

北京课工场教育科技有限公司 出品



新技术技能人才培养系列教程

大数据开发实战系列

基于Hadoop与Spark的

大数据开发实战



肖睿 丁科 吴刚山 / 主编

鄢长青 张琪 / 副主编



视频资源



案例素材



交流社区



学习 APP

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

北京课工场教育科技有限公司 出品



新技术技能人才培养系列教程
大数据开发实战系列

基于Hadoop与 Spark的

大数据开发实战

肖睿 丁科 吴刚山 / 主编
鄢长青 张琪 / 副主编



人民邮电出版社

北京

图书在版编目 (C I P) 数据

基于Hadoop与Spark的大数据开发实战 / 肖睿, 丁科,
吴刚山主编. — 北京: 人民邮电出版社, 2018. 4
新技术技能人才培养系列教程
ISBN 978-7-115-47764-4

I. ①基… II. ①肖… ②丁… ③吴… III. ①数据处
理软件—教材 IV. ①TP274

中国版本图书馆CIP数据核字(2018)第010078号

内 容 提 要

大数据技术让我们以一种前所未有的方式, 对海量数据进行分析, 从中获得有巨大价值的产品和服务, 最终形成变革之力。本书围绕 Hadoop 和 Spark 两个主流大数据技术进行讲解, 主要内容包括 Hadoop 环境配置、Hadoop 分布式文件系统 (HDFS)、Hadoop 分布式计算框架 MapReduce、Hadoop 资源调度框架 YARN 与 Hadoop 新特性、Hadoop 分布式数据库 HBase、数据仓库 Hive、大数据离线处理辅助系统、Spark Core、Spark SQL、Spark Streaming 等知识。

本书紧密结合实际应用, 运用大量案例说明和实践, 提炼含金量十足的开发经验。另外, 本书配以多元的学习资源和支持服务, 包括视频教程、案例素材下载、学习交流社区、讨论组等学习内容, 为读者带来全方位的学习体验。

本书适合作为计算机、大数据相关专业的教材使用, 也适合具有一定 Linux、Java 开发经验, 并且想从事大数据开发的人员使用, 也可作为大数据分析 with 运维人员的参考用书。

-
- ◆ 主 编 肖 睿 丁 科 吴刚山
副 主 编 鄢长青 张 琪
责任编辑 祝智敏
责任印制 马振武
 - ◆ 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>
北京鑫正大印刷有限公司印刷
 - ◆ 开本: 787×1092 1/16
印张: 24 2018 年 4 月第 1 版
字数: 566 千字 2018 年 4 月北京第 1 次印刷
-

定价: 66.80 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

大数据开发实战系列

编 委 会

主 任：肖 睿

副 主 任：相洪波 韩 露

委 员：孙 苹 李 娜 张惠军 杨 欢

潘贞玉 庞国广 张德平 王丙辰

课 工 场：周 嵘 孙正哲 刘 尧 董 海

崔建瑞 冯娜娜 李 真 陈 璇

尚永祯 于学杰 陈 燕 刁志星

刘校锋 吉志星 曹紫涵 霍荣慧

序 言

丛书设计

准备好了吗？进入大数据时代！大数据已经并将继续影响人类生产生活的方方面面。2015年8月31日，国务院正式下发《关于印发促进大数据发展行动纲要的通知》。企业资本则以BAT互联网公司为首，不断进行大数据创新，实现大数据的商业价值。本丛书根据企业人才的实际需求，参考以往学习难度曲线，选取“Java+大数据”技术集作为学习路径，首先从Java语言入手，深入学习理解面向对象的编程思想、Java高级特性以及数据库技术，并熟练掌握企业级应用框架——SSM、SSH，熟悉Java Web应用和Hadoop大数据开发，积累企业实战经验，通过实战项目对大型分布式应用有所了解和认知，为“大数据核心技术系列”的学习打下坚实基础。本丛书旨在为读者提供一站式实战型大数据应用开发学习指导，帮助读者踏上由开发入门到大数据实战的“互联网+大数据”开发之旅！

丛书特点

1. 以企业需求为设计导向

满足企业对人才的技术需求是本丛书的核心设计原则，为此课工场大数据开发教研团队，通过对数百位BAT一线技术专家进行访谈、上千家企业人力资源情况进行调研、上万个企业招聘岗位进行需求分析，从而实现对技术的准确定位，达到课程与企业需求的强契合度。

