大数据创新人才培养系列

# 大数据技术基础

## 基于 Hadoop 与 Spark

BIG DATA TECHNOLOGY:
HADOOP AND SPARK

◎ 罗福强 李瑶 陈虹君 编著

关注大数据处理系统**三大要求**——“**存储**”、“**计算**”与“**容错**”
将 **Hadoop** 和 **Spark** 组合起来进行剖析，呈现完整**大数据技术方案**
提供**大量案例**，全部通过 **JDK 1.8** 调试，并给出**源码**和**运行效果**

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

大数据创新人才培养系列

# 大数据技术基础

## 基于 Hadoop 与 Spark

BIG DATA TECHNOLOGY:
HADOOP AND SPARK

◎ 罗福强 李瑶 陈虹君 编著



人民邮电出版社
北京

**图书在版编目（CIP）数据**

大数据技术基础 ：基于Hadoop与Spark / 罗福强，李瑶，陈虹君编著. -- 北京 ：人民邮电出版社，2017.6
（大数据创新人才培养系列）
ISBN 978-7-115-45410-2

Ⅰ. ①大… Ⅱ. ①罗… ②李… ③陈… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第073225号

## 内 容 提 要

大数据处理系统必须关注存储、计算与容错问题。本书以此为起点，系统地介绍了 Hadoop 和 Spark 技术原理以及应用编程方法。本书主要内容包括：大数据概述、Hadoop 和 Spark 原理、HDFS 与 HDFS API 编程与应用、YARN 与 MapReduce API 编程与应用、Spark Streaming 和 Spark SQL 编程等。

本书旨在帮助初学者迅速掌握 Hadoop 和 Spark 原理及其应用，提升读者的大数据应用与开发能力，同时本书极强的系统性、可操作性以及大量精心设计的案例对于有一定基础的中高级读者有非常好的参考价值。

◆ 编　　著　罗福强　李　瑶　陈虹君
责任编辑　刘　博
责任印制　杨林杰

◆ 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
大厂聚鑫印刷有限责任公司印刷

◆ 开本：787×1092　1/16
印张：18.75　　2017 年 6 月第 1 版
字数：494 千字　　2017 年 6 月河北第 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

随着互联网应用的井喷式发展，人类进入了信息大爆炸和海量信息处理时代，大数据成为近几年来国内外最热门的话题之一。但是，究竟什么是大数据，大数据有什么意义，什么是大数据技术，有哪些大数据技术，大数据技术发展状况如何，如果升级现有的信息系统又该从何处入手，应该采用哪一种大数据解决方案……所有这些问题都令每一个关注大数据话题的人感到困惑。

四川大学图像信息研究所创始人、四川大学锦城学院电子信息学院院长陶德元教授很早以前就敏锐地察觉到，要解决这些问题就必须大量培养掌握大数据技术的人才。为此，当大家还在争论要不要在本科开设大数据专业时，陶教授于 2013 年已率先在四川大学锦城学院的高年级本科生中引入了大数据课程，依托物联网工程专业在本科生层次建立了大数据专业方向的人才培养试点。没有他山之石可采，只能摸着石头过河，不断地总结经验和教训，一路走来甚为艰辛。我们深知，在众多国内大数据专家、学者面前，我们的水平仍显得不足，因此深怕因技术水平浅薄，所汇集的文字材料漏洞百出而贻笑大方。

如今，从四川大学锦城学院已经走出了 3 届初步掌握大数据技术的学生，还有 3 个年级的在校生正在接受大数据技术的洗礼。经过 3 年的摸索，大数据专业方向的相关课程和内容已逐步稳定下来。特别是连续 3 届学生的成功就业，让我们有了写作本书的勇气。

