大数据科学丛书
BIG DATA SCIENCE

Scala

语言基础与开发实战

王家林 等 编著

机械工业出版社
CHINA MACHINE PRESS

大数据科学丛书

Scala 语言基础与开发实战

王家林　管祥青　等编著

机 械 工 业 出 版 社

本书分为基础篇、中级篇、高级篇及分布式框架四大部分，从 Scala 零基础入门，步步深入，引导读者由浅入深地学习 Scala 及其应用。本书从手把手指引读者搭建 Scala 语言开发环境开始，详细介绍了 Scala 的语法基础，以代码实例形式分别讲解了 Scala 面向对象开发及函数式编程；在此基础上进一步深入讲解了 Scala 的中高级语法特性，包括模式匹配、集合、类型参数、高级类型、隐式转化及各语法特性在 Spark 源码中的应用解析，并引出 Scala 的 Actor 模型及其应用详解。本书还详细介绍了以 Scala 为基础的两大框架——Akka 和 Kafka。

本书每章开始均有重点介绍，以引导读者有目的、有重点地阅读或查阅。另外，针对不同语法特性的源码及应用解析是本书的另一大特点。

本书适合具备一定编程语言基础、对大数据开发有兴趣的在校学生，同时，对有面向对象编程或函数式编程经验的人员，本书也可以作为开发实例的参考书籍。

图书在版编目（CIP）数据

Scala 语言基础与开发实战/王家林等编著 . —北京：机械工业出版社，2016.6
（大数据科学丛书）
ISBN 978-7-111-54169-1

Ⅰ. ①S… Ⅱ. ①王… Ⅲ. ①JAVA 语言 - 程序设计 Ⅳ. ①TP312

中国版本图书馆 CIP 数据核字（2016）第 151317 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：王 斌 责任编辑：王 斌
责任校对：张艳霞
印刷（ 装订）

2016 年 7 月第 1 版 · 第 1 次印刷
184mm × 260mm · 26.5 印张 · 640 千字
0001 - 3000 册
标准书号：ISBN 978-7-111-54169-1
定价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
电话服务
社 服 务 中 心：(010)88361066
销 售 一 部：(010)68326294
销 售 二 部：(010)88379649
读者购书热线：(010)88379203
网络服务
门户网：http://www.cmpbook.com
教材网：http://www.cmpedu.com
封面无防伪标均为盗版

前　言

采用 Scala 语言编写实现的，大数据领域最火爆的计算框架 Spark（其实 Spark 在 Apache 下的数据处理领域也是最火爆的计算框架），正在以迅雷不及掩耳之势快速发展。很少有一门语言能够像 Scala 这样，因其作为大数据框架 Spark 的核心和首选开发语言而爆发式地普及起来。Spark 本身起源于 2009 年，是美国加州大学伯克利分校 AMP 实验室的一个研究性项目，于 2010 年开源，在 2014、2015 年大数据领域软件排名中，Spark 都以绝对优势遥遥领先！虽然基于 Spark 平台可以采用 Scala、Java、Python、R 等 4 种语言开发，但据 Spark 官方统计，2014 年和 2015 年全世界范围内基于 Spark 开发采用最多的语言一直都是 Scala。另外，在大数据领域越来越多的其他技术框架，例如 Kafka 等也都把 Scala 作为实现和开发语言。因此，为了打好大数据领域学习的基础，本书面向广大 Scala 爱好者和大数据开发者，以实战为主导，并用实战与理论相结合的方式来帮助读者学习 Scala 语言。

从 2012 年美国政府的“大数据研发计划”，到 2015 年我国国务院发布的《促进大数据发展行动纲要》，可以说，大数据已经迎来了它的黄金时代。本书紧跟时代潮流，除了讲解 Scala 语言之外，还额外挑选了当前在大数据领域中应用非常广泛 Akka 和 Kafka 两大框架进行讲解，并且详细讲解了 Scala 语言在其中的应用。Akka 是一个在 JVM 上构建高并发、分布式和可快速恢复的消息驱动应用的工具包；Kafka 是高产出的分布式消息系统，它实现了生产者和消费者之间的无缝连接，实现了处理速度快、高可扩展性的分布式实时系统。

本书编写的主线是以 Scala 实战实例为主导，由浅入深，从 Scala 的基础篇、中级篇直至高级篇，对 Scala 各个知识点加以详细分析并给出相应的实例及解析。然后更进一步地引入分布式框架篇，针对当前大数据领域使用非常广泛的大分布式框架 Akka 和 Kafka，通过介绍 Scala 语言在开发分布式框架时的实战案例，为读者进一步学习大数据领域各个框架打好基础。