2. 以任务驱动为讲解方式

丛书中的技能点和知识点都由任务驱动，读者在学习知识时不仅可以知其然，而且可以知其所以然，帮助读者融会贯通、举一反三。

3. 以实战项目来提升技术

每本书均增设项目实战环节，以综合运用每本书的知识点，帮助读者提升项目开发能力。每个实战项目都有相应的项目思路指导、重难点讲解、实现步骤总结和知识点梳理。

4. 以“互联网+”实现终身学习

本丛书可配合使用课工场APP进行二维码扫描，观看配套视频的理论讲解和案例操作。同时课工场(www.kgc.cn)开辟教材配套版块，提供案例代码及作业素材下载。此外，课工场也为读者提供了体系化的学习路径、丰富的在线学习资源以及活跃的学习社区，欢迎广大读者进入学习。

读者对象

1. 大中专院校学生
2. 编程爱好者
3. 初中级程序开发人员
4. 相关培训机构的老师和学员

致谢

本丛书由课工场大数据开发教研团队编写。课工场是北京大学旗下专注于互联网人才培养的高端教育品牌。作为国内互联网人才教育生态系统的构建者，课工场依托北京大学优质的教育资源，重构职业教育生态体系，以学员为本，以企业为基，构建“教学大咖、技术大咖、行业大咖”三咖一体的教学矩阵，为学员提供高端、实用的学习内容！

读者服务

读者在学习过程中如遇疑难问题，可以访问课工场官方网站（www.kgc.cn），也可以发送邮件到 ke@kgc.cn，我们的客服专员将竭诚为您服务。

感谢您阅读本丛书，希望本丛书能成为您踏上大数据开发之旅的好伙伴！

“大数据开发实战系列”丛书编委会

前 言

本书以 Hadoop 和 Spark 为核心，阐述了基于这两种通用大数据处理平台的应用开发技术。

在 Hadoop 生态圈中，从 HDFS 初识分布式存储系统；以 MapReduce 详解分步式计算的步骤；利用 HBase 分析适合非结构化数据存储的分布式数据库；利用 Hive 分析将 SQL 查询转化为分布式计算的过程；并结合项目案例“音乐排行榜”练习 Hadoop 核心技能点的运用；同时，介绍了几种离线处理系统中常用的辅助工具。

在 Spark 生态圈中，从 Scala 开始介绍多范式编程；并从 Spark Core、Spark SQL、Spark Streaming 三个方面来分析对比 Hadoop 生态圈中的分布式计算、Hive、流式计算的可替换方案和它们各自的优势。

技能训练

- 掌握 Hadoop 运行环境的部署。
- 掌握大数据文件在 HDFS 中的存储。
- 掌握 MapReduce 编程模型以及 MapReduce 应用开发方法。
- 掌握 YARN 的运行原理。
- 掌握 HBase 数据库的操作方法。
- 掌握 Hive 数据仓库的操作方法。
- 掌握常用离线处理辅助系统 Sqoop 和 Azkaban 的用法。
- 掌握 Scala 基本编程方法。
- 掌握 Spark RDD 创建与操作。
- 掌握 DataFrame 编程方法。
- 掌握 Spark Streaming 对 Socket、HDFS 数据进行流式处理的方法。
- 了解 Spark Streaming 与 Flume、Kafka 的整合。

设计思路

本书共 12 章，内容包括 Hadoop 初体验、Hadoop 分布式文件系统、Hadoop 分布式计算框架、Hadoop 新特性、Hadoop 分布式数据库、Hadoop 综合实战——音乐排行榜、数据仓库 Hive、大数据离线处理辅助系统、Spark 基础、Spark Core、Spark SQL 和 Spark Streaming。具体内容安排如下。

