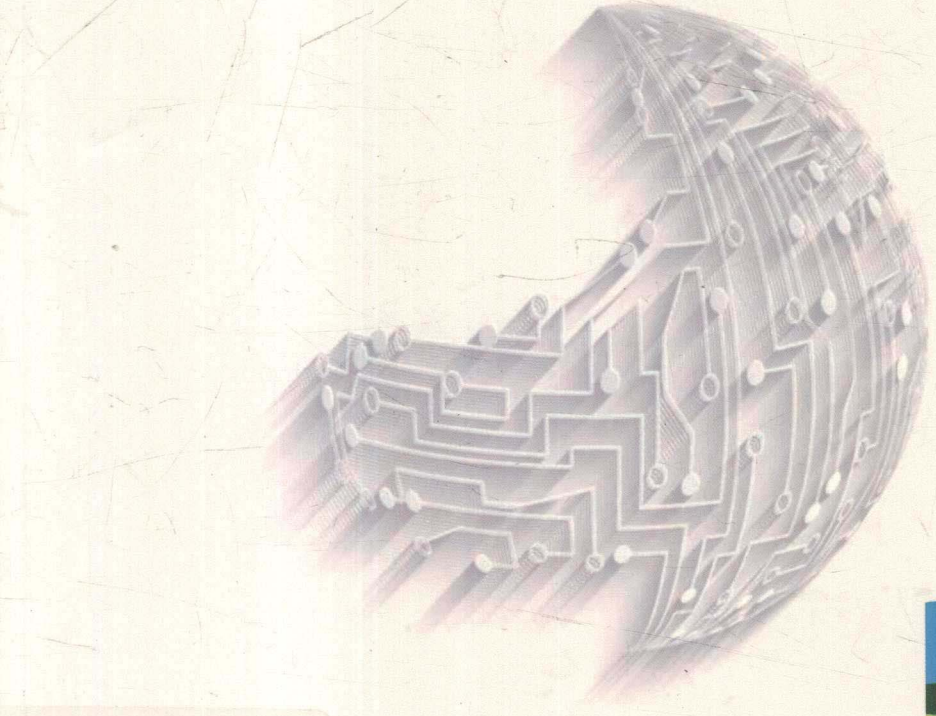
BIG DATA

大数据处理技术与系统研究

郑睿颖 著



-级出版社 中国纺织出版社 全国百佳图书出版单位

大数据处理技术与系统研究

郑睿颖◎著

中国纺织出版社

图书在版编目（CIP）数据

大数据处理技术与系统研究 / 郑睿颖著 . -- 北京 : 中国纺织出版社 , 2019.4
ISBN 978-7-5180-6010-8

Ⅰ . ①大… Ⅱ . ①郑… Ⅲ . ①数据处理—研究 Ⅳ . ① TP274

中国版本图书馆 CIP 数据核字 (2019) 第 050143 号

责任编辑：闫　星　　**责任印制：**王跃杰

中国纺织出版社出版发行
地　　址：北京市朝阳区百子湾东里 A407 号楼　　**邮政编码：**100124
销售电话：010-67004422　　**传真：**010-87155801
http://www.c-textilep.com
E-mail：faxing@c-textilep.com
中国纺织出版社天猫旗舰店
官方微博 http://weibo.com/2119887771
北京虎彩文化传播有限公司印刷　　各地新华书店经销
2019 年 4 月第 1 版第 1 次印刷
开　　本：880mm × 1230mm　1/16　　**印张：**8.5
字　　数：129 千字　　**定价：**39.80 元

凡购买本书，如有缺页、倒页、脱页由本社图书营销中心调换

引 言

随着计算机和信息技术的迅猛发展和普及应用，行业数据爆炸性增长，全球已经进入了大数据时代。大数据已引起全球业界、学术界和各国政府的高度关注。大数据已经渗透到各行各业，巨大的数据资源已成为国家和企业的战略资源。

大数据给全球带来了重大的发展机遇与挑战。一方面，大规模数据资源蕴含着巨大的商业价值和社会价值，有效地管理和利用这些数据、挖掘数据的深度价值，对国家治理、社会管理、企业决策和个人生活将带来巨大的影响。