Spark 刚出道时，有人预言 Hadoop 终将被 Spark 所淘汰，从而走向消亡。事实果却非如此。实际上，自 2007 年被推出以来，Hadoop 一直在不断发展和演化，如今已经发展成一个由 60 多个技术组件组成的庞大生态系统，核心组件包括 HDFS、MapReduce、YARN、HBase、Hive、Spark 等。其中，HDFS 用于实现数据的分布式存储。MapReduce 用于实现数据的分布式计算，其计算任务的输入和输出是依靠文件读写操作来实现的，由于大量的磁盘 I/O 操作会影响系统性能，因此它常常满足不了实时处理数据的需求。Spark 改进了 MapReduce 的计算模式，它在减少磁盘 I/O 操作的同时把数据缓存在内存中，从而提升数据的处理性能，理论上要比 MapReduce 的速度快 100 倍。但是，Spark 没有提供分布式存储解决方案。可见，Hadoop 和 Spark 可以做到优势互补。也正是基于这一点，本书将 Hadoop 和 Spark 技术组合起来进行剖析，目的也是希望能够向读者呈现一个比较完整的大数据技术解决方案。

本书以大数据处理系统所关注的三大要求——“存储”“计算”与“容错”为起点，全面介绍了它们所代表的大数据技术的原理以及应用编程方法。全书分为 4 个部分，共 10 章。第 1 部分只包括第 1 章，主要介绍大数据概念与特征、大数据技术的发展、大数据存储与计算模式、大数据的典型应用等；第 2 部分包括第 2~4 章，重点介绍了 Hadoop 平台的部署方法、HDFS 的分布式存储原理及其 Shell 操作、HDFS API 编程与应用；第 3 部分包括第 5~7 章，主要介绍了 MapReduce v2( YARN )

的分布式计算原理、MapReduce API 基本编程与高级编程方法及应用；第 4 部分为最后 3 章，主要介绍了 Spark 组成和原理、Spark Streaming 和 Spark SQL 编程与应用方法等。

本书具有 4 个鲜明特点：第一，重点突出，避免因面面俱到而缺乏技术深度；第二，内容结构完整，文字流畅，循序渐进，符合人的认知规律；第三，案例丰富，可操作性强，有助于快速培养大数据开发能力；第四，全书配备了丰富的、符合初学者习惯的思考和实践任务。

本书旨在帮助大数据技术的初学者快速掌握 Hadoop 和 Spark 原理及其应用，提升读者的大数据应用与开发能力。同时本书极强的系统性、可操作性以及大量精心设计的案例对于有一定基础的中高级读者有非常好的参考价值。

参与本书编写工作的有四川大学锦城学院的罗福强、李瑶、陈虹君、熊永福等老师。其中，熊永福编写了第 1 章，罗福强编写了第 2～4 章，李瑶编写了第 5～7 章，陈虹君编写了第 8、9、10 章。本书由罗福强负责主编工作。本书获得了科研项目资助，也获得了四川大学锦城学院和人民邮电出版社各级领导的重视与支持，特别得到了电子信息学院陶德元教授的支持和大量帮助。在此，我们对支持本书编写并提供过帮助的所有人表示诚挚的感谢！

由于作者水平有限，本书虽经多次校对，仍难免有疏漏之处，我们殷切地期望读者提出中肯的意见，联系方式：LFQ501@sohu.com。

编　者

2017 年 1 月

# 目 录

# 第 1 章 大数据技术概述

**本章目标：**

- 了解大数据的发展过程以及大数据对国内外各行各业的影响。
- 掌握大数据的概念及其特征。
- 了解大数据的来源，理解大数据在技术、安全等方面面临的挑战和研究大数据的意义。
- 掌握大数据的存储与计算模式的相关概念，了解其中的关键技术及基本思想。
- 了解大数据的典型应用场景，学会用创新性思维来看待大数据。
- 了解 Hadoop 的发展过程和优势。
- 熟悉 Hadoop 的生态系统以及其中的基本概念。
- 了解 Hadoop 的版本发行状况。

**本章重点和难点：**

- 大数据的概念与特征。
- 大数据的存储与计算模式及其相关技术。
- Hadoop 的生态系统及其基本概念。

我们生活在一个数据大爆炸的时代，很难估算全球电子设备中存储的数据总共有多少。根据中国最大的企业级 IT 网站 ZDNET（至顶网）的年度技术报告——《数据中心 2013：硬件重构与软件定义》，2013 年中国产生的数据总量超过 0.8ZB（相当于 8 亿 TB），2 倍于 2012 年中国的数据总量，相当于 2009 年全球的数据总量。该报告预计，到 2020 年，中国产生的数据总量将是 2013 年的 10 倍，超过 8.5ZB。本章将深入介绍大数据的发展、概念、特征、典型应用，以及 Hadoop 大数据平台的发展、基本概念及体系结构。