- 第 1 章是对 Hadoop 的总体概述，介绍大数据基本概念、Hadoop 生态圈、

Hadoop 与大数据的关系以及 Hadoop 安装部署的详细步骤。

- 第 2 章是对 HDFS 的介绍，主要包括 HDFS 的体系结构、Shell 操作以及通过 Java API 实现访问。
- 第 3 章是对 MapReduce 分布式计算框架的讲解，包括 MapReduce 的编程模型、编写和运行 MapReduce 程序。
- 第 4 章是对 Hadoop 新的资源调度框架 YARN 及 Hadoop 新特性的讲解，以及如何实现 Hadoop 高可用集群。
- 第 5 章是对 HBase 数据库的讲解，介绍 HBase 的安装及其使用方法。
- 第 6 章通过案例“音乐排行榜”的实现，对前面各章的技能点做一个阶段回顾与总结，介绍如何通过 HDFS、MapReduce 与 HBase 的结合使用完成 Hadoop 离线批处理应用开发。
- 第 7 章是对 Hive 的讲解，介绍如何使用类似于 SQL 查询的方式来执行 MapReduce 计算。
- 第 8 章介绍 Sqoop、Azkaban 这两种在开发离线处理系统时常用的辅助工具。
- 第 9 章是对 Spark 的基本介绍，包括 Spark 的安装与运行、Spark 的开发语言 Scala。
- 第 10 章是对 Spark 的核心 RDD 的详解，介绍 Spark Core 的编程模型以及 Spark 应用程序的开发。
- 第 11 章是对 Spark SQL 的详解，包括常用的 SQL on Hadoop 框架、Spark SQL 的编程方法以及 Spark SQL 对多种外部数据源的操作。
- 第 12 章是对 Spark Streaming 的详解，包括 Spark Streaming 核心概念、常用的流处理系统，以及使用 Spark Streaming 进行流处理应用的开发。

章节导读

- 技能目标：本章要达成的学习目标，可以作为检验学习效果的标准。
- 本章任务：本章要完成的学习内容及要求，通过任务描述引导读者思考，进而引导读者全面了解章节内容。
- 案例代码：通过代码让读者掌握如何应用本章讲解的技能点。
- 本章总结：本章内容的概括和总结。
- 本章练习：针对本章学习内容的补充性练习，用于加强对本章知识的理解和运用。

本书由课工场大数据开发教研团队编写，参与编写的还有丁科、吴刚山、鄢长青、张琪等院校老师。由于编者水平有限，书中不妥或错误之处在所难免，殷切希望广大读者批评指正！

编者

2017 年 12 月

关于引用作品的版权声明

为了方便读者学习，促进知识传播，本书选用了一些知名网站的相关内容作为学习案例。为了尊重这些内容所有者的权利，特此声明，凡在书中涉及的版权、著作权、商标权等权益均属于原作品版权人、著作权人、商标权人。

为了维护原作品相关权益人的权益，现对本书选用的主要作品的出处给予说明（排名不分先后）。

序号	选用的网络作品	版权归属
1	Hadoop	hadoop.apache.org
2	HBase	hbase.apache.org
3	ZooKeeper	zookeeper.apache.org
4	Hive	hive.apache.org
5	Sqoop	sqoop.apache.org
6	Spark	spark.apache.org

以上列表并未全部列出本书所选用的作品。在此，本书创作团队衷心感谢所有原作品的相关版权权益人及所属公司对职业教育的大力支持！

目 录

序言

前言

关于引用作品的版权声明

第1章 Hadoop初体验 1

任务 1 初识大数据	2
1.1.1 大数据基本概念	2
1.1.2 大数据带来的挑战	3
任务 2 初识 Hadoop	3
1.2.1 Hadoop 概述	4
1.2.2 Hadoop 生态圈	6
1.2.3 Hadoop 应用案例	8
任务 3 安装 Hadoop 平台	9
1.3.1 安装虚拟机	10
1.3.2 安装 Linux 系统	13
1.3.3 安装 Hadoop 伪分布式环境	30
本章总结	34
本章练习	34

第2章 Hadoop分布式文件系统 35

任务 1 HDFS 入门	36
2.1.1 认识 HDFS	36
2.1.2 HDFS 基础	38
2.1.3 HDFS 架构	40
任务 2 HDFS 基本操作	41
2.2.1 使用 HDFS shell 访问	41
2.2.2 使用 Java API 访问	45

任务 3 HDFS 运行原理	48
2.3.1 HDFS 读写流程	49
2.3.2 HDFS 副本机制	50
2.3.3 HDFS 负载均衡	51
2.3.4 HDFS 机架感知	52
任务 4 HDFS 高级知识	53
2.4.1 Hadoop 序列化机制	53
2.4.2 SequenceFile	58
2.4.3 MapFile	63
本章总结	65
本章练习	66