另一方面，大数据带来新的发展机遇的同时，也带来很多技术挑战。格式多样、形态复杂、规模庞大的大数据给传统的计算技术带来了巨大挑战，传统的信息处理与计算技术已难以有效地应对大数据的处理。因此，从计算技术的多个层面出发，采用新的技术方法，才能提供有效的大数据处理技术手段和方法。

目　录

试读结束：需要全本请在线购买：　www.ertongbook.com

第一章　大数据的概念和发展背景

第一节　大数据的发展背景

在20世纪90年代后期，当气象学家在做气象地图分析、物理学家在建立大物理仿真模型、生物学家在建立基因图谱的分析过程中，由于数据量巨大，他们已经不能再用传统的计算技术来完成这些任务时，大数据的概念在这些科学研究领域首先被提出来。面对大量科学数据在获取、存储、搜索、共享和分析中遇到的技术难题，一些新的分布式计算技术陆续被研究和开发了出来。

2008年，随着互联网和电子商务的快速发展，当Yahoo、Google等大型互联网和电子商务公司不能用传统手段解决他们的业务问题时，大数据的理念和技术开始被他们实际应用。他们共同遇到的问题是，处理的数据量通常很大（那时是PB级，1个PB的数据相当于全美学术研究图书馆的藏书和资讯内容的50%），数据的种类很多（文档、日志、博客、视频等），数据的流动速度很快（包括流文件数据、传感器数据和移动设备的数据的快速流动），而且，这些数据经常是不完备甚至是不可理解的（需要从预测分析中推演出来）。大数据的新技术和新架构正是在这种背景下被不断开发出来的，以有效地解决这些现实的互联网数据处理问题。

2010年，全球进入Web 2.0时代，Twitter（推特）、Facebook（脸书）、博客、微博、微信等社交网络将人类带入自媒体时代，互联网数据激增。随着智能手机的普及，移动互联网时代也已经到来，移动设备所产生的海量数据涌入网络。为

了实现更加智能的应用，物联网技术也逐步被推广，随之而来的是更多实时获取的视频、音频、电子标签（RFID）、传感器等数据也被联入互联网，数据量进一步暴增。根据美国国际数据公司 IDC 的预测，人类产生的数据量正在呈指数级增长，大约每两年翻一番，这个速度在 2020 年之前会继续保持下去。全球在 2010 年已正式进入 ZB 时代（1 个 ZB 的数据相当于全世界海滩上的沙子数量的总和），预计到 2020 年，全球将总共拥有 35ZB 的数据量，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。人类真正进入了一个数据的世界，大数据技术有了用武之地，大数据技术和应用空前繁荣起来。

2011 年，全球著名战略咨询公司麦肯锡的全球研究院 （MGI）发布了《大数据：创新、竞争和生产力的下一个新领域》研究报告，这份报告分析了数字数据和文档的爆发式增长的状态，阐述了处理这些数据能够释放出的潜在价值，分析了大数据相关的经济活动和业务价值链。这篇报告在商业界引起了极大的关注，为大数据从技术领域进入商业领域吹响了号角。

2012 年 3 月 29 日奥巴马政府以“大数据是一个大生意（Big Data is a Big Deal）”为题发布新闻，宣布投资 2 亿美元启动“大数据研究和发展计划”，涉及美国国家科学基金、美国国防部等 6 个联邦政府部门，大力推动和改善与大数据相关的收集、组织和分析工具及技术，以提高从大量的、复杂的数据集合中获取知识和洞见的能力。美国政府认为大数据技术事关美国国家安全、科学和研究的步伐。

2012 年 5 月，联合国发布了一份大数据白皮书，总结了各国政府如何利用大数据更好地服务公民，指出大数据对于联合国和各国政府来说是一个历史性的机遇，联合国还探讨了如何利用包括社交网络在内的大数据资源造福人类。

2012 年 12 月“世界经济论坛”发布《大数据，大影响》报告，阐述大数据为国际发展带来的新的商业机会，建议各国与工业界、学术界、非营利性机构与管理者一起利用大数据创造机会。

2012 年以来，大数据成为全球投资界所青睐的领域之一，IBM 公司通过并购数据仓库厂商 Netezza、软件厂商 Info Sphere Big Insights 和 Streams 等来

