普通高等院校数据科学与大数据技术专业“十三五”规划教材

Spark大数据

SPARK BIG DATA | PROGRAMMING FOUNDATIONS

编程基础(Scala版)

高建良 盛 羽 编著

本书配有教学课件和源代码

中南大学出版社
www.csupress.com.cn

普通高等院校数据科学与大数据技术专业“十三五”规划教材

Spark大数据

SPARK BIG DATA | PROGRAMMING FOUNDATIONS

编程基础(Scala版)

高建良 盛 羽 编著

本书配有教学课件和源代码

中南大学出版社
www.csupress.com.cn
·长 沙·

图书在版编目（CIP）数据

Spark 大数据编程基础：Scala 版 / 高建良，盛羽编著. --长沙：中南大学出版社，2019.3

ISBN 978-7-5487-3574-8

Ⅰ.①S… Ⅱ.①高… ②盛… Ⅲ.①数据处理—教材 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2019)第 033042 号

Spark 大数据编程基础

Spark DASHUJU BIANCHENG JICHU

（Scala 版）

（Scala BAN）

高建良　盛　羽　编著

□**责任编辑**　韩　雪

□**责任印制**　易红卫

□**出版发行**　中南大学出版社

社址：长沙市麓山南路　　邮编：410083

发行科电话：0731-88876770　　传真：0731-88710482

□**印　　装**　长沙雅鑫印务有限公司

□**开　　本**　787×1092　1/16　□**印张** 24.25　□**字数** 616 千字

□**版　　次**　2019 年 3 月第 1 版　□2019 年 3 月第 1 次印刷

□**书　　号**　ISBN 978-7-5487-3574-8

□**定　　价**　65.00 元

普通高等院校数据科学与大数据技术专业“十三五”规划教材

编委会

总序

Preface

随着移动互联网的兴起，全球数据呈爆炸式增长，目前90%以上的数据是近年产生的，数据规模大约每两年翻一番，而随着人工智能下物联网生态圈的形成，数据的采集、存储及分析处理、融合共享等技术需求都能得到响应，各行各业都在体验大数据带来的革命，“大数据时代”真正来临。这是一个产生大数据的时代，更是需要大数据力量的时代。

大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据从产生、分析到利用提出了前所未有的新要求。高等教育只有转变观念，更新方法与手段，寻求变革与突破，才能在大数据与人工智能的信息大潮面前立于不败之地。据预测，中国近年来大数据相关人才缺口达200万人，全世界相关人才缺口更超过1000万人之多。我国教育部门为了适应社会发展需要，率先于2016年开始正式开设“数据科学与大数据技术”本科专业及“大数据技术与应用”专科专业，近几年，全国形成了申报与建设大数据相关专业的热潮。随着专业建设的深入，大家发现了一个共同的难题：没有成系列的大数据相关教材。

中南大学作为首批申报大数据专业的学校，2015年在我校计算机科学与技术专业设立大数据方向时，信息科学与工程学院院领导便意识到系列教材缺失的严重问题，因此院领导规划由课程团队在教学的同时积累素材，形成面向大数据专业知识体系与能力体系、老师自己愿意用、同学觉得买得值、关联性强的系列教材。经过两年的准备，针对2017年《教育部办公厅关于推荐新工科研究与实践项目的通知》的精神，中南大学出版社组织对系列教材文稿进行相应的打磨，最终于2018年底出版“普通高等院校数据科学与大数据技术专业‘十三五’规划教材”。

该套系列教材具有如下特点：

1. 本套教材主要参照“数据科学与大数据技术”本科专业的培养方案，综合考虑专业的来源，如从计算机类专业、数学统计类专业以及经济类专业发展而来，同时适当兼顾了专科类偏向实际应用的特点。

2. 注重理论联系实际，注重能力培养。该系列教材中既有理论教材也有配套的实践教程，力图通过理论或原理教学、案例教学、课堂讨论、课程实验与实训实习等多个环节，训练学生掌握知识、运用知识分析并解决实际问题的能力，以满足学生今后就业或科研的需求，同时兼顾“全国工程教育专业认证”对学生基本能力的培养要求与复杂问题求解能力的要求。

3. 在规范教材编写体例的同时，注重写作风格的灵活性。本套系列教材中每本书的内容都由教学目的、本章小结、思考题或练习题、实验要求等组成。每本教材都配有 PPT 电子教案及相关的电子资源，如实验要求及 DEMO、配套的实验资源管理与服务平台等。本套系列教材的文本层次分明、逻辑性强、概念清晰、图文并茂、表达准确、可读性强，同时相关配套电子资源与教材的相关性强，形成了新媒体式的立体型系列教材。