## 1.1 大数据技术的发展背景

大数据，即 Big Data，一个如今人们已经耳熟能详的概念，其实早在 2008 年就已经被提出来了。2008 年，在 Google 成立 10 周年之际，世界著名杂志《自然》出版了一期专刊，专门讨论与未来的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。

大数据的概念能广为人知其实要归功于以下两件事情：2011 年麦肯锡全球研究院发布的研究报告《大数据：下一个创新、竞争和生产力的前沿》，该报告系统地阐述了大数据概念，并详细列举了大数据的核心技术。之后，经 Gartner 新兴技术成熟度曲线（见图 1-1 和图 1-2）和 2012 年维克托·迈尔-舍恩伯 格《大数据时代：生活、工作与思维的大变革》的宣传推广，大数据概念开

始风靡全球。

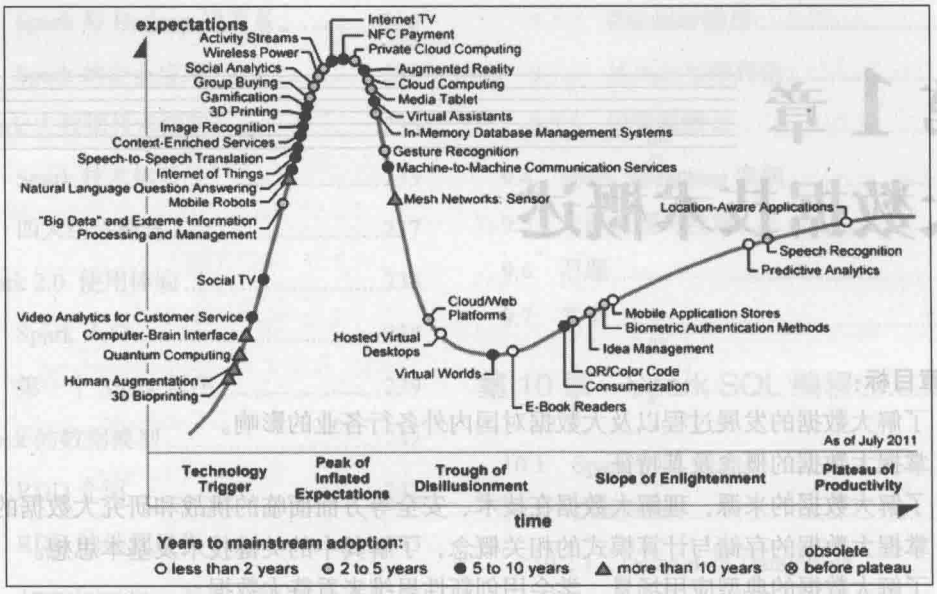


图 1-1　Gartner 曲线 2011 年针对 Big Data 的预测情况

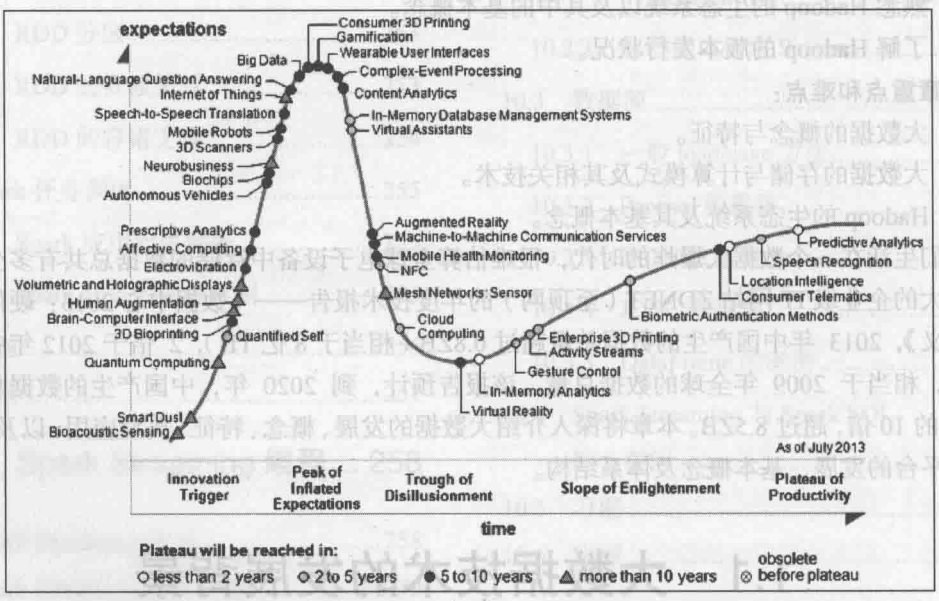


图 1-2　Gartner 曲线 2013 年针对 Big Data 的预测情况

### 1.1.1　大数据技术的发展过程

大数据技术的出现比大数据的概念被正式提出要早得多，到目前为此已经历了多个不同的发展阶段。

**1. 萌芽阶段**

20 世纪 90 年代至 21 世纪初，是大数据技术发展的萌芽期。在此阶段，数据库技术已逐步成

熟，数据挖掘理论也不断完善，因此也被称为数据挖掘技术阶段。在这期间，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。此时，对于大数据处理的研究主要集中于算法（algorithms）、模型（model）、模式（patterns）、标识（identification）等领域。

**2. 突破阶段**

2003 至 2006 年是大数据技术发展的突破期。在此阶段，学术界和企业界开始从多角度对数据处理系统、数据库架构进行重新思考。以 2004 年 Facebook 的创立为标志，Web 2.0 应用（如社交网络、电子商务等）的流行，直接导致了非结构化数据的大量涌现，使得传统数据库处理方法难以应对，从而导致了大数据技术的异军突起。该阶段也被称为非结构化数据阶段。此时，非结构化数据处理得到了广泛而深入的探索和研究，但仍然没有形成共识。

**3. 成熟阶段**

2006 至 2009 年，是大数据技术发展的成熟阶段。首先，在 2003 年和 2004 年，Google 公司先后公开发表了两篇论文——*The Google File System*（《谷歌文件系统》）、*MapReduce: Simplified Data Processing on Large Clusters*（《基于集群的简单数据处理：MapReduce》），公开了 Google 搜索引擎基于大数据处理的解决方案。