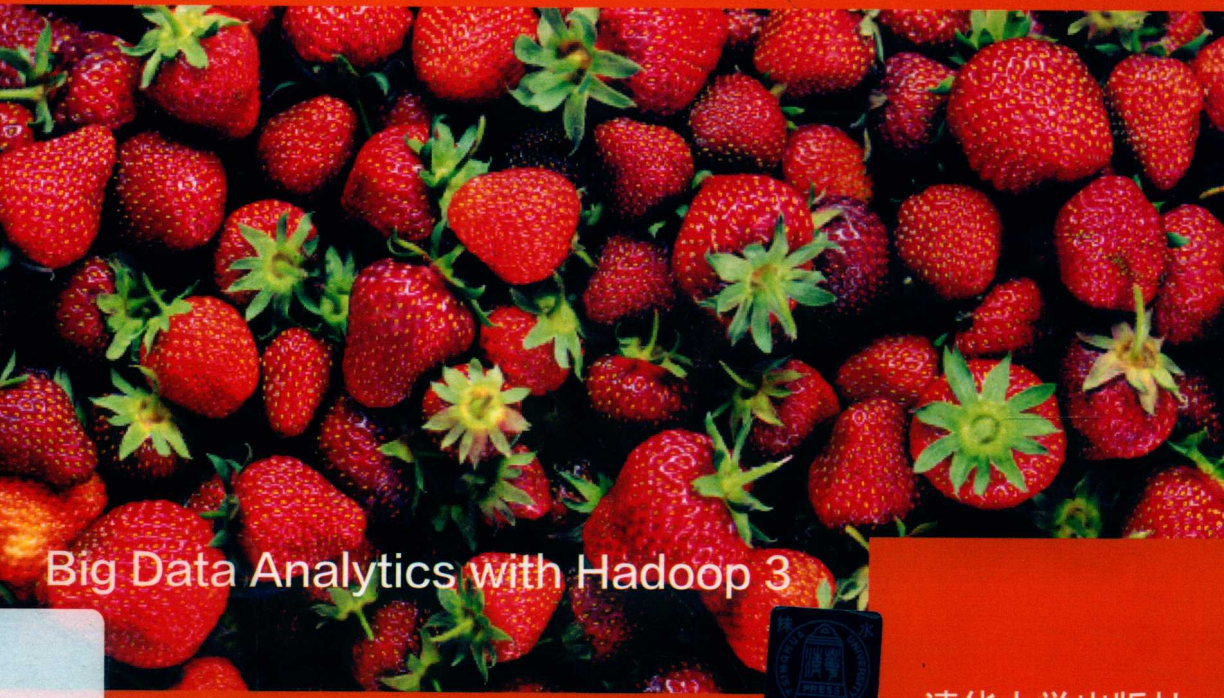
利用Hadoop 3构建高效的大数据分析方案

Hadoop 大数据分析实战

[美] 斯里达尔·奥拉 著　　李 垚 译



Big Data Analytics with Hadoop 3

清华大学出版社

Hadoop 大数据分析实战

[美] 斯里达尔·奥拉 著

李 垚 译

清华大学出版社

北 京

内 容 简 介

本书详细阐述了与 Hadoop 3 大数据分析相关的基本解决方案，主要包括 Hadoop 简介、大数据分析概述、基于 MapReduce 的大数据处理、Python-Hadoop 科学计算和大数据分析、R-Hadoop 统计数据计算、Apache Spark 批处理分析、Apache Spark 实时数据分析、Apache Flink 批处理分析、Apache Flink 流式处理、大数据可视化技术、云计算简介、使用亚马逊 Web 服务等内容。此外，本书还提供了相应的示例、代码，以帮助读者进一步理解相关方案的实现过程。

本书适合作为高等院校计算机及相关专业的教材和教学参考书，也可作为相关开发人员的自学教材和参考手册。

北京市版权局著作权合同登记号 图字：01-2018-6267

图书在版编目（CIP）数据

Hadoop 大数据分析实战/（美）斯里达尔·奥拉（Sridhar Alla）著；李垚译. —北京：清华大学出版社，2019

书名原文：Big Data Analytics with Hadoop 3

ISBN 978-7-302-52789-3

Ⅰ. ①H…　Ⅱ. ①斯…　②李…　Ⅲ. ①数据处理软件-高等学校-教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 076915 号

