针对常见问题的快速指南，囊括60多种Spark开发技巧

Spark Cookbook 中文版

Spark Cookbook

[印度] Rishi Yadav 著
顾星竹 刘见康 译

中国工信出版集团
人民邮电出版社 POSTS & TELECOM PRESS

Spark Cookbook 中文版

[印度] Rishi Yadav 著
顾星竹 刘见康 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Spark Cookbook 中文版 / (印) 亚达夫 (Rishi Yadav) 著 ; 顾星竹, 刘见康译. -- 北京 : 人民邮电出版社, 2016.10
ISBN 978-7-115-42966-7

Ⅰ. ①S… Ⅱ. ①亚… ②顾… ③刘… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第189836号

版权声明

◆ 著 [印度] Rishi Yadav
译 顾星竹 刘见康
责任编辑 胡俊英
责任印制 焦志炜

◆ 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
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：12.75
字数：251 千字 2016 年 10 月第 1 版
印数：1 – 3 000 册 2016 年 10 月河北第 1 次印刷

著作权合同登记号 图字：01-2016-2854 号

定价：45.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

Spark 是一个基于内存计算的开源集群计算系统，它非常小巧玲珑，让数据分析更加快速，已逐渐成为新一代大数据处理平台中的佼佼者。

本书内容分为 12 章，从认识 Apache Spark 开始讲解，陆续介绍了 Spark 的使用、外部数据源、Spark SQL、Spark Streaming、机器学习、监督学习中的回归和分类、无监督学习、推荐系统、图像处理、优化及调优等内容。

本书适合大数据领域的技术人员，可以帮助他们更好地洞悉大数据，本书也适合想要学习 Spark 进行大数据处理的人员，它将是一本绝佳的参考教程。

译者简介

顾星竹，南京大学软件学院软件工程硕士，曾就职于 eBay（上海）中国研发中心，现就职于上海平安陆金所，从事大数据开发工作，有丰富的 Hadoop 开发和优化经验，主攻数据挖掘和机器学习，擅长使用 MR、Hive、Impala 和 Spark 等。业余爱好瑜伽、烹饪、写作和翻译。从读研期间开始从事计算机领域的翻译，主攻大数据领域和数据挖掘领域。非常荣幸能有机会翻译本书，也希望大家喜欢本书。

刘见康，南京大学软件学院软件工程学士。曾就职于 eBay（上海）中国研发中心，在数据基础架构平台从事软件开发工作，目前在一家创业公司从事数据挖掘相关的工作。

他爱好广泛，热爱编程，钟爱开源，喜欢研究大数据相关的开源框架，并且对数据、算法、函数式编程、机器学习、软件架构等有一定的探索和学习。目前专注于大数据领域的开发工作，主要使用 Java、Scala、Python 等语言，平时还喜欢研究 Clojure、Haskell、Go 语言，认为不同编程范式能带来不同的编程思想和抽象思维。业余爱好篮球、羽毛球和弹吉他，是名副其实的“书虫”，也是 Kindle 的重度使用者，Coursera 和 edX 蹭课爱好者。

作者简介

Rishi Yadav 拥有 17 年设计和开发企业级应用的经验。他是一位开源软件专家，引领了美国公司的大数据趋势。Rishi 被评为 2014 年 40 位 40 岁以下硅谷杰出工程师之一。他于 1998 年获得著名的印度理工学院（Indian Institute of Technology，IIT）德里分校的学士学位。

大约在 10 年前，Rishi 创办了 InfoObjects，这是一家以深度探索数据为宗旨的企业。

InfoObjects 结合了开源和大数据技术来解决客户的业务问题，并特别注重 Apache Spark 技术。该公司已连续 4 年被列入发展最快公司 5000 强。InfoObjects 也被授予了 2014 及 2015 年度湾区最佳工作地点第一名的桂冠。