第3章 Hadoop分布式计算框架 67

任务 1 认识 MapReduce 编程模型	68
3.1.1 MapReduce 基础	68
3.1.2 MapReduce 编程模型	69
3.1.3 MapReduce 词频统计编程实例	70
任务 2 MapReduce 应用开发	75
3.2.1 MapReduce 输入 / 输出类型	75
3.2.2 MapReduce 输入格式	76
3.2.3 MapReduce 输出格式	78
3.2.4 Combiner 操作	79
3.2.5 Partitioner 操作	82
3.2.6 自定义 RecordReader	86
任务 3 MapReduce 高级应用	92
3.3.1 使用 MapReduce 实现 join 操作	93
3.3.2 使用 MapReduce 实现排序	100
3.3.3 使用 MapReduce 实现二次排序	103
3.3.4 使用 MapReduce 合并小文件	108
本章总结	113
本章练习	113

第4章 Hadoop新特性 115

任务 1 初识 YARN	116
4.1.1 YARN 产生背景	116
4.1.2 YARN 简介	117

4.1.3	YARN 架构设计	119
任务 2	了解 HDFS 新特性	121
4.2.1	HDFS NameNode 高可用机制	121
4.2.2	HDFS NameNode Federation	129
4.2.3	HDFS Snapshots	130
4.2.4	HDFS REST API	134
4.2.5	DistCp 工具	134
任务 3	了解 YARN 新特性	135
4.3.1	ResourceManager 自动重启	135
4.3.2	ResourceManager 高可用机制	136
	本章总结	139
	本章练习	139

第5章 Hadoop分布式数据库 141

任务 1	认识 HBase	142
5.1.1	HBase 简介	142
5.1.2	HBase 体系结构	143
5.1.3	HBase 数据模型	145
5.1.4	HBase 的安装	148
任务 2	HBase Shell 操作	155
5.2.1	HBase Shell 简介	155
5.2.2	HBase Shell 的使用	156
任务 3	HBase 编程	162
5.3.1	开发 HBase 应用程序	162
5.3.2	HBase 数据存储管理 API	163
	本章总结	175
	本章练习	175

第6章 Hadoop综合实战——音乐排行榜 177

任务 1	MapReduce 与 HBase 的集成	178
6.1.1	MapReduce 与 HBase 的集成环境	178
6.1.2	批量数据导入 (Bulk Loading)	181
任务 2	HBase MapReduce API	182
6.2.1	HBase MapReduce API 简介	182
6.2.2	TableMapper 的使用	183
6.2.3	TableReducer 的使用	195

任务3 实现音乐排行榜	197
6.3.1 程序的结构与实现	198
6.3.2 HBase 数据库设计优化	205
6.3.3 MapReduce 全局共享数据	205
本章总结	207
本章练习	207

第7章 数据仓库Hive 209

任务1 Hive 基础	210
7.1.1 认识 Hive	210
7.1.2 Hive 架构设计	211
7.1.3 Hive 与 Hadoop	212
7.1.4 Hive 与传统关系型数据库	212
7.1.5 Hive 数据存储模型	213
7.1.6 Hive 部署	213
任务2 掌握 Hive 操作	214
7.2.1 Hive DDL	214
7.2.2 Hive DML	217
7.2.3 Hive shell	222
任务3 Hive 高级应用	223
7.3.1 Hive 函数	224
7.3.2 Hive 调优策略	227
本章总结	232
本章练习	232

第8章 大数据离线处理辅助系统 233

任务1 认识并使用数据迁移框架 Sqoop	234
8.1.1 Sqoop 简介	234
8.1.2 使用 Sqoop 导入 MySQL 数据到 HDFS	239
8.1.3 使用 Sqoop 导出 HDFS 数据到 MySQL	246
8.1.4 使用 Sqoop 导入 MySQL 数据到 Hive	248
8.1.5 Sqoop Job	250
任务2 使用 Azkaban 实现 workflow 调度	250
8.2.1 Azkaban 概述	250
8.2.2 Azkaban 环境部署	252
8.2.3 Azkaban 应用实例	256

本章总结.....	259
本章练习.....	259

第9章 Spark基础 261

任务1 Spark 入门	262
9.1.1 Spark 简介	262
9.1.2 Spark 优势	262
9.1.3 Spark 生态圈	264
任务2 Scala 基础	267
9.2.1 Scala 简介	268
9.2.2 Scala 函数定义	271
9.2.3 Scala 面向对象操作	272
9.2.4 Scala 集合的使用	275
9.2.5 Scala 高阶函数	278
任务3 编译 Spark	281
9.3.1 下载 Spark 源码	281
9.3.2 编译 Spark 源码	282
任务4 Spark 初体验	284
9.4.1 Spark 环境部署	284
9.4.2 spark-shell	285
本章总结.....	286
本章练习.....	286

