教育部—邦飞产学合作协同育人项目
大数据与人工智能技术丛书

Hadoop+Spark 大数据技术 微课版

◎ 刘彬斌 主编　李柏章 周磊 李永富 编著

- 以实例讲解知识点：Hadoop与Spark集群搭建、MapReduce、Hive、HBase、Sqoop、Flume、Scala、RDD、Spark SQL、KafKa、Spark-ML等
- 完整的大数据实战项目解析

微课版
250分钟
教学视频

清華大學出版社

教育部—邦飞产学合作协同育人项目
大数据与人工智能技术丛书

Hadoop+Spark 大数据技术 微课版

◎ 刘彬斌 主编　李柏章 周磊 李永富 编著

清華大學出版社
北京

内 容 简 介

本书从初学者角度出发，通过丰富的实例，详细介绍了大数据开发环境和基本知识点的应用。全书内容包括：大数据系统基础篇、Hadoop 技术篇、Spark 技术篇和项目实战篇。大数据系统基础篇讲解 Linux 的安装、Linux 的使用和在 Linux 系统上安装并使用 MySQL；Hadoop 技术篇讲解 Hadoop 集群的搭建、Hadoop 两大核心的原理与使用、Hadoop 生态圈的工具原理与使用（Hive、HBase、Sqoop、Flume 等）；Spark 技术篇讲解 Spark 集群的搭建、Scala 语言、RDD、Spark SQL、Spark streaming 和机器学习；项目实战篇将真实的电力能源大数据分析项目作为实战解读，帮助初学者快速入门。

本书所有知识点都结合具体实例和程序讲解，便于读者理解和掌握。本书适合作为高等院校计算机应用、大数据技术及相关专业的教材；也适合作为大数据开发入门者的自学用书，可快速提高开发技能。

图书在版编目（CIP）数据

Hadoop+Spark 大数据技术：微课版/刘彬斌主编. —北京：清华大学出版社，2018
（大数据与人工智能技术丛书）
ISBN 978-7-302-51427-5

Ⅰ. ①H…　Ⅱ. ①刘…　Ⅲ. ①数据处理软件－高等学校－教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 242160 号

责任编辑：付弘宇　薛　阳
封面设计：刘　键
责任校对：焦丽丽
责任印制：杨　艳

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
课 件 下 载：http://www.tup.com.cn，010-62795954

印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张：**22.5　　**字　数：**537
版　　次：2018 年 11 月第 1 版　　**印　次：**2018 年 11 月第 1 次印刷
印　　数：1～2000
定　　价：69.00 元

产品编号：079663-01

前 言

随着信息技术的不断发展，以及物联网、社交网络、移动终端等新兴技术与服务的不断涌现和广泛应用，数据种类日益增多，数据的规模急剧增大，大数据时代已悄然来临。由于大数据对政府决策、商业规划和危险预防等方面所起的重大作用，大数据逐渐成为一种重要的国家战略性资源，受到政府、能源及信息领域的普遍关注。大数据的多样性（Variety）、规模性（Volume）和高速性（Velocity）等特点，使得传统的数据存储、管理、分析技术已经无法满足大数据的处理要求。

时至今日，无论你是来自互联网、通信行业，还是来自金融业、服务业或零售业，相信你都不会对大数据感到陌生。调查显示，32.5%的公司正在搭建大数据平台，29.5%的公司已经在生产环境实践大数据技术，并有成功的用例/产品；24.5%的公司已经做了足够的了解，开发准备就绪；基本不了解的只占调查对象的 13.5%。根据某知名数据公司的调查数据，目前国内市场的 IT 人才缺口已经高达几十万，到 2025 年，这一数字还会增加至 200 万，“尤其是大数据技术方面的人才”。在智联、58 同城等大型招聘网站最新发布的招聘职位中，大数据相关岗位占比已经超过 50%，薪酬比软件工程师高 10%以上。由此可见，大数据人才的培养是一份重大的责任和使命。

1．高校大数据人才培养的背景

（1）高校教育中，大数据人才培养存在起步晚、规模化不足的问题，而且高校学生从大学入学到研究生毕业需要相当长的一段时间。本书从实用的角度出发，为高校快速培养大数据人才提供可行性。

