HZ BOOKS
华章IT

科大讯飞和百分点资深大数据专家实践经验结晶，秉承老庄哲学，从开发、数据分析、计算、管理和性能优化等多角度系统、深度地讲解了Spark的核心技术与高级应用

大数据
技术丛书

Spark Core Technology and Advanced Applications

Spark核心技术与高级应用

于俊　向海　代其锋　马海平◎著

机械工业出版社
China Machine Press

技术丛书

Spark Core Technology and Advanced Applications

Spark核心技术与高级应用

于俊　向海　代其锋　马海平◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Spark 核心技术与高级应用 / 于俊等著 . —北京：机械工业出版社，2015.12
（大数据技术丛书）

ISBN 978-7-111-52354-3

I. S… II. 于… III. 数据处理软件 IV. TP274

中国版本图书馆 CIP 数据核字（2015）第 305661 号

Spark 核心技术与高级应用

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：高婧雅
责任校对：董纪丽
印　　刷：北京诚信伟业印刷有限公司
版　　次：2016 年 1 月第 1 版第 1 次印刷
开　　本：186mm × 240mm 1/16
印　　张：19.75
书　　号：ISBN 978-7-111-52354-3
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
客服热线：（010）88379426 88361066
投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259
读者信箱：hzit@hzbook.com

Preface 前　言

上善若水，水善利万物而不争。

数据一如水，无色无味，非方非圆，以百态存于自然，于自然无违也。绵绵密密，微则无声，巨则汹涌；与人无争却又容纳万物。

生活离不开水，同样离不开数据，我们被数据包围，在数据中生活，在数据中入梦和清醒。

某夜入梦时分，趴桌而眠，偶遇庄周那只彩色翅膀的蝴蝶飞入梦中，在数据上翩翩起舞；清醒时分，蝴蝶化身数据，继续在眼前飞舞，顿悟大数据之哲学。本书从《道德经》和《庄子》各精选 10 句名言，并结合大数据相关内容，对名言加以讲解，引导大家以老庄的思考方式来认识大数据的内涵，探求老子道之路和庄子智慧之路。

为什么要写这本书

2014 年春天，我所在的知识云团队聚焦大数据，调研过程中，深深感觉到国内资料匮乏，可供参考的资料仅是 Spark 官方文档。团队人员英文水平参差不齐，Spark 官方文档门槛比较高，学习起来困难重重。

当时和几个同事一起，对 Spark 官方文档进行了翻译，参考了机械工业出版社《 Spark 快速数据处理》的小册子，编了一本《 Spark 数据处理》内部文档，解决了一部分问题，并将 Spark 应用推向具体业务。在实际业务中，相比传统的数据处理，尤其是实时处理和迭代计算，MapReduce 在 Spark 面前显得苍白无力。随着 Spark 的应用越来越多，深深感觉到《 Spark 数据处理》内部文档的不足，遗憾的是，一直没有时间进行补充和完善，俨然成了一块心病。

2014 年 9 月，在机械工业出版社华章公司福川兄的指导下，开始重点思索：Spark 解决哪些问题、优势在哪里、从业人员遇到哪些困难、如何解决这些困难等问题，并得到了吴爱华、吕劲松、代其锋、马海平、向海、陈明磊等几位同事的支持。怀着一颗“附庸风雅”之

心，我决定和大家一起写一本具有一定实战价值的 Spark 方面的书籍。

当前大数据从业者，有数据科学家、算法专家、来自互联网的程序员、来自传统行业的工程师等，无论来自哪里，作为新一代轻量级计算框架，Spark 集成 Spark SQL、Spark Streaming、MLlib、GraphX、SparkR 等子框架，都提供了一种全新的大数据处理方式，让从业者的工作变得越来越便捷，也让机器学习、数据挖掘等算法变得“接地气”。数据科学家和算法专家越来越了解社会，程序员和工程师有了逆袭的机会。

本书写作过程中，Spark 版本从 1.0 一直变化到 1.5，秉承大道至简的主导思想，我们尽可能地按照 1.5 版本进行了统筹，希望能抛砖引玉，以个人的一些想法和见解，为读者拓展出更深入、更全面的思路。

本书只是一个开始，大数据之漫漫雄关，还需要迈步从头越。

本书特色

本书虽是大数据相关书籍，但对传统文化进行了一次缅怀，吸收传统文化的精华，精选了《道德经》和《庄子》各 10 句名言，实现大数据和文学的有效统一。结合老子的“无为”和庄子的“天人合一”思想，引导读者以辩证法思考方式来认识大数据的内涵，探求老子道之路和庄子智慧之路，在大数据时代传承“老庄哲学”，让中国古代典籍中的瑰宝继续发扬下去。

从技术层面上，Spark 作为一个快速、通用的大规模数据处理引擎，凭借其可伸缩、基于内存计算等特点，以及可以直接读写 HDFS 上数据的优势，实现了批处理时更加高效、延迟更低，已然成为轻量级大数据快速处理的统一平台。Spark 集成 Spark SQL、Spark Streaming、MLlib、GraphX、SparkR 等子框架，并且提供了全新的大数据处理方式，让从业者的工作变得越来越便捷。本书从基础讲起，针对性地给出了实战场景；并围绕 DataFrame，兼顾在 Spark SQL 和 Spark ML 的应用。