其核心技术包括分布式文件系统 GFS、分布式计算系统框架 MapReduce、分布式锁机制 Chubby 以及分布式数据库 BigTable 等。以此为基础，从 2006 年开始，Apache 基金会的开源社团和企业纷纷推出了各种各样的 Google 大数据技术的开源实现，从而推动大数据技术逐渐走向了成熟。在此期间，大数据技术研究的焦点是性能（performance）、云计算（cloud computing）、大规模数据集并行运算算法（MapReduce）以及开源分布式系统基础架构（Hadoop）等。

**4. 应用阶段**

2009 年至今，大数据技术架构和大数据技术生态系统越来越完善，尤其是 Hadoop 大数据技术平台的成熟，标志着大数据技术的发展正式进入了落地应用阶段。学术界和企业界纷纷开始从大数据技术的基础性研究转向大数据技术的应用研究。到 2013 年时，大数据技术开始向商业、科技、医疗、政府、教育、经济、交通、物流、人文以及社会的其他领域进行全面深入渗透，从而引起了整个社会的变革。因此，2013 年被称为“大数据元年”。如今，大数据正在影响社会的方方面面，并已成长为一种能催生各行各业变革的巨大力量。

## 1.1.2　大数据技术的影响

近年来，大数据不断向社会各行各业渗透，使得大数据的技术领域和行业边界越来越模糊，应用创新已超越技术本身而受到更多青睐。大数据技术已经为每一个领域带来了变革性影响，并且正在成为各行各业颠覆性创新的原动力和助推器。

**1. 大数据技术在国外**

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得到全球各行业的高度关注和重视，掀起了一个可与 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究计划，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和信息技术顾问委员会就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了体现国家意志的战略高度。该报告列举了 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。该报告指出：“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”。该报告建议：“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、卫生健康总署、能源部、国防部等 6 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。该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，更好地为国民服务，并建议成员国建立“脉搏实验室”，挖掘大数据的潜在价值。