增强自己在大数据处理上的实力；EMC 公司陆续收购 Greenplum（Pivotal）、VMware、Isilo 等公司，展开大数据和云计算产业的战略布局；惠普公司通过并购 3PAR、Autonomy、Vertica 等公司实现了大数据产业链的全覆盖。业界主要的信息技术巨头都纷纷推出大数据产品和服务，力图抢占市场先机。

2012 年以来，国内互联网企业和运营商率先启动大数据技术的研发和应用，如淘宝、百度、腾讯、京东、中国移动、中国联通等企业纷纷启动了大数据试点应用项目，推广大数据应用。

2013 年，第 4 期《求是》杂志刊登中国工程院邬贺铨院士的《大数据时代的机遇与挑战》一文，阐述中国科技界对大数据的重视。郭华东、李国杰、倪光南、怀进鹏等院士也纷纷撰文阐述大数据的战略意义，清华大学、北京大学等高校纷纷设立大数据方面的学院和专业，以推进大数据技术的研发。

2015 年，《促进大数据发展行动纲要》正式颁布，提出大数据已成为国家基础性战略资源，是推动经济转型和发展的新动力，是重塑城市竞争优势的新机遇，是提升政府治理能力的新途径，中国正式启动和实施国家大数据战略。

第二节　大数据的概念和特征

一、大数据的概念

大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。

虽然这个定义并不严谨，但这是各种学术和应用领域最广泛引用的一个定义，如果接着以大数据的四个特征作为补充，就能给出一个较为清晰的大数据的概念。《促进大数据发展行动纲要》指出，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。

二、大数据的特征

大数据有四个主要特征。

（1）Volume：数据容量大。容量大是大数据区分于传统数据最显著的特征。一般关系型数据库处理的数据容量在 TB 级，大数据技术所处理的数据容量通常在 PB 级以上。

（2）Variety：数据类型多。大数据技术所处理的计算机数据类型早已不是单一的文本形式或者结构化数据库中的表，它包括网络日志、音频、视频、机器数据等各种复杂结构的数据。

（3）Velocity：数据存取速度快。存取速度是大数据区分于传统数据的重要特征。在海量数据面前，需要快速实时存取和分析需要的信息，处理数据的效率就是组织的生命。

（4）Value：数据应用价值高。在研究和技术开发领域，上述三个特征已经足够表征大数据的特点。但在商业应用领域，第四个特征就显得非常关键。投入如此巨大的研究和技术开发的努力，就是因为大家都洞察到了大数据潜在的巨大应用价值。如何通过强大的机器学习和高级分析更迅速地完成数据的价值“提纯”，挖掘出大数据的应用价值，是目前大数据技术应用的发展重点。

第三节　大数据的产生

大量数据的产生是计算机和网络通信技术（ICT）广泛应用的必然结果，特别是互联网、云计算、移动互联网、物联网、社交网络等新一代信息技术的发展，起到了巨大的作用，它带来了数据产生的四大变化：一是数据产生由企业内部向企业外部扩展；二是数据产生由 Web 1.0 向 Web 2.0 扩展；三是数据产生由互联网向移动互联网扩展；四是数据产生由计算机 / 互联网（IT）向物联网（IOT）扩展。这四大变化，让数据的产生源头成倍地增长，数据量也相

应的大幅度快速增长。

一、数据产生由企业内部向企业外部扩展

在企业内部的企业资源计划（ERP）、管理和决策分析系统、办公自动化（OA）等业务所产生的数据，主要存储在关系型数据库中。内部数据是企业内最成熟并且被熟知的数据。这些数据已经通过多年的ERP、数据仓库（DW）、商业智能（BI）和其他相关应用积累，实现了内部数据的收集、集成、结构化和标准化处理，可以为企业决策提供分析报表和商业智能。