第10章 Spark Core 287

任务1 Spark RDD	288
10.1.1 RDD 介绍	288
10.1.2 RDD 的创建	289
10.1.3 RDD 的转换算子	291
10.1.4 RDD 的动作算子	293
10.1.5 RDD 的依赖关系	295
任务2 RDD 高级应用	297
10.2.1 RDD 缓存机制	297
10.2.2 共享变量	300
10.2.3 Spark 架构设计	302
任务3 基于 RDD 的 Spark 应用程序开发	303
10.3.1 准备工作	303

10.3.2 词频计数实例	304
10.3.3 年龄统计实例	308
本章总结	309
本章练习	309

第11章 Spark SQL 311

任务1 认识 Spark SQL	312
11.1.1 SQL	312
11.1.2 SQL on Hadoop 框架	312
11.1.3 Spark SQL 简介	314
任务2 Spark SQL 编程基础	315
11.2.1 Spark SQL 编程入口	315
11.2.2 DataFrame 基础	317
11.2.3 DataFrame 编程实例	318
任务3 Spark SQL 编程进阶	325
11.3.1 Spark SQL 操作外部数据源	325
11.3.2 Spark SQL 函数	329
11.3.3 Spark SQL 调优	332
本章总结	334
本章练习	335

第12章 Spark Streaming 337

任务1 流处理框架及 Spark Streaming	338
12.1.1 流处理框架简介	338
12.1.2 Spark Streaming 简介	340
任务2 使用 Spark Streaming 编程	343
12.2.1 Spark Streaming 核心	343
12.2.2 Spark Streaming 编程实例	348
任务3 Spark Streaming 高级应用	352
12.3.1 使用 Spark Streaming 整合 Flume	353
12.3.2 使用 Spark Streaming 整合 Kafka	356
12.3.3 Spark Streaming 优化策略	361
本章总结	363
本章练习	363

附录 365

Hadoop 初体验

技能目标

- ❖ 了解大数据和 Hadoop 是什么
- ❖ 掌握 Hadoop 的核心构成
- ❖ 了解 Hadoop 生态圈
- ❖ 掌握虚拟机、CentOS 和 Hadoop 的安装

本章任务

学习本章,需要完成以下 3 个工作任务。记录学习过程中遇到的问题,通过自己的努力或访问 kgc.cn 解决。

任务 1: 初识大数据

了解大数据的基本概念和基本特征,以及大数据带给企业的挑战有哪些。

任务 2: 初识 Hadoop

了解 Hadoop 是什么,掌握 Hadoop 的核心构成,了解 Hadoop 生态圈中各个组件的功能。

任务 3: 安装 Hadoop 平台

掌握虚拟机、CentOS、Hadoop 的安装。



任务1 初识大数据

关键步骤如下。

- 了解大数据是什么。
- 了解大数据的特征。
- 了解大数据带给企业哪些方面的挑战。

1.1.1 大数据基本概念

1. 大数据概述

相信大家经常会在各种场合听到“大数据”这个词，被誉为数据仓库之父的 Bill Inmon 早在 20 世纪 90 年代就将大数据挂在嘴边了。那么到底什么是大数据呢？这是我们在本章要了解的。

我们现在生活的时代是一个数据时代，近年来随着互联网的高速发展，每分每秒都在产生数据，那么产生的这些数据如何进行存储和相应的分析处理呢？各大公司纷纷研发和采用一批新技术来应对日益庞大的数据处理需求，主要包括分布式文件系统、分布式计算框架等，这些都是我们需要学习和掌握的。

《互联网周刊》对大数据的定义为：“大数据”的概念远不止大量的数据（TB）和处理大量数据的技术，或者所谓的“4 个 V”之类的简单概念，而是涵盖了人们在大规模数据的基础上可以做的事情，这些事情在小规模数据的基础上是无法实现的。换句话说，大数据让我们以一种前所未有的方式，通过对海量数据进行分析，来获得有巨大价值的产品和服务，或深刻的洞见，最终形成变革之力。

2. 大数据特征

(1) 数据量大 (Volume)

随着网络技术的发展和普及，每时每刻都会产生大量的数据。在我们的日常生活中，比如说在电商网站购物、在直播平台看直播、在线阅读新闻等，都会产生很多的日志，汇在一起每分每秒产生的数据量将是非常巨大的。

(2) 类型繁多 (Variety)

大数据中最常见的类型是日志，除了日志之外常见的还有音频、视频、图片等。由于