从适合读者阅读和掌握知识的结构安排上讲，分为“基础篇”、“实战篇”、“高级篇”、“扩展篇”四个维度进行编写，从基础引出实战，从实战过渡高级，从高级进行扩展，层层推进，便于读者展开讨论，深入理解分析，并提供相应的解决方案。

本书的案例都是实际业务中的抽象，都经过具体的实践。作为本书的延续，接下来会针对 Spark 机器学习部分进行拓展，期待和读者早点见面。

读者对象

（1）对大数据非常感兴趣的读者

伴随着大数据时代的到来，很多工作都变得和大数据息息相关，无论是传统行业、IT 行

业以及移动互联网行业，都必须要了解大数据的概念，对这部分人员来说，本书的内容能够帮助他们加深对大数据生态环境和发展趋势的理解，通过本书可以了解 Spark 使用场景和存在价值，充分体验和实践 Spark 所带来的乐趣，如果希望继续学习 Spark 相关知识，本书可以作为一个真正的开始。

（2）从事大数据开发的人员

Spark 是类 Hadoop MapReduce 的通用并行计算框架，基于 MapReduce 算法实现的分布式计算，拥有 Hadoop MapReduce 所具有的优点，并且克服了 MapReduce 在实时查询和迭代计算上较大的不足，对这部分开发人员，本书能够拓展开发思路，了解 Spark 的基本原理、编程思想、应用实现和优缺点，参考实际企业应用经验，减少自己的开发成本，对生产环境中遇到的技术问题和使用过程中的性能优化有很好的指导作用。

（3）从事大数据运维的人员

除了大数据相关的开发之外，如何对数据平台进行部署、保障运行环境的稳定、进行性能优化、合理利用资源，也是至关重要的，对于一名合格的大数据运维人员来说，适当了解 Spark 框架的编程思想、运行环境、应用情况是十分有帮助的，不仅能够很快地排查出各种可能的故障，也能够让运维人员和开发人员进行有效的沟通，为推进企业级的运维管理提供参考依据。

（4）数据科学家和算法研究者

基于大数据的实时计算、机器学习、图计算等是互联网行业比较热门的研究方向，这些方向已经有一些探索成果，都是基于 Spark 实现的，这部分研究人员通过本书的阅读可以加深对 Spark 原理与应用场景的理解，对大数据实时计算、机器学习、图计算等技术框架研究和现有系统改进也有很好的参考价值，借此降低学习成本，往更高层次发展。

如何阅读本书

本书分为四篇，共计 20 章内容。

基础篇（第 1 ~ 10 章），详细说明什么是 Spark、Spark 的重要扩展、Spark 的部署和运行、Spark 程序开发、Spark 编程模型以及 Spark 作业执行解析。

实战篇（第 11 ~ 14 章），重点讲解 Spark SQL 与 DataFrame、Spark Streaming、Spark MLlib 与 Spark ML、GraphX、SparkR，以及基于以上内容实现大数据分析、系统资源统计、LR 模型、二级邻居关系图获取等方面的实战案例。

高级篇（第 15 ~ 18 章），深入讲解 Spark 调度管理、存储管理、监控管理、性能调优。

扩展篇（第 19 ~ 20 章），介绍 Jobserver 和 Tachyon 在 Spark 上的使用情况。

其中，第二部分实战篇为本书重点，如果你没有充足的时间完成全书的阅读，可以选择性地进行重点章节的阅读。如果你是一位有着一定经验的资深人员，本书有助于你加深基础概念和实战应用的理解。如果你是一名初学者，请在从基础篇知识开始阅读。

勘误和支持

由于笔者的水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。如果你有更多的宝贵意见，可以通过 Spark 技术 QQ 交流群 435263033，或者邮箱 ustcyujun@163.com 联系到我，期待能够得到你们的真挚反馈，在技术之路上互勉共进。

致谢

感谢 Spark 官方文档，在写作期间提供给我们最全面、最深入、最准确的参考材料。

感谢我亲爱的搭档向海、代其锋、马海平三位大数据专家，在本书写作遭遇困惑的时候，一直互相鼓励，对本书写作坚持不放弃。

感谢知识云团队的范仲毅、杨志远、万文强、张东明、周熠晨、吴增锋、韩启红、吕劲松、张业胜，以及贡献智慧的陈明磊、林弘杰、王文庭、刘君、汪黎、王庆庆等小伙伴，由于你们的参与使本书完成成为可能。

感谢机械工业出版社华章公司的首席策划杨福川和编辑高婧雅，在近一年的时间中始终支持我们的写作，你们的鼓励和帮助引导我们顺利完成全部书稿。

最后，特别感谢我的老婆杨丽静，在宝宝出生期间，因为麻醉意外躺在病房一个月多的时间里，以微笑的生活态度鼓励我，时时刻刻给我信心和力量；还有我可爱的宝宝于潇杨，让我的努力变得有意义。

谨以此书献给我亲爱的家人，知识云团队的小伙伴，以及众多热爱 Spark 技术的朋友们！

于　俊