一些企业已经关注到交易行为数据的潜在价值，如利用一些非结构化数据的分析方法，挖掘在客户交易过程、业务处理流程和电子邮件中所获得的内部日志等数据，为企业提供客户分析、绩效分析和风险管理等方面的更多洞察力。还有一些大型企业内部的数据量也很大，如电信运营商、石油勘探企业等，这些企业使用大数据有很多年了。例如，一家全球电信公司每天从120个不同系统中收集数十亿条详细呼叫记录，并保存至少9个月时间；一家石油勘探公司分析几万亿字节的地质数据。对于这些公司，大数据虽然是一个新概念，但要做的事情却并不新鲜。他们早就在使用大数据，但由于没有合适的技术手段对这些大数据进行分析，导致这些大数据中的大部分被丢弃了。

对于所有企业而言，信息化的应用环境在发生着变化，外部数据迅速扩展，企业和互联网、移动互联网、物联网的融合越来越快。企业需要通过互联网来服务客户、联系外部供应商、沟通上下游的合作伙伴，并在互联网上实现电子商务和电子采购的交易。企业需要开通微博、博客等社交网络来进行网络化营销、客户关怀和品牌建设。企业的产品被贴上了电子标签，在制造、供应链和物流的过程中进行跟踪和反馈。伴随着自带设备（BYOD）工作模式的兴起，企业员工自带设备进行工作，个人的数据进一步与企业数据相融合，必将产生更多来自企业外部的数据。

二、数据产生从 Web 1.0 向 Web 2.0、从互联网向移动互联网扩展

随着社交网络的发展，互联网进入了 Web 2.0 时代，每个人从数据的使用者，变成了数据的生产者，数据规模迅速扩张，每时每刻都在产生大量的新数据。例如，从全球统计数据来看，全球每秒钟发送 290 万封电子邮件，每秒钟电子商务公司 Amazon 上将产生 72.9 笔商品订单，每分钟会有 20 个小时的视频上传到视频分享网站 YouTube，Google 上每天需要处理 24PB 的数据，Twitter 上每天发布 5000 万条消息，每天被每个家庭消费的数据有 375MB，网民每个月在 Facebook 上要花费 7000 亿分钟……

从中国来看，数据规模也十分巨大，淘宝网目前已拥有近 5 亿的注册会员，在线商品 8.8 亿，每天交易超过数千万笔，其单日数据产生量超过 20TB。百度目前数据总量接近 1000PB，存储网页数量接近 1 万亿，每天大约要处理 60 亿次搜索请求，几十拍字节数据。新浪微博每天有数十亿外部网页和 API 接口访问需求，服务器群在晚上高峰期每秒要接收 100 万个以上的响应请求。

移动互联网的发展让更多人成为了数据的生产者，据统计，全球每个月移动互联网使用者发送和接收的数据高达 1.3EB。在中国，中国联通用户上网记录条数为 83 万条 / 秒，即一万亿条 / 月，对应数据量为 300TB/ 月，或 3.6PB/ 年。

三、数据产生从计算机 / 互联网（IT）向物联网（IOT）扩展

随着视频设备、传感器、智能设备和 RFID 等技术的增长，视频、音频、RFID、机器对机器（M2M）、物联网和传感器等数据大量产生，其数据量更是巨大。根据 IDC 公布的数据，2005 年仅由 M2M 产生的数据就占全世界数据总量的 11%，预计到 2020 年这一数值将增加到 42%。

试读结束：需要全本请在线购买： www.ertongbook.com

第四节　数据的量级

一、数据大小的量级

数据量的大小是用计算机存储容量的单位来计算的，基本的单位是字节（Byte），每一级按照千分位递进，如下所示：

1Byte（B）相当于一个英文字母；1Kilobyte（KB）＝ 1024B 相当于一则短篇故事的内容；1Megabyte（MB）＝ 1024KB，相当于一则短篇小说的文字内容；1Gigabyte（GB）＝ 1024MB，相当于贝多芬第五乐章交响曲的乐谱内容；1Terabyte（TB）＝ 1024GB，相当于一家大型医院中所有的 X 光图片内容；1Petabyte（PB）＝ 1024TB，相当于 50% 的全美学术研究图书馆藏书信息内容；1Exabyte（EB）＝ 1024PB，5EB 相当于至今全世界人类所讲过的话语；1 Zettabyte（ZB）＝ 1024EB，相当于全世界海滩上的沙子数量的总和；1 Yottabyte（YB）＝ 1024ZB，相当于 1024 个像地球一样的星球上的沙子数量的总和。