Rishi 是一位开源社区贡献者和活跃的博主。

“我要特别感谢我的另一半 Anjali 对我努力挤出时间进行冗长而艰巨的创作的包容，感谢我 8 岁的儿子 Vedant 每天追踪我的进度，感谢 InfoObjects 的 CTO 兼我的合伙人 Sudhir Jangir 领导公司的大数据工作，感谢 Helma Zargarian、Yogesh Chandani、Animesh Chauhan 和 Katie Nelson 的高效运营使得我可以专注于本书的创作，感谢我们的内部审查团队，尤其是 Arivoli Tirouvingadame、Lalit Shravage 和 Sanjay Shroff 的审查，没有你们的支持我绝对写不成本书。同时还要感谢 Marcel Izumi 对于本书插图的贡献。”

——Rishi

审阅者简介

Thomas W. Dinsmore 是一位独立顾问，为分析软件供应商提供产品咨询服务，他拥有30多年为世界各地企业提供分析解决方案的经验，独具动手分析经验和领导分析项目并提供解析结果的能力。

Thomas之前的服务对象包括SAS、IBM、波士顿咨询公司、普华永道和奥纬咨询公司。

Thomas与他人合著了《现代分析方法》（Modern Analytics Methodologies）和《高级分析方法》（Advance Analytics Methodologies），由培生出版社于2014年出版，目前正在和Apress出版社商谈一本关于商业分析的新书的出版事宜。他的名为“大数据分析”的博客地址为www.thomaswdinsmore.com。

“我要感谢Packt出版社的编辑和制作团队全体人员，是你们的不懈努力才能给公众带来这本制作精良的图书。”

——Thomas

连城不仅是一位来自Databricks公司的Apache Spark贡献者，还是一位来自中国的软件工程师。他的主要技术领域包括大数据分析、分布式系统和函数式编程语言。

连城也是《Erlang/OTP并发编程实战》和《Erlang并发编程（第一部分）》中文版的译者。

“我要感谢来自AsiaInfo的Yi Tian帮我检查第6章的部分内容。”

——连城

Amir Sedighi 是一位经验丰富的软件工程师，一位求知若渴的学习狂，一位积极主动

的问题解决者。他的经验涵盖了软件开发领域，包括跨平台开发、大数据处理、数据流、信息检索和机器学习。他是一位在伊朗工作的大数据讲师和专家，拥有软件工程的本科和硕士学位。Amir 目前是 Rayanesh Dadegan Ekbatan 公司的 CEO，该公司是 Amir 历经数年为多家私营公司制定和实施分布式大数据和数据流媒体解决方案后，于 2013 年和他人共同创办的提供数据解决方案的公司。

“我要感谢 Packt 出版社的所有人，正是因为你们的辛苦工作，才能有这么多杰出的书，才能让读者们的职业技能增长。”

——Amir

前言

随着 Hadoop 这个大数据平台的成功，用户的期望也水涨船高，他们既希望解决不同分析问题的功能提高，又希望减少延迟。由此，各类工具应运而生。Apache Spark 这个可以解决所有问题的单一平台也出现在了 Hadoop 的大舞台上。“Spark 一出，谁与争锋”，它终结了需要使用多种工具来完成复杂挑战和学习曲线的局面。通过使用内存进行持久化存储和计算，Apache Spark 避免了磁盘上的中间存储过程并将速度提高了 100 倍，并且提供了一个单一平台用来完成诸如机器学习、实时 streaming 等诸多分析作业。

本书包含了 Apache Spark 的安装和配置，以及 Spark 内核、Spark SQL、Spark Streaming、MLlib 和 GraphX 库的构建方案。

