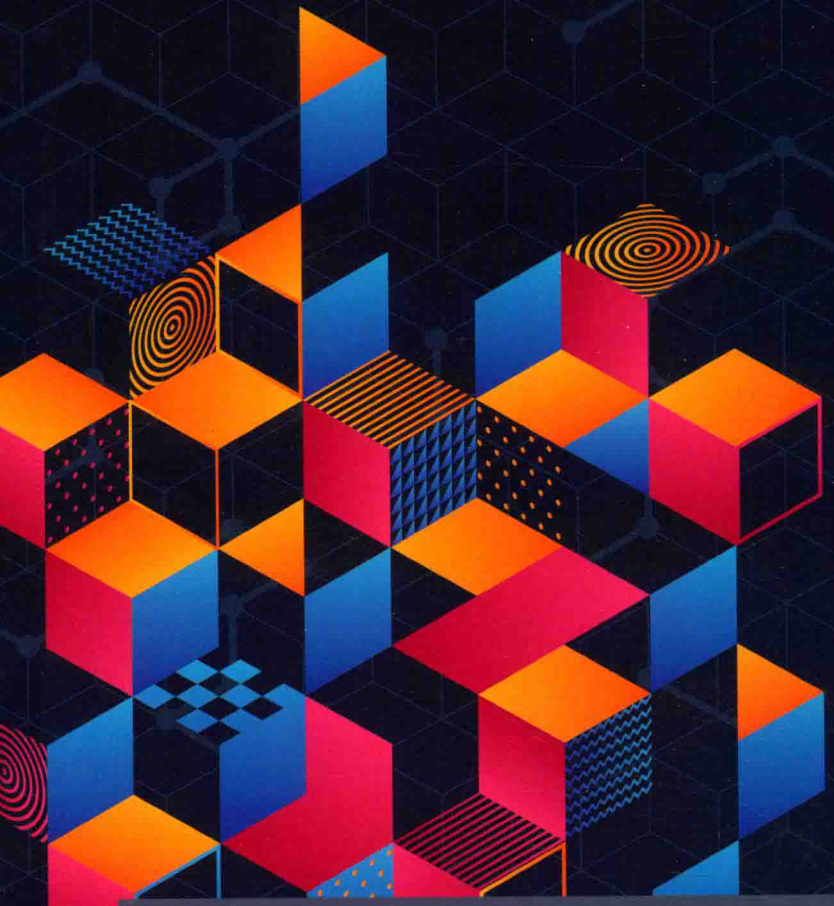
大数据科学与应用丛书



· 深入分析组件原理 ·
· 充分展示搭建过程 ·
· 详细指导应用开发 ·

Hadoop
大数据实战权威指南

黄东军　编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据科学与应用丛书

Hadoop
大数据实战权威指南

黄东军　编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

大数据贵在落实！

本书是一本讲解大数据实战的图书，按照“深入分析组件原理、充分展示搭建过程、详细指导应用开发”编写。全书分为三篇，第一篇为大数据的基本概念和技术，主要介绍大数据的背景、发展及关键技术；第二篇为 Hadoop 大数据平台搭建与基本应用，内容涉及 Linux、HDFS、MapReduce、YARN、Hive、HBase、Sqoop、Kafk、Spark 等；第三篇为大数据处理与项目开发，包括交互式数据处理、协同过滤推荐系统、销售数据分析系统，并就京东的部分销售数据应用大数据进行处理分析。

本书适合初学者入门和进阶，也可供希望全面、系统地理解并掌握大数据实际应用的读者参考，对从事大数据项目开发的专业人员也有参考价值。

为了方便读者实践，本书配有开发资源包，读者可登录华信教育资源网（www.hxedu.com.cn）免费注册后下载。

图书在版编目（CIP）数据

Hadoop 大数据实战权威指南 / 黄东军编著. — 北京：电子工业出版社，2017.7
（大数据科学与应用丛书）
ISBN 978-7-121-31821-4

I. ①H… II. ①黄… III. ①数据处理软件－指南 IV. ①TP274-62

中国版本图书馆 CIP 数据核字（2017）第 129534 号

责任编辑：田宏峰
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：23.5　　字数：526 千字
版　　次：2017 年 7 月第 1 版
印　　次：2017 年 7 月第 1 次印刷
定　　价：68.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：tianhf@phei.com.cn。