二、大数据的量级

目前，传统企业的数据量基本在 TB 级以上，一些大型企业达到了 PB 级，Google、百度、腾讯、阿里巴巴这些企业的数据量在 PB 级以上。

大数据技术和应用擅长处理的数量级一般都在 PB 级以上。但数据量的巨大是相对处理这些数据的计算设备而言的，例如，对一台小型机或 PC 服务器，PB 级数据是大数据，但可能对一台智能手机而言，GB 级的数据就是“大数据”。就目前大数据技术架构所处理的数据来看，通常是指 PB 级以上的数据。

摩尔定律是由英特尔（Intel）创始人之一戈登·摩尔（Gordon Moore）提出来的，其内容为：当价格不变时，集成电路上可容纳的晶体管数目约每隔 18 个月便会增加一倍，性能将提升一倍。这一定律揭示了信息技术进步的速度。吉姆·格

雷（Jim Gray）的新摩尔定律认为，每 18 个月全球新增的信息量是计算机有史以来全部信息量的总和，数据容量每 18 个月就翻一番。据 IDC 统计，全球在 2010 年正式进入 ZB 时代，预计到 2020 年，全球将总共拥有 35ZB 的数据量。在过去的 50 年，数据存储的成本大概每两年就能降一半，而存储密度却增加了 5000 万倍。

因此，我们的世界正在成为一个数据的世界，我们正处于大数据时代，像水、空气、石油一样，数据正成为这个世界中的一种重要资源。

第二章　大数据应用的总体架构和处理技术

第一节　总体架构

一、业务目标

大数据产生、聚集、分析和利用的各个阶段都提出了一些需求，这些需求需要通过大数据技术来实现，这能从架构层面实现业务需求向技术要求的映射。

大数据应用的总体架构被业务需求逐步勾勒了出来：大数据应用需要采用一个统一集成的大数据平台，使得用户能够快速处理和加载海量数据，能够在统一平台上对不同类型的数据进行处理和存储；大数据应用需要采用一个数据集成和管理平台，集成各种工具和服务来管理异构存储环境下的各类数据，并建立一个实时预测分析解决方案，整合结构化的数据仓库和非结构化的分析工具。在平台上用户可以在任何时间、任何地点通过任何设备进行大数据的集中共享、协同和分析，大数据应用总体架构能够支撑组织对新的业务战略进行建模，提升组织的洞察力。

二、架构设计原则

企业级大数据应用架构需要满足业务的需求：一是要能够满足基于数据容量大、数据类型多、数据存取速度快的大数据基本处理需求，能够支持大数据的

采集、存储、处理和分析；二是要能够满足企业级应用在可用性、可靠性、可扩展性、容错性、安全性和保护隐私等方面的基本准则；三是要能够满足用原始技术和格式来实现的数据分析的基本要求。

（一）满足大数据 V3 的要求

1. 大容量数据的加载、处理和分析

要求大数据应用平台经过扩展可以支持 GB、TB、PB、EB 甚至 ZB 的数据集。

2. 各种类型数据的加载、处理和分析

支持各种各样的数据类型，包括支持处理交易数据、各种非结构化数据、机器数据及其他新结构数据。支持极端的混合工作负载，包括数以千计的地理上分布的在线用户和程序，这些用户和程序执行各种各样的请求，范围从临时性的请求到战略分析的请求，同时以批量或流的方式加载数据。

3. 大数据的处理速度

在很高速度（GB/ 秒）的加载过程中集成来自多个来源的数据。对高度限定的标准 SQL 查询的亚秒级响应时间。以至少每秒千兆字节的速度高速加载数据，随时进行分析。以满负荷速度就地更新数据。在传入的加载数据上实时执行某些“流”分析查询。

（二）满足企业级应用的要求

1. 高可扩展性

要求平台符合企业未来业务发展要求及对新业务的响应，能够支持大规模数据计算的节点可扩展，能适应将来数据结构的变化、数据容量的增长、用户的增加、查询要求和服务内容的变化，要求大数据架构具备支持调度和执行数百上千节点的复杂工作流。