关于本书教程的更多内容，请访问 infoobjects.com/spark-cookbook。

内容概要

第 1 章　开始使用 Apache Spark。介绍了如何在多种环境和集群管理上安装 Spark。

第 2 章　使用 Spark 开发应用。介绍了在不同的 IDE 中使用不同的构建工具开发 Spark。

第 3 章　外部数据源。介绍了如何读写各种数据源。

第 4 章　Spark SQL。带你浏览 Spark SQL 模块，帮助你通过 SQL 接口使用 Spark 的功能。

第 5 章　Spark Streaming。探索 Spark Streaming 库以分析实时数据源（比如 Kafka）

的数据。

第 6 章　机器学习——MLlib。介绍机器学习以及诸如矩阵、向量之类的基本概念。

第 7 章　监督学习之回归——MLlib。连续输出变量的监督学习。

第 8 章　监督学习之分类——MLlib。离散输出变量的监督学习。

第 9 章　无监督学习——MLlib。介绍例如 k-means 等无监督学习。

第 10 章　推荐系统。介绍使用多种技术（比如 ALS）构建推荐系统。

第 11 章　图像处理——GraphX。介绍 GraphX 的多种图像处理算法。

第 12 章　优化及调优。介绍多种 Spark 调优方法和性能优化技术。

阅读须知

你需要使用 InfoObjects 大数据沙箱软件运行本书的例子，该软件可以从 `http://www.infoobjects.com` 下载。

目标读者

只要你是一个希望使用 Apache Spark 更好地洞悉大数据的数据工程师、应用开发工程师或者数据科学家，那本书就是为你而写。

体例

本书将经常出现如下标题——准备工作、具体步骤、工作原理、更多内容和参考资料。

为了更清晰地撰写每一篇教程，我们会使用如下标题。

准备工作

本节介绍教程的大致内容，以及所需的软件和初步配置。

具体步骤

本节包含教程的具体步骤。

工作原理

本节通常由之前章节的具体解释组成。

更多内容

本节由教程相关的更多信息组成，帮助读者了解更多的知识。

参考资料

本节提供有用的链接和其他教程相关的有用信息。

本书约定

在本书中，我们将会用不同的格式区分不同的信息。

文本代码、数据表名、文件夹名、文件名、文件扩展名、路径名、虚拟 URL、用户输入和 Twitter 用户名格式如下所示："安装 Spark 之前，需要安装 Java 并配置好 JAVA_HOME 环境变量。"

代码段如下所示：

```
lazy val root = (project in file("."))
  settings(
    name := "wordcount"
  )
```

任何命令行输入输出如下所示：

```
$ wget http://d3kbcqa49mib13.cloudfront.net/spark-1.4.0-bin-hadoop2.4.tgz
```

新术语和重要的话用粗体显示，例如菜单或对话框显示如下："打开右上角你的账户名下拉框并点击**安全证书**"。

警告或重要信息如此所示。

提示和技巧如此所示。

读者反馈

读者反馈总是很受欢迎的。无论你喜不喜欢本书，都请让我们知道。读者的反馈对我们而言很重要，它会帮助我们提供对读者最有效的信息。

反馈请发送到 feedback@packtpub.com，并请在邮件主题上注明本书书名。

如果你对其中某一主题很有经验，想要写成一本书的话，请访问作者指南 www.packtpub.com/authors。

客户支持

现在，你可以很自豪地说自己是 Packt 图书的拥有者了，我们会最大程度地支持我们图书的客户。

下载本书彩图

我们还为你提供了本书的 PDF 文件，其中有本书截图的彩图。彩图可以帮助你更好地理解输出变化。你可以从 https://www.packtpub.com/sites/default/files/downloads/7061OS_ColorImages.pdf 下载。

勘误

虽然我们已经尽力确保内容的准确性，但是错误是难免的。如果你在我们的书中发现错误并报告给我们，不管是文本的或是代码的，我们将不胜感激。这样可以帮助减少其他读者的困扰并帮助我们提高后续版本的质量。如果你发现任何错误请访问 http://www.packtpub.com/submit-errata 并报告。选择你所读的书，点击勘误提交表格链接并输入你的勘误细节。一旦你的勘误被确认，提交就会被接受，勘误内容将在标题下的勘误章节中呈现。

要查看先前提交的勘误表，请访问 https://www.packtpub.com/books/content/support 并输入书名搜索。所需内容将出现在勘误章节中。

盗版

互联网上的盗版现象是一直存在的问题。在 Packt，我们非常重视版权和许可保护。如

果你在互联网上看到任何形式的非法复制内容，请立即向我们提供网址或站名，以便我们补救。

请通过 copyright@packtpub.com 联系我们提供盗版材料地址。

非常感谢你对我们的作者和我们有价值的内容的保护。

联系我们

如果你对本书有任何方面的疑问，请通过 questions@packtpub.com 联系我们，我们将会竭尽所能地帮助你。

目录

第 1 章
开始使用 Apache Spark

在本章中，我们将介绍安装和配置 Spark，包括如下内容。

- 通过二进制可执行文件安装 Spark。
- 通过 Maven 构建 Spark 源码。
- 在 Amazon EC2 上安装 Spark。
- 在集群上以独立模式部署 Spark。
- 在集群上使用 Mesos 部署 Spark。
- 在集群上使用 YARN 部署 Spark。
- 使用 Tachyon 作为堆外存储层。

