场的理解到如何挖掘经营信息，大数据能洞察每项转变。一个致力于收集和分析大数据的行业业已形成，对现有公司产生了深刻影响。据有关调查，有 10%的公司认为在过去的 5 年中，大数据彻底改变了它们的经营方式。46%的公司认同大数据是其决策的一项重要支持因素。

大数据应用在经历了喊口号、布局深耕之后，开始显现出巨大的商业价值，触角延伸到国防、市政、金融、教育、医疗、体育、汽车、影视、智能硬件、社交网络等各个层面。据 IDC 数据显示，目前大数据形成的市场规模达到约 51 亿美元，到 2017 年，这一数字将会增

长到 530 亿美元。

近日，国内大数据精准营销平台亿玛公司联合中关村大数据产业联盟举办了 2015 亿玛智慧峰会，智能穿戴设备、智能汽车、互联网金融、大数据精准营销、智能家居、大数据医疗与健康、移动智能大数据应用等大数据应用代表企业共同围绕“大数据，智未来”的主题，讨论了“如何依托大数据提供更符合社会需求的产品和服务”“大数据如何为精准营销提供强大的驱动力”等业界最关注的热点话题。

截至目前，大数据应用的商业价值已经在互联网金融、智能可穿戴设备、人工智能、智慧城市、精准营销等多个领域体现。其中，在互联网金融领域，通过分析大量的网络交易及行为数据，可对用户进行信用评估，从而帮助互联网金融企业对用户还款意愿及还款能力得出结论，继而为用户提供快速授信及现金分期服务。

而智能可穿戴设备“真正”的主线产品是由云端大数据引出的软件与服务。将来在新的模式里，硬件和服务都要具备快速迭代的能力，同时集成更多的传感器，数据来源将越来越丰富，而数据分析服务将利用更多种类的数据来交叉分析，无须用户干预而通过数据智能化学习就能把人一天重要的生理活动描绘出来，通过数据把行为量化。

大数据一个主要特性是复杂，这就意味着它的多元性。大数据不再是结构化数据，因此针对数据分析的模型和理论都必须重新构建，甚至分析大数据行为特征所依托的软硬件都必须进行变革。

1.1.4 大数据技术特点和研究内容

根据国际数据公司（IDC）的测算，2011 年数字世界将产生 1800EB 的数据，2012 年会增长 40%，达到 2500EB。截止 2020 年，会达到 35000EB，似乎没有足够的磁盘空间存储。就传统 IT 企业来看，其结构化和非结构化的数据增长也是惊人的。2005 年企业存储的结构化数据为 4EB，到 2015 年将增至 29EB，年复合增长率逾 20%。非结构化数据发展更猛。2005 年为 22EB，2015 年将增至 1600EB，年复合增长率约 60%，远远快于摩尔定律。

大数据具有 5 个主要的技术特点，人们将其总结为 5V 特征：

- Volume（大体量）：可从数百 TB 到数十数百 PB，甚至 EB 的规模。
- Variety（多样性）：大数据包括各种格式和形态的数据。
- Velocity（时效性）：很多大数据需要在一定的时间限度下得到及时处理。
- Veracity（准确性）：处理的结果要保证一定的准确性。
- Value（大价值）：大数据包含很多深度的价值，大数据分析挖掘和利用将带来巨大的商业价值。

传统的数据库系统主要面向结构化数据的存储和处理，但现实世界中的大数据具有各种不同的格式和形态。据统计，现实世界中 80%以上的数据都是文本和媒体等非结构化数据；同时，大数据还具有很多不同的计算特征。我们可以从多个角度分类大数据的类型和计算特征：

- 从数据结构特征角度看，大数据可分为结构化与非结构化/半结构化数据。
- 从数据获取处理方式看，大数据可分为批处理与流式计算方式。

- 从数据处理类型看，大数据处理可分为传统的查询分析计算和复杂数据挖掘计算。
- 从大数据处理响应性能看，大数据处理可分为实时/准实时与非实时计算，或者是联机计算与线下计算。前述的流式计算通常属于实时计算，此外查询分析类计算通常也要求具有高响应性能，因而也可以归为实时或准实时计算。而批处理计算和复杂数据挖掘计算通常属于非实时或线下计算。
- 从数据关系角度看，大数据可分为简单关系数据（如 Web 日志）和复杂关系数据（如社会网络等具有复杂数据关系的图计算）。
- 从迭代计算角度看，现实世界的数据处理中有很多计算问题需要大量的迭代计算，诸如一些机器学习等复杂的计算任务会需要大量的迭代计算，为此需要提供具有高效的迭代计算能力的大数据处理和计算方法。
- 从并行计算体系结构特征角度看，由于需要支持大规模数据的存储和计算，因此目前绝大多数大数据处理都使用基于集群的分布式存储与并行计算体系结构和硬件平台。MapReduce 是最为成功的分布式存储和并行计算模式。然而，基于磁盘的数据存储和计算模式使 MapReduce 难以实现高响应性能。为此人们从分布计算体系结构层面上又提出了内存计算的概念和技术方法。

大数据的研究与分析应用的意义和价值十分重大，带来巨大的挑战、技术创新与商机。维克托·迈克-舍恩伯格在《大数据时代》中列举了大量翔实的案例，指出了大数据的发展思路，大数据开启了生活、工作和创新的思维模式，影响了我们的经济、政治、科技和社会发展的各个领域，由于大数据应用行业需求的日益增长，大数据的并行计算技术越来越多地渗透到每个涉及大规模数据和复杂计算的应用领域。因此，以大数据处理为中心的技术变革，直接刺激计算机体系结构、操作系统、数据库、编译技术、程序设计、软件工程、多媒体信息处理、人工智能以及其他计算机应用技术，融合传统技术产生很多相应的新的研究课题与热点。

1.1.5　大数据计算与系统

大数据中蕴含的宝贵价值成为人们存储和处理大数据的驱动力。维克托·迈克-舍恩伯格在《大数据时代》一书中指出了大数据时代处理数据理念的三大转变，即要全体不要抽样，要效率不要绝对精确，要相关不要因果。因此，海量数据的处理对于当前存在的技术来说是一种极大的挑战。目前，人们对大数据的处理形式主要是对静态数据的批量处理、对在线数据的实时处理，以及对图数据的综合处理。其中，在线数据的实时处理又包括对流式数据的处理和实时交互计算两种。

MapReduce 计算模式的出现有力地推动了大数据技术和应用的发展，使其成为目前大数据处理最成功的主流大数据计算模式。然而，现实世界中的大数据处理问题复杂多样，难以有一种单一的计算模式能涵盖所有不同的大数据计算需求。研究和实际应用中发现，由于 MapReduce 主要适合于进行大数据线下批处理，在面向低延迟和具有复杂数据关系和复杂计算的大数据问题时有很大的不适应性。因此，近几年来学术界和业界在不断研究并推出多种不同的大数据计算模式。所谓大数据计算模式，是指根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象（Abstraction）和模型