2. 高可用性（容错）

要求平台能够具备实时计算环境所具备的高可用性，在单点故障的情况下能够保证应用的可用性，具备处理节点故障时的故障转移和流程继续的能力。

3. 安全性和保护隐私

系统在数据采集、存储、分析架构上，保证数据、网络、存储和计算的安全性，具备保护个人和企业隐私的措施。

4. 开放性

要求平台能够支持计算和存储分布到数以千计的、地理位置可能不同的、可能异构的计算节点，能够识别和整合不同技术和不同厂商开发的工具和应用，能够支持移动应用、互联网应用、社交网络、云计算、物联网、虚拟化、网络、存储等多种计算设备、计算协议和计算架构。

5. 易用性

系统功能操作易用，能满足大多数企业业务、管理和技术人员的操作习惯。平台具有可编程性，能够支持不同编程工具和语言的集成，具备集成编译环境。能在处理请求内嵌入任意复杂的用户定义函数（UDF），以各种行业标准过程语言执行 UDF，组合大部分或全部使用案例的大量可复用 UDF 库，在几分钟内对 PB 级大小的数据集执行 UDF“关系扫描”。

（三）满足对原始格式数据进行分析的要求

系统具备对复杂的原始格式数据进行整合分析的能力，如对文本数据、数学数据、统计数据、金融数据、图像数据、声音数据、地理空间数据、时序数据、机器数据等进行分析的能力。

三、总体架构参考模型

大数据的产生、组织和处理主要是通过分布式文件处理系统来实现的，主流的技术是 Hadoop+MapReduce，其中 Hadoop 的分布式文件处理系统（HDFS）作为大数据存储的框架，分布式计算框架 MapReduce 作为大数据处理的框架。

（一）大数据存储的框架

HDFS：即 Hadoop 分布式文件处理系统，分布式文件处理系统运行于大规模集群之上，集群使用廉价的通用服务器构建，整个文件系统采用的是元数据集中管理与数据块分散存储相结合的模式，并通过数据的复制来实现高度容错。分

布式文件处理系统架构在通用的服务器、操作系统或虚拟机上。

（二）大数据处理的框架

MapReduce：一个分布式并行计算软件框架，基于 Map（可理解为“任务分解”）和 Reduce（可理解为“综合结果”）的 Java 函数，基于 MapReduce 写出来的应用程序能够运行在由上千个通用服务器组成的大型集群上，并以一种可靠容错的方式并行处理 TB 级别以上的数据集。Mapper 和 Reducer 的主代码可以用多种语言书写。Hadoop 的原生语言是 Java，但是 Hadoop 公开 API 用于以 Ruby 或 Python 等其他语言编写代码。它提供了与 C++ 的接口，其名称为 Hadoop Pipes。在底层进行 MapReduce 编程显然提供了最大的潜力，但这种编程层次非常像汇编语言的编程，属于低级编程语言。

（三）大数据访问的框架

在 Hadoop+MapReduce 之上，架构的是网络层。网络层之上，是大数据访问的框架层。大数据访问的框架实现了对传统关系型数据库和 Hadoop 的访问，主流技术包括 Pig、Hive、Sqoop、Cascading 等。

Pig：是基于 Hadoop 的并行计算高级编程语言，它提供一种类 SQL 的数据分析高级文本语言，称为 Pig Latin，该语言的编译器会把类 SQL 的数据分析请求转换为一系列经过优化处理的 MapReduce 运算。Pig 支持的常用数据分析主要有分组、过滤、合并等。Pig 为创建 Apache MapReduce 应用程序提供了一款相对简单的工具，它有效简化了编写、理解和维护程序的工作，还优化了任务自动执行功能，并支持使用自定义功能进行接口扩展。

Hive：是由 Facebook 贡献的数据仓库工具，是 MapReduce 实现的用来查询分析结构化数据的中间件。Hive 的类 SQL 查询语言 ——Hive QL 可以查询和分析储存在 Hadoop 中的大规模数据。

Sqoop：是由 Cloudera 开发的一种开源工具，用于在 Hadoop 与传统的数据库间进行数据的传递，允许将数据从关系型数据库导入 HDFS 及从 HDFS 导出到关系型数据库。MapReduce 等函数都可以使用由 Sqoop 导入 HDFS 中的数据。

试读结束：需要全本请在线购买： www.ertongbook.com