2013 年 5 月，麦肯锡全球研究院（McKinsey Global Institute）发布了一份名为《颠覆性技术：技术进步改变生活、商业和全球经济》的研究报告。该报告指出未来的 12 种新兴技术有望在 2025 年带来 14 万亿至 33 万亿美元的经济效益。出人意料的是，在这份报告中最为热门的大数据技术却未被列入其中。麦肯锡的解释是，大数据已成为 12 种技术中许多技术的基石，包括移动互联、知识工作自动化、物联网、云计算、机器人、自动汽车、基因组学等。

2014 年 5 月，美国政府发布了 2014 年全球大数据白皮书的研究报告《大数据：抓住机遇、守护价值》。该报告提出要使用数据来推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究来保护个人隐私、公平以及反歧视的社会信仰。

2014 年 4 月，世界经济论坛也以“大数据的回报与风险”为主题发布了《全球信息技术报告（第 13 版）》。该报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要。

**2. 大数据技术在我国**

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予以了高度关注。中央电视台分别于 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，充分体现了大数据技术已经成为国家和社会普遍关注的焦点。

国内的学术界和企业界也都迅速地行动了起来，广泛地开展了对大数据技术的研发。为了推动我国大数据技术的研究发展，2012 年中国计算机学会发起并组织了大数据专家委员会，该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，撰写发布了《2013 年中国大数据技术与产业发展白皮书》。2013 年以后，国家自然科学基金、973 计划、核高基、863 等重大研究计划都已经把大数据研究列为重大研究课题。

2015 年 9 月，国务院印发《促进大数据发展行动纲要》，系统部署了大数据发展工作。该纲要明确提出要推动大数据发展和应用，在未来 5~10 年打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态。该纲要部署了三方面主要任务。一要加快政府数据开放共享，推动资源整合，提升治理能力。大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商事服务便捷化，促进安全保障高效化，加快民生服务普

惠化。二要推动产业创新发展，培育新兴业态，助力经济转型。发展大数据在工业、新兴产业、农业农村等行业领域应用，推动大数据发展与科研创新有机结合，推进基础研究和核心技术攻关，形成大数据产品体系，完善大数据产业链。三要强化安全保障，提高管理水平，促进健康发展。健全大数据安全保障体系，强化安全支撑。

2016 年 3 月 17 日，国家“十三五”规划纲要发布。该纲要明确指出：一是加快政府数据开放共享。全面推进重点领域大数据高效采集、有效整合，深化政府数据和社会数据关联分析、融合利用，提高宏观调控、市场监管、社会治理和公共服务的精准性和有效性。依托政府数据统一共享交换平台，加快推进跨部门数据资源共享。加快建设国家政府数据统一开放平台，推动政府信息系统和公共数据互联开放共享。制定政府数据共享开放目录，依法推进数据资源向社会开放。统筹布局建设国家大数据平台、数据中心等基础设施。研究制定数据开放、保护等法律法规，制定政府信息资源管理办法。二是促进大数据产业健康发展。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。促进大数据软硬件产品发展。完善大数据产业公共服务支撑体系和生态体系，加强标准体系和质量技术基础建设。

### 1.1.3　大数据发展的重大事件

2005 年 Hadoop 项目诞生。Hadoop 最初只是雅虎公司用来解决网页搜索问题的一个项目，后来因其技术的高效性，被 Apache 基金会引入并成为开源应用。Hadoop 本身不是一个软件产品，而是由多个软件产品组成的一个生态系统，这些产品共同实现了功能全面和灵活的大数据分析。Hadoop 由两个核心构成：HDFS 和 MapReduce。HDFS 是 Hadoop 分布式文件系统，用于提供可靠数据存储服务。MapReduce 则用于提供高性能的并行数据处理服务。

2008 年年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟（Computing Community Consortium），发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》，使人们的思维不再局限于进行数据处理的机器，并提出“大数据真正重要的是新用途和新见解，而非数据本身”。