1.1 简介

Apache Spark 是一个用于处理大数据工作流的多功能集群计算系统。Spark 在速度、易用性以及分析能力上都强于它的前辈们（如 MapReduce）。

Apache Spark 最初在 2009 年，由加州大学伯克利分校的 AMPLab 实验室研发，在 2010 年按照 BSD 协议实现开源，并在 2013 年转为 Apache 2.0 协议。到 2013 年下半年，Spark 的创始人建立了 Databricks，专注于 Spark 的研发和未来的公开发行。

谈到速度，Spark 大数据工作流的处理可以达到亚秒级别的延迟。为了达到如此低的延迟，Spark 充分利用了内存。在 MapReduce 中，内存仅仅用于实际计算，而 Spark 不仅使用内存进行计算，而且还用于存储对象。

Spark 也提供一个连接各种大数据存储源的统一运行时接口，例如 HDFS、Cassandra、Hbase 和 S3。它同时也提供大量的用于不同的大数据计算任务的顶层库，例如机器学习、SQL 处理、图像处理以及实时数据流。这些库加快了开发速度，可以任意组合。

虽然 Spark 是用 Scala 所写，本书也只关注 Scala 部分的教程，但是 Spark 也支持 Java 和 Python 语言。

Spark 是一个开源社区产品，每个人都是用 Apache 纯开源分布部署，不像 Hadoop，有大量开发商改进的分布部署。

图 1-1 展示了 Spark 的生态圈。

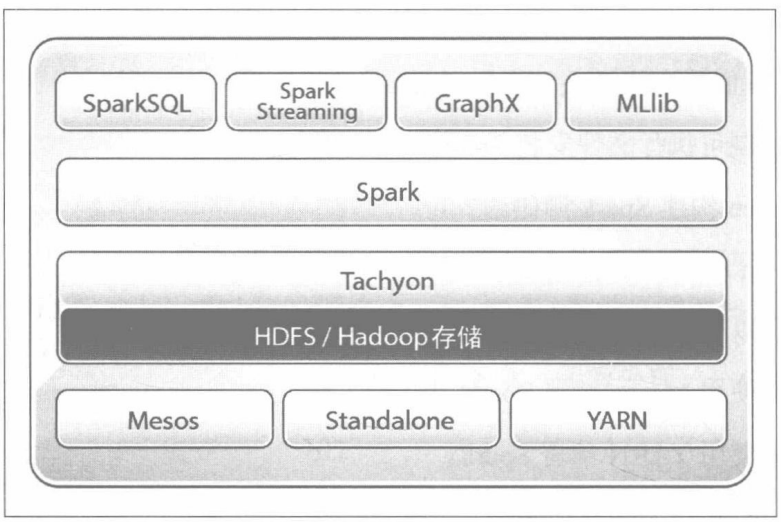

图 1-1 Spark 生态圈

Spark 运行时运行在一系列集群管理器的基础之上，包括 YARN（Hadoop 的计算框架）、Mesos 以及 Spark 自己的被称为独立模式的集群管理器。Tachyon 是一个内存层的分布式文件系统，使得集群架构之间的文件共享速度能够可靠到达内存级别。（译注：Tachyon 现已更名为 alluxio，官网地址：www.alluxio.org。本书的其他部分仍会按照原文写作 Tachyon，后续不再赘述。）简而言之，它是内存上的一个堆外存储层，用于在任务和用户之间分享数据。Mesos 是一个涉及数据中心处理系统的集群管理器。YARN 是一个有着健壮的资源管理特性的 Hadoop 计算框架，Spark 可以与它无缝连接使用。

1.2 使用二进制文件安装 Spark

Spark 既可以通过源码安装也可以通过预编译二进制安装，下载地址为 `http://spark.`