4. 响应了教育部“新工科”研究与实践项目的要求。本套教材从专业导论课开始设立相关的实验环节，作为知识主线与技术主线把相关课程串接起来，力争让学生尽早具有培养自己动手能力的意识、综合利用各种技术与平台的能力。同时为了避免新技术发展太快、教材纸质文字内容容易过时的问题，在相关技术及平台的叙述与实践中，融合了网络电子资源容易更新的特点，使新技术保持时效性。

5. 本套丛书配有丰富的多媒体教学资源，将扩展知识、习题解析思路等内容做成二维码放在书中，丰富了教材内容，增强了教学互动，有利于提高学生的学习积极性与主动性。

本套丛书吸纳了数据科学与大数据技术教育工作者多年的教学与科研成果，凝聚了作者们的辛勤劳动，同时也得到了中南大学等院校领导和专家的大力支持。我相信本套教材的出版，对我国数据科学与大数据技术专业本科、专科教学质量的提高将有很好的促进作用。

桂卫华

2018 年 11 月

前言

Foreword

大数据被称为“未来的新石油”，那么如何开采“新石油”是各个领域处理大数据面临的核心问题。工欲善其事，必先利其器。大数据编程为处理大数据提供了最有效的“器”，本书全面介绍了大数据编程基础。Apache Spark 已经成为大数据处理的首选平台，因此本书的大数据编程将基于 Spark 平台进行。

本书成体系地介绍了 Spark 大数据编程技术。本书分为三个部分共 10 章，从介绍 Spark 开发环境开始，再以 Spark 编程入门基础为承接，最后具体到每一个 Spark 编程组件。这三部分内容由浅入深自成体系，可以方便地学习 Spark 编程的每个具体知识点。

第一部分包含第 1 ~ 2 章，讲述了 Spark 的开发环境。其中，第 1 章对 Spark 的背景和运行架构进行了概述；第 2 章对 Spark 开发环境的搭建进行了详细介绍。这是学习后续章节的基础。

第二部分包含第 3 ~ 5 章，讲述了 Spark 编程入门基础部分，重点介绍了 Scala 编程基础和弹性分布式数据集(resilient distributed dataset，RDD)编程。本书采用 Scala 编程语言，第 3 章和第 4 章分别介绍了 Scala 语言基础和 Scala 面向对象编程。RDD 是 Spark 对数据的核心抽象，第 5 章介绍了 RDD 编程。

第三部分包含第 6 ~ 10 章，讲述了 Spark 编程组件部分，重点介绍了 Spark SQL、Spark Streaming、Spark GraphX、Spark ML 四个组件的编程。其中，第 6 章介绍了 Spark SQL，它可以高效地处理结构化数据；第 7 章介绍了 Spark Streaming，它可以高效地处理流式数据；第 8 章介绍了 Spark GraphX，它可以高效地处理图数据；第 9 章和第 10 章介绍了 Spark ML，它们分别以 Spark 机器学习原理和 Spark 机器学习模型为重点进行介绍。

本书在编写过程中力求深入浅出、重点突出、简明扼要，尽可能方便不同专业背景和知识层次的读者阅读。本书编写过程中，中南大学研究生杜宏亮、田玲、熊帆、高俊、吕腾飞、蒋志怡、应晓婷等做了大量的资料收集整理、书稿校对等工作，在此，对这些同学的辛勤工作表示感谢。

本书配套的官方网站是 http://aibigdata.csu.edu.cn，免费提供全部课件资源、源代码和数据集。相关资料也可以从中南大学出版社的网站下载。

另外，本书部分内容参考了大量的公开资料和网络上的资源，对他们的工作致以衷心的

感谢。需要指出的是，数据科学与大数据技术是一个全新的专业，因此编写一本完美的大数据编程教材绝非易事。由于水平有限，书中难免存在疏漏或者错误，希望广大读者不吝赐教。如有任何建议、意见或者疑问，请及时联系作者，以期在后续版本中加以改进和完善。

编　者

2019 年 3 月

目 录

Contents

第 1 章　Spark 概述

Spark 在大数据处理中发挥着越来越重要的作用，本章将对 Spark 进行概述。首先介绍 Spark 的发展背景和特点；其次介绍 Spark 的生态系统，即 Spark 的核心组件；再次对 Spark 的运行架构进行介绍，为理解 Spark 的运行流程与原理提供帮助；最后结合代码实例对 Spark 的编程进行简单介绍。