前　言

本书内容

本书分为三篇，共有 12 章。

第一篇　大数据的基本概念和技术

第 1 章　绪论，描述大数据的时代背景与国家大数据战略，探讨大数据的概念和特性，重点阐述大数据支撑体系，包括数据采集、存储、分布式计算和应用，并讨论大数据人才特点与能力要求。

第 2 章　Hadoop 大数据关键技术，详细介绍大数据系统涉及的主流技术，主要包括数据采集与生成、数据分布式存储、分布式计算框架、数据分析与挖掘等方面的技术和工具。

第二篇　Hadoop 大数据平台搭建与基本应用

第 3 章　Linux 操作系统与集群搭建，介绍 Linux 集群的安装、Java 开发包 JDK 的安装，以及集群的配置方法。

第 4 章　HDFS 安装与基本应用，介绍 Hadoop HDFS 的架构、工作原理，以及 Hadoop 安装、配置、启动和程序的运行。

第 5 章　MapReduce 与 YARN，介绍 MapReduce 的工作原理，描述 MapReduceV2（也就是 YARN）的架构和执行流程。本章重点介绍如何设计 MapReduce 程序，给出了在 Eclipse 中实现 Java 语言 MapReduce 程序的具体过程。

第 6 章　Hive 和 HBase 的安装与应用，主要介绍 Hive 和 HBase 的安装配置和应用方法，同时也介绍 MySQL 和 ZooKeeper 的安装与应用。

第 7 章　Sqoop 和 Kafka，介绍 Sqoop 和 Kafka 组件的安装及其基本应用方法。

第 8 章　Spark 集群安装与开发环境配置，介绍 Spark 架构及其工作原理，详细介绍 Spark 开发环境的安装与配置，包括热门的 IntelliJ IDEA 集成开发环境的安装与基本应用。

第 9 章　Spark 应用基础，介绍 Spark 程序的运行模式和应用设计方法，通过编写计算圆周率 Pi、基于随机森林模型的贷款风险预测 Scala 程序，展示了在集成开发环境 IDEA 中编写 Spark 程序的流程。

第三篇　大数据处理与项目开发

第 10 章　交互式数据处理，介绍如何利用 Hive 进行大数据处理和分析。Hive 是建立在 Hadoop MapReduce 基础上的数据仓库工具，用户借助 SQL 语句，可完成很多处理和分析，因此，对实际工作者有很大帮助。

第 11 章　协同过滤推荐系统，介绍推荐算法的基本概念和应用，展示基于 Spark 的机器学库 MLlib 实现的协同推荐应用。

第 12 章　销售数据分析系统，通过一个完整的销售数据分析系统设计，展示如何利用 Hadoop 的各种组件开发实际的大数据应用系统。本章运用到的组件包括 HDFS、MySQL、Eclipse、Phoenix、HBase、WebCollector、Sevlet、Tomcat 等，所展示的数据和应用均来自真实场景，对读者有较高参考价值。

本书特点

把原理、架构、运行流程分析与实际应用融合起来介绍，融合性阐述框架优于单纯的原理分析，因为原理最终要付诸应用。

本书高度重视实践能力的培养，对系统安装、配置和应用过程给出了十分详细的描述，所有实验都是基于实际完成的操作介绍的，并配有现场截图，为读者展示了真实、详尽、可重现的场景，十分方便读者自学和钻研。

与很多大数据技术书籍不同，本书突出了数据处理本身，深入介绍了如何运用技术进行实际的数据分析，所采用的数据样本来自生产一线，所展示的项目具有实用的参考价值，读者掌握这些技术之后，就可以开始进行项目开发了。

本书的读者群

本书十分适合初学者入门和进阶。

本书也可供那些已经学习过 Hadoop 组件技术，但希望全面、系统地理解并掌握实际应用的读者参考。

本书对从事大数据项目开发的专业人员也有参考价值，书中所描述的 Hadoop 组件应用中遇到的各种问题及其解决办法，十分实用。

本书特别适合自学，读者完全可以利用本书给出的资源和示例，一步一步地完成各项操作和应用，体验一种登堂入室的成就感。

致谢

感谢大数据时代，感谢开源社区，感谢 Apache 基金会，感谢 Google，感谢所有关心和热爱大数据的人们！

作者在创作本书中借鉴了中科普开（北京）科技公司的部分培训资源，在此谨表示衷心的感谢。特别感谢中南大学郑瑾副教授，本书的部分内容使用了她编撰的书稿。由衷地感谢王建新教授、李建彬教授、张祖平教授，他们耐心地审阅了本书，提出了中肯的意见和建议。非常感谢电子工业出版社田宏峰编辑，他细心专业的工作方式，给作者留下深刻印象，并为本书的高质量印装提供了保障。