2009 年，印度政府建立了用于身份识别管理的生物识别数据库，而联合国全球脉冲项目也已研究了如何利用手机和社交网站的数据源来分析预测从螺旋 CT 价格到疾病暴发之类的问题。

2009 年，美国政府通过启动 data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供了 4 万多个各种各样的政府数据集，这些数据集可以面向一些智能手机应用程序，提供从航班到产品召回再到特定区域内失业率的跟踪信息。这一行动推动从肯尼亚到英国范围内的政府相继推出了类似举措。

2009 年，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系，致力于改善在互联网上获取科学数据的简易性。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到：世界上有着无法想象的海量数字信息，并以极快的速度增长。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到了这种海量信息的影响。科学家和计算机工程师已经为这个现象创造了一个新词：“大数据”。库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2011 年 2 月，IBM 最新研发的沃森超级计算机每秒可扫描并分析 4TB（约 2 亿页文字量）的数据量，并在美国著名智力竞赛电视节目《危险边缘》上击败两名人类选手而夺冠。后来纽约时

报认为这一刻是一次“大数据计算的胜利”。

2011 年 5 月，全球知名咨询公司麦肯锡全球研究院（MGI）发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，大数据开始备受关注，这也是专业机构第一次全方面地介绍和展望大数据。该报告指出，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。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。报告还提到，“大数据”源于数据生产和收集的能力和速度的大幅提升——由于越来越多的人、设备和传感器通过数字网络连接起来，产生、传送、分享和访问数据的能力也得到了彻底变革。

2011 年 11 月，工业和信息化部发布了《物联网“十二五”发展规划》，在关键技术创新工程部分，信息处理技术作为四项之一被提出来，其中的核心内容是海量数据存储、数据挖掘、图像视频智能分析，而这些都是大数据的重要组成部分。

2012 年 1 月，在瑞士达沃斯举办的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》（*Big Data, Big Impact*）宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。

2012 年 3 月，美国奥巴马政府在白宫网站上发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。2012 年 3 月 22 日，奥巴马政府宣布将 2 亿美元投资于大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭。在 23 日的电话会议中，美国政府将数据比喻为“未来的新石油”，并表示大数据技术领域的竞争事关国家安全和未来，即：国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用的能力，国家数字主权体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。

2012 年 4 月，美国软件公司 Splunk 于 19 日在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低靡、股市持续振荡的大背景，Splunk 股份在首日就暴涨了一倍多的突出交易表现尤其令人们印象深刻。Splunk 是一家领先的提供大数据监测和分析服务的软件提供商，成立于 2003 年。Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快了大数据布局。

2012 年 7 月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明了在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求。例如，通过对价格关注和更好服务的渴望，个人提供数据和众包[①]信息，并对隐私和退出权力提出需求；公共部门出于改善服务、提升效益的目的，提供了诸如统计数据、设备信息、健康指标及税务和消费信息等，并对隐私和退出权力提出需求；私人部门出于提升客户认知和预测趋势的目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。白皮书还指出，人们如今可以使用丰富的数据资源，包括旧数据和新数据，来对社会人口进行前所未有的实时分析。联合国还以爱尔兰和美国的社交网络活跃度增长可以作为失业率上升的早期征兆为例表明，政府如果能合理分析所掌握的数据资源，将能“与数俱进”，快速应变。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台——“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。马云强调：“假如我们有一个数据预报台，就像为企业装上了一个 GPS 和雷达，你们出海将会更有把握。”因此，阿

① 众包指的是一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的大众网络的做法。

里巴巴集团希望通过分享和挖掘海量数据，为国家和中小企业提供价值。此举是国内企业最早把大数据提升到企业管理层高度的一个重大里程碑。阿里巴巴也是最早提出通过数据进行企业数据化运营的企业。

2013 年 1 月 24 日，英国商业、创新和技能部宣布，英国政府将注资 6 亿英镑（1 英镑约合 1.57 美元），发展大数据、合成生物等 8 类高新技术。其中，1.89 亿英镑用来发展大数据技术。同年 7 月，中国上海市发布了《上海推进大数据研究与发展三年行动计划》（2013—2015 年）。2016 年 9 月，上海市又发布了《上海市大数据发展实施意见》，并于同年 10 月获批成立国家大数据示范综合试验区。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”为主题发布了《全球信息技术报告（第 13 版）》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要。报告表示，接下来将针对数据保密和网络管制等议题展开积极讨论。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到了大数据在推动经济发展、改善公共服务、增进人民福祉，乃至保障国家安全方面的重大意义。