参与本书编写的有王家林、段智华、管祥青、徐奔、张敏、徐香玉等。

本书能顺利出版，离不开出版社编辑们的大力支持与帮助，在此表示诚挚的感谢。

非常感谢本书的技术审核徐香玉为审核本书技术相关内容所做出的努力。

在阅读本书的过程中，若发现任何问题或有任何疑问，可以加入本书的阅读群（QQ：418110145）提出讨论，会有专人帮忙答疑。同时，该群中也会提供本书所用实例代码。

如果读者想要了解或者学习更多大数据相关技术，可以通过以下方式参与互动交流：

关注 DT 大数据梦工厂微信公众号：DT_Spark 及 QQ 群：163728659，或者通过扫描下方二维码咨询，也可以通过 YY 客户端登录 68917580 永久频道直接体验。

我的新浪微博是 http://wei bo. com/ilovepains/，也欢迎大家在微博上进行互动。
由于时间仓促，书中难免存在不妥之处，请读者谅解，并提出宝贵意见。

王家林
2016. 1. 18 日于北京

目　录

中　级　篇

高 级 篇

分布式框架篇

基础篇

作为本书的开篇，本篇将从零开始引导读者由浅入深，循序渐进，依照图例的指引一步一步地真正动手搭建 Scala 编程环境及大数据实验环境。本篇由最基本的 Scala 语法开始讲起，让读者在理解语法的基础上动手编程实验，逐步学会 Scala 基本编程，为后续学习 Scala 中高级篇的高级语法，以及基于 Scala 的三大主流分布式大数据框架打下坚实的基础。

本篇共3章，第1章介绍了 Scala 零基础入门的环境搭建及 Scala 的基础语法知识；第2章基于基础语法，以实例分析的形式详解 Scala 面向对象的编程开发；第3章讲解了 Scala 的函数式编程，并以实例分析的形式详解其中的高阶函数应用。

第 1 章　Scala 零基础入门

本章将通过详细的图例，清晰地介绍了 Scala 实验环境的搭建（包括 Scala 安装、Hadoop 安装及 MapReduce 词频统计、Spark 安装、Scala IDE 安装），Scala 的基础知识（变量、函数、流程控制、异常处理），Scala Tuple、Array、Map 与文件操作，以及 Scala apply 应用等内容。

Section 1.1 Scala 概述

曾经有人问 Java 的创始人高斯林（James Gosling）这样一个问题："除了 Java 语言以外，您现在还使用 JVM（Java Virtual Machine，Java 虚拟机）上的哪种编程语言?"他毫不犹豫地说是 Scala。Scala 是一门集成面向对象和函数式特性的语言，它的创始人是 Martin Ordersky。Martin Ordersky 也是 Javac 编译器的作者，他曾将泛型引入到 Java 语言中。因为 Java 是强约束的语言，出于 Java 本身设计上的局限性，诸如函数式编程等特性在 Java 8 之前都没有实现，因此 Martin Ordesky 设想基于 JVM 和 Java 类库，创建一个比 Java 更高级的语言。他于 2001 年开始基于 Funnel 语言设计 Scala，在 2003 年年底 Scala 发布正式（基于 Java 平台）版本。

截至 2016 年 1 月，Scala 2.12 系列的最新版本是 Scala 2.12.0 – M3，Scala 2.11 系列的最新版本是 Scala 2.11.7。2016 年，Scala 2.11 系列预计于一季度实现 Scala 2.11.8 版本，在三季度实现 Scala 2.11.9 版本；Scala 2.12 系列预计在二季度重点实现 Scala 2.12.0 – RC1 版本，在 2.12 系列的两个里程碑 Scala 2.12.0 – M4 及 Scala 2.12.0 – M5 Scala 版本中，Scala 编译器将会有很大的变化，包括 Lambda 表达式和使用新的编码方式，充分利用 Java 8 的新特性，以及 Scala 的优化等。

Scala 是一个运行在 Java 虚拟机环境（JVM）中将面向对象编程和函数式编程的最佳特性完美结合起来具有强大类型系统的，优雅、简洁的语言。

Scala 语言的优势如下：

1）Scala 是面向对象编程语言：在 Scala 中，所有预先定义的类型是对象，所有用户定义的类型也是对象，每个操作都是方法调用。

2）Scala 是函数式编程语言：在 Scala 中，函数是"头等公民"，函数可以被当成参数传递给其他函数，也可以当成结果返回或保存为变量。Scala 编写的代码简洁、优雅，同时又具备丰富的表达能力，在实际开发过程中推荐使用大量不可变的变量及无副作用的函数进行代码编写；Scala 支持匿名函数、高阶函数、柯里化函数、函数嵌套的应用；相对于指令式编程语言而言，Scala 可以使用更少的代码量来实现相同的功能，充分体现出 Scala 语言简洁、精炼、优雅的特点。