由于作者水平有限，本书的错误和疏漏在所难免，恳请广大读者提出宝贵意见和建议。联系邮箱：djhuang@csu.edu.cn。

作　者

2017 年 6 月于长沙

目　录

第一篇　大数据的基本概念和技术

第二篇 Hadoop 大数据平台搭建与基本应用

第三篇 数据处理与项目开发术

第一篇

大数据的基本概念和技术

第 1 章 绪　论

最早提出“大数据”时代到来的全球知名咨询公司麦肯锡称：“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”。

本章主要分析大数据的时代背景与国家大数据战略，给出大数据的概念，并分析其特性，重点介绍大数据技术支撑体系，包括数据采集、存储、分布式计算和应用，最后简要讨论大数据人才特点与能力要求。

1.1 时 代 背 景

1.1.1 全球大数据浪潮

为什么最近几年里大数据变得如此引人注目？大数据到底有多大？

一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满 1.68 亿张 DVD；发出的邮件有 2940 亿封之多；发出的社区帖子达 200 万个（相当于《时代》杂志 770 年的文字量）；卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万。

目前，全球数据量已经从 TB（1024 GB=1 TB）级别跃升到 PB（1024 TB=1 PB）、EB（1024 PB=1 EB）乃至 ZB（1024 EB=1 ZB）级别。国际数据公司（IDC）的研究结果表明，2008 年全球产生的数据量为 0.49 ZB，2009 年的数据量为 0.8 ZB，2010 年增长到 1.2 ZB，2011 年的数量更是高达 1.82 ZB，相当于全球每人产生 200 GB 以上的数据。而

到 2016 年，人类生产的所有印刷材料的数据量是 300 PB，全人类历史上说过的所有话的数据量大约是 5 EB。IBM 的研究称，整个人类文明所获得的全部数据中，有 90%是过去两年内产生的。而到了 2020 年，全世界所产生的数据规模将达到今天的 44 倍。

这样的趋势将会持续下去。我们现在还处于大数据的初级阶段，随着技术的进步，我们的设备、交通工具和迅速发展的“可穿戴”科技将能互连互通。科技的进步已经使创造、采集和管理信息的成本降至十年前的六分之一，而从 2005 年起，用在硬件、软件、人才及服务之上的商业投资也增长了整整 50%，达到了 4000 亿美元。

正如《纽约时报》2012 年 2 月的一篇专栏文章所称，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉。哈佛大学社会学教授加里金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

越来越多的政府、企业等机构开始意识到数据正在成为组织最重要的资产，数据分析能力正在成为组织的核心竞争力。

2012 年 3 月 22 日，美国政府宣布投资 2 亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。美国政府将数据定义为“未来的新石油”，并表示一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来，对数据的占有和控制甚至将成为陆权、海权、空权之外的另一种国家核心资产。

联合国也在 2012 年发布了大数据政务白皮书，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，人们如今可以使用极为丰富的数据资源，来对社会经济进行前所未有的实时分析，帮助政府更好地响应社会和经济运行。

最为积极的还是众多的 IT 企业。麦肯锡在一份名为“大数据，是下一轮创新、竞争和生产力的前沿”的专题研究报告中提出，“对于企业来说，海量数据的运用将成为未来竞争和增长的基础”，该报告在业界引起广泛反响。麦肯锡的报告发布后，大数据迅速成为了计算机行业争相传诵的热门概念，也引起了包括金融界在内的各行各业的高度关注。随着互联网技术的不断发展，数据本身是资产，这一点在业界已经形成共识。如果说云计算为数据资产提供了保管、访问的场所和渠道，那么如何盘活数据资产，使其为国家治理、企业决策乃至个人生活服务，则是大数据的核心议题，也是云计算内在的灵魂和必然的升级方向。事实上，全球互联网巨头都已意识到了“大数据”时代，数据的重要意义，包括谷歌、苹果、惠普、IBM、微软在内的全球 IT 巨头纷纷通过收购“大数据”相关厂商来实现技术整合，可见其对“大数据”的重视。