（2）如前文所述，大数据人才紧缺的现象在全球越来越突出。在此背景下，本书旨在弥补高校大数据教材的不足，以模拟真实生产环境为教学目标，为企业培养“到岗就能用”的大数据实用型人才。

（3）经济社会的高速发展，对 IT 产业（尤其是软件产业）提出了更高的要求，对大数据开发人才从数量和质量方面提出了更高的要求。

（4）教育技术的进步和移动互联网时代的到来，打破了高校进行知识传播的技术壁垒。大量的资本和风险投资涌进 IT 培训产业。达内、传智播客等实体 IT 培训机构，开课吧、慕课网、极客网等在线 IT 培养机构纷纷引入先进的教学理念、强大的技术支持，再加上商业化运作，对高校 IT 人才培养带来巨大的挑战和竞争压力。

（5）教学环境的变化。教室、实验室硬件配置齐全，实现了高速稳定的互联网接入，笔记本电脑和手机等互联网接入设备日渐普及，这些都为先进教学理念和教学模式（如微课）的实施提供了硬件和软件上的准备。

（6）教育参与者。教师应该树立“教育就是服务”的教育观念，贯彻工程教育的教育理念，从注重“教师教什么”转移到“学生学到了什么”。学生作为“数字原住民”，对新鲜事物、新技术、新教学方式（人性化学习、泛在学习等）有着天然的渴望，教师应尽量多利用新的教学手段，提升课程的吸引力。

综上所述，IT 产业、软件技术以及软件人才培养中的教学理念、教学模式、教学环境、教学对象等因素的发展变化倒逼着高校进行教学改革，教师必须围绕以上因素进行教学创新，传统教材形式的革新也势在必行。

2．本书内容

全书内容分为大数据系统基础、Hadoop 技术、Spark 技术和项目实战 4 部分。其中，Linux 是学习大数据技术的基础，先从 Linux 入手，打下坚实的基础，之后才能更好地学习 Hadoop 和 Spark。4 部分内容分别介绍如下。

大数据系统基础篇通过大数据概述、Linux 系统安装、Linux 系统基础命令、Shell 编程和 MySQL 数据操作，为以后编程奠定坚实的基础。

Hadoop 技术篇以 Hadoop 生态圈为中心，详细介绍 Hadoop 高可用集群搭建、HDFS 技术、MapReduce 技术、Hive 技术，为读者学习大数据开发技术提供便利，并以实用的方式简单介绍 HBase、Sqoop、Flume 工具的使用，使读者在精通一门技术的前提下，能扩展了解相关知识，真正成为一专多能的专业型人才。

Spark 技术篇从 Spark 概述、Scala 语言、环境搭建、RDD 核心技术、Spark SQL 和机器学习等多方面讲解 Spark 大数据的开发，从基础的 Scala 语言开始学习，并以 Hadoop 环境为基础搭建 Spark 大数据集群，从最基础、最常用、最容易理解的思路出发，帮助读者逐步掌握 Spark 大数据技术。

项目实战篇从真实项目“电力能源大数据分析”中抽取一部分业务作为实战解读，通过简洁的流程讲解，使读者了解大数据项目开发的整个过程。

3．本书特色

本书不是对相关原理进行纯理论的阐述，而是提供了丰富的上机实践操作和范例程序，极大地降低了读者学习大数据技术的门槛。对于需要直接上机实践的读者而言，本书更像是一本大数据学习的实践上机手册。书中首先展示了如何在单台 Windows 系统上通过 VirtualBox 虚拟机安装多台 Linux 虚拟机，而后建立 Hadoop 集群，再建立 Spark 开发环境。搭建这个上机实践的平台并不限制于单台实体计算机，主要是考虑个人读者上机实践的实际条件和环境。对于有条件的公司和学校，参照这个搭建过程，同样可以将实践平台搭建在多台实体计算机上。