2014 年 5 月，美国白宫发布了 2014 年全球大数据白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

2015 年 9 月，国务院正式印发《促进大数据发展行动纲要》，以推动大数据发展和应用。

2016 年 3 月 17 日，国家“十三五”规划纲要发布。该纲要提出要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，加快政府数据开放共享，促进大数据产业健康发展。

2016 年 7 月 14 日，首届中国大数据应用大会在成都拉开帷幕，国内外行业专家、龙头企业、行业用户及主流媒体云集成都，共商大数据应用之道。该大会以“大数据与智能时代”为主题，围绕智能制造、大数据核心技术、地理信息与大数据、大数据与健康医疗、大数据与互联网金融、宏观经济大数据等当前热点领域展开了讨论。

## 1.2　大数据的概念、特征及意义

### 1.2.1　什么是大数据

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，不同的机构或个人有不同的理解，难以有一个非常定量的定义。

美国咨询公司——麦肯锡公司是研究大数据的先驱。该公司在其报告《大数据：创新、竞争和生产力的下一个前沿领域》中针对大数据给出的定义是：大数据指的是大小超出常规的数据库工具能获取、存储、管理和分析的数据集。该报告同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。

国际数据公司（IDC）从 4 个特征定义大数据，即海量的数据规模（volume）、快速的数据流转和动态的数据体系（velocity）、多样的数据类型（variety）和巨大的数据价值（value）。

亚马逊公司的大数据科学家 John Rauser 给出了大数据的简单定义：Big data is any amount of

data that's too big to be handled by one computer（大数据是任何超出了一台计算机处理能力的数据量）。

维基百科对大数据的定义是：大数据指的是所涉及的资料量规模巨大到无法通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实现更积极目的的信息。

《大数据时代的历史机遇》一书的作者认为：大数据是“在多样的或者大量数据中，迅速获取信息的能力”。

可见，大数据是一个宽泛的概念，见仁见智，有些人可能强调数据的规模，即“大”字；有些人则可能强调大数据的作用，即大数据能帮助人们做什么；甚至有些人更强调新数据处理技术的应用。综合而言，本书采用“百度百科”的定义：大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，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。

### 1.2.2 大数据的特征

大数据是一种数据量增长速度极快，用传统的数据处理方法或工具无法在用户所要求的时间内完成采集、处理、存储和计算的数据集合，它具有以下五大特征。

**1. 数据量大（volume）**

大数据的第一个特征是数据量大，包括采集、存储和计算的量都非常大。大数据的起始计量单位至少是 PB，也可采用更大的单位 EB 或 ZB。相关信息单位的换算关系如下。

1 Byte =8 bit

1 KB = 1 024 Bytes = 8192 bit

1 MB = 1 024 KB = 1 048 576 Bytes

1 GB = 1 024 MB = 1 048 576 KB

1 TB = 1 024 GB = 1 048 576 MB

1 PB = 1 024 TB = 1 048 576 GB

1 EB = 1 024 PB = 1 048 576 TB

1 ZB = 1 024 EB = 1 048 576 PB

**2. 类型繁多（variety）**

大数据的第二个特征是种类和来源多样化。大数据可以是结构化、半结构化和非结构化的数据，具体表现为网络日志、音频、视频、图片、地理位置信息等，多类型的数据对数据的处理能力提出了更高的要求。

**3. 价值密度低（value）**

大数据的第三个特征是数据价值密度相对较低。有人把大数据比喻成金矿，金矿只有经过反复清洗与筛查，才能获取其中的黄金，大数据是浪里淘沙却又弥足珍贵。特别是，随着互联网以及物联网的广泛应用，智能感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的数据挖掘与机器学习算法来挖掘数据价值，是大数据时代最需要解决的问题。

**4. 速度快时效高（velocity）**

大数据的第四个特征数据增长速度快，处理速度也快，时效性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。