1.1　Spark 的背景

Spark 起源于一个学术研究项目，仅用几年时间就成为了大数据领域应用最广泛的项目之一。本节简要介绍 Spark 发展史和 Spark 的特点。

1.1.1　Spark 发展史

Spark 是一种快速、通用、可扩展的大数据分析引擎。它以高效的方式处理分布式数据集，为分布式数据集的处理提供了一个有效的框架。2009 年，Spark 在加州大学伯克利分校 AMP 实验室(Algorithms, Machines and People Lab)形成雏形。2010 年，Spark 正式开放源代码。2013 年，Spark 进入 Apache 孵化器项目，成为孵化项目，不久便成为顶级项目并不断发展和完善。2018 年 11 月，最新的 Spark 2.4.0 版本。Spark 的演进路线如图 1-1 所示，可见其发展速度非常之快。

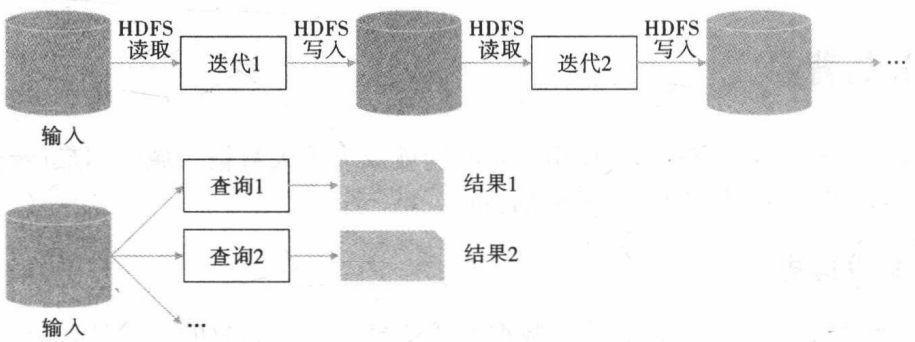
图 1-1　Spark 演进路线图

Spark 的发展态势迅猛，已经成为当前大数据分析的主流平台之一。Spark 作为一个开源项目，越来越多的开发人员参与其中作出贡献，共同推动 Spark 继续快速地发展。

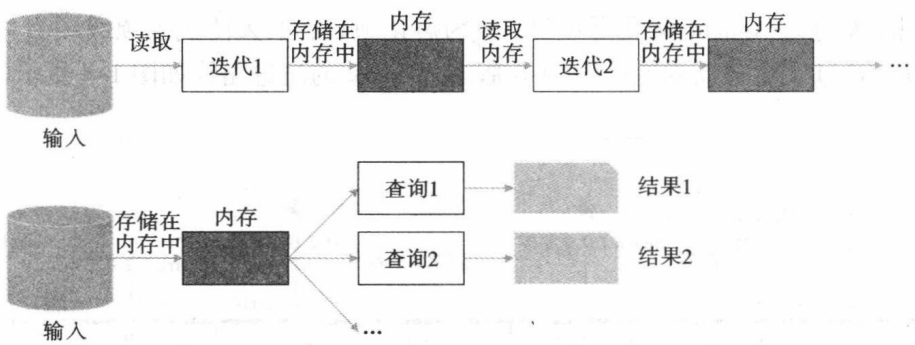
1.1.2　Spark 的特点

Spark 的一个含义是“电光火石”，表示运行速度非常快。Spark 官网提供的数据表明，如果数据是从内存中读取，它的速度可以达到 Hadoop MapReduce 的 100 多倍。

Spark 默认情况下迭代的中间结果放在内存中，后续的运行作业利用这些结果进一步计算，如图 1-2 所示。而 Hadoop 的计算结果都需要存储到磁盘中，后续的计算需要从磁盘中读取之前的计算结果。由于从内存中读取数据要比从磁盘读取数据快得多，所以 Spark 运行速度会快得多，尤其是在需要多次迭代计算的情况下。另外，Spark 基于 JVM(Java virtual machine)进行了优化。Hadoop 中的每次 MapReduce 操作，启动一个 Task 便会启动一次 JVM，这是基于进程的操作；而 Spark 的每次 MapReduce 操作是基于线程的，只在启动 Executor 时启动一次 JVM，内存的 Task 操作是在线程复用的。每次启动 JVM 的时间大约需要几秒甚至十几秒，那么当 Task 多了，Hadoop 就比 Spark 花了更多时间。



(a) Hadoop MapReduce执行流程



(b) Spark执行流程

图 1-2　Spark 与 Hadoop 的执行流程比较

Spark 在很多方面借鉴 Hadoop MapReduce，并克服了 Hadoop MapReduce 的很多不足，它具有以下优点：

(1)Spark 运算效率高。MapReduce 在数据 Shuffle 之前，要花费大量时间排序，而 Spark 不需要对所有情景进行排序，由于采用有向无环图(directed acyclic graph，DAG)执行计划，每次输出结果可以缓存在内存中，所以迭代运算效率高。MapReduce 的计算结果保存在磁盘