搭建好大数据上机实践的软硬件环境之后，就可以在各个章节的学习中结合本书提供的范例程序逐一设置、修改、调试和运行，从中体会大数据实践应用的真谛——对大数据进行高效的“加工”，萃取大数据中蕴含的“智能和知识”，实现数据的“增值”，并最终将其应用于实际工作或者商业项目中。

4．本书的使用

第 1 篇讲解 Linux 系统和 Linux 系统上的软件应用。本篇是学习大数据技术的第一步，就如同你要学习 Java 开发，必须先学会操作 Windows 系统一样。

第 2 篇讲解 Hadoop 大数据技术。Hadoop 大数据集群要求在 CentOS 6.9 版本的系统上搭建，JDK 版本为 JDK 1.8，Hadoop 版本为 Hadoop 2.6.5，Zookeeper 版本为 Zookeeper 3.4.10。

第 3 篇讲解在 Hadoop 大数据技术的基础上搭建 Spark 环境，所以读者在学习本篇内容之前，需要熟悉第 2 篇中的 Hadoop 大数据集群搭建的内容。

第 4 篇讲解电力大数据项目，是基础 HDFS 的离线分析项目，读者需要掌握 Java 知识、

Hadoop 技术和 Web 前端知识。

5. 作者与致谢

本书由刘彬斌主编。参与本书的编写、资料整理、书稿校对、课件制作等工作的还有李永富、李柏章、周磊、汪磊等。另外，感谢清华大学出版社相关编辑专业和严谨的工作，为本书的顺利出版提供了宝贵的意见，并付出了辛勤的劳动。

编　者

2018 年 3 月

目 录

第 1 篇 大数据系统基础

第3篇 Spark技术

第 4 篇 项目实战

第1篇　大数据系统基础

第1章 大数据概述

大数据是目前最热门的技术之一。顾名思义，大数据的“大”是指相对于传统意义上的数据量来说，它的数据体量很大，这些数据信息是海量信息，且在动态变化和快速增长。多样性：大数据是异构的，且是多样性的；价值密度大，大量的数据包含了大量的数据信息，对这些看似不相关的数据信息进行对未来趋势和模式的可预测分析具有巨“大”的价值。速度：对于这些海量数据，需要快速获取信息，这就需要奇“快”的速度。大数据的这些特征能带来预测未来、决胜千里的能力，给生活特别是企业带来巨大的价值。

要想实现大数据的这些价值，就需要大数据技术的支撑，包括数据的存储、数据的管理、数据的计算和建模、数据分析以及数据的可视化。

为了解决这些问题，Hadoop 和 Spark 应运而生，作为大数据处理系统，它们利用分布式文件存储系统、高可靠性以及高可用的分布式计算框架可以构建庞大的高可靠、高效率的集群，为快速地从大体量、庞杂的数据中获取有价值的数据提供了强有力的技术支撑。本书立足于生产实践的实际应用，从最基础的 Linux 平台的学习入手，再到 Hadoop 体系和 Spark 体系。本书包含了想要走向或者正在走向大数据之路的读者所要掌握的绝大部分基础知识。

1.1 数据的产生与发展

随着计算机和信息技术的迅猛发展，人们从工业时代迈向了互联网时代。随着人们获取信息的方式和方法的改变，各行业应用系统的规模迅速扩大，所产生的数据呈井喷式增长。很多应用产生的数据量每天达到数十 TB、数百 TB 甚至数 PB 的规模，各行业所应用的大数据已远远超出了传统计算机和信息技术的处理能力。因此，需要迫切寻求有效的大数据处理技术、方法和手段。

2003 年，Google 发布了有关 Google File System 的论文，详细阐述了一个可扩展的分布式文件系统，用于大型的、分布式对大量数据进行访问的应用。在此基础上，Google 又陆续发表了有关 MapReduce 和 Bigtable 的论文。

2005 年，基于 Google 发表的前两篇论文的理论，Hadoop 应运而生。Hadoop 最初只是雅虎公司用来解决网页搜索问题的一个项目，后因其技术的高效性，被 Apache Software Foundation 公司引入并成为开源应用，为其发展和广泛使用打下了坚实基础。

2008 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