责任编辑： 贾小红
封面设计： 刘　超
版式设计： 魏　远
责任校对： 马子杰
责任印制： 丛怀宇

出版发行： 清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者： 清华大学印刷厂
经　　销： 全国新华书店
开　　本： 185mm×230mm　　**印　　张：** 24　　**字　　数：** 480 千字
版　　次： 2019 年 5 月第 1 版　　**印　　次：** 2019 年 5 月第 1 次印刷
定　　价： 129.00 元

产品编号：081453-01

译 者 序

Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构，它可以使用户在不了解分布式底层细节的情况下开发分布式程序，充分利用集群的威力进行高速运算和存储。Hadoop 解决了两大问题：大数据存储和大数据分析，也就是 Hadoop 的两大核心内容：HDFS 和 MapReduce。目前，Hadoop 已经成为业界大数据平台的首选方案之一，Hadoop 人才的需求量也是越来越大

本书旨在令读者具备 Hadoop 3 生态系统的分析能力，并能够构建强大的解决方案来执行大数据分析，同时毫不费力地从大数据分析结果中获得敏锐的洞察力。本书涉及 R 语言、Python 语言、Spark、Flink、Hadoop 的综合运用，同时实现了大数据分析的可视化结果。

在本书的翻译过程中，除李垚外，王辉、刘璋、刘晓雪、张博、刘祎、张华臻等人也参与了部分翻译工作，在此一并表示感谢。

由于译者水平有限，难免有疏漏和不妥之处，恳请广大读者批评指正。

译 者

前　言

Apache Hadoop 是一类流行的大数据处理平台，并可与大多数大数据工具集成，以构建功能强大的数据分析方案。本书将围绕这一点对相关软件展开讨论，同时辅以大量的操作实例。

在本书阅读过程中，读者将会系统学习 HDFS、MapReduce、YARN 方面的知识，以及如何实现快速、高效的大数据处理方案。此外，本书还将 Hadoop 与其他开源工具集成，例如 Python 和 R 语言，进而分析和可视化数据，同时针对大数据进行统计计算。一旦读者掌握了这些内容，即可尝试在 Apache Spark 和 Apache Flink 的基础上应用 Hadoop，最终实现实时数据分析和流式处理。除此之外，本书还将讨论如何在云端和端到端管道上利用 Hadoop 构建数据分析方案，并通过操作实例执行大数据分析任务。

在阅读完本书后，读者将具备基于 Hadoop 生态系统的分析能力，同时可构建强大的解决方案执行大数据分析，并拥有自己的技术观点。

适用读者

如果读者希望使用 Hadoop 3 的强大功能为企业或业务构建高性能的分析解决方案，或者您是一名大数据分析新手，那么本书将十分适合于您。另外，本书需要读者具备 Java 编程方面的基础知识。

本书内容

第 1 章将介绍 Hadoop 环境及其核心组件，包括 HDFS 和 MapReduce。

第 2 章将讨论大型数据集的检测处理过程，从中发现数据的模式，生成相应的报告并采集有价值的内容。

第 3 章将讨论 MapReduce，这也是大多数计算/处理系统中的基本概念。

第 4 章探讨 Python 语言，并在此基础上通过 Hadoop 对大数据进行分析。

第 5 章介绍了 R 语言，同时阐述了如何使用 R 语言并借助于 Hadoop 执行大数据统计计算。

第 6 章将考查 Apache Spark，同时根据批处理模型使用 Spark 进行大数据分析。

第 7 章将对 Apache Spark 的流式处理模型进行分析，以及如何打造基于流式的实时分析应用程序。

第 8 章主要介绍 Apache Flink，及其基于批处理模型的、针对大数据分析的应用方式。

第 9 章讨论 DataStream API 和基于 Flink 的流处理。其中，Flink 用于接收和处理实时事件流，并在 Hadoop 集群中存储聚合和结果。

第 10 章考查数据可视化问题，并通过各种工具和技术实现这一功能，例如 Tableau。

第 11 章讲述云计算以及各种概念，例如 IaaS、PaaS 和 SaaS。除此之外，本章还将对云供应商加以简要介绍。

第 12 章介绍 AWS 和 AWS 中的各种服务，这些服务使用 Elastic MapReduce（EMR）在 AWS 云中建立 Hadoop 集群，这对执行大数据分析非常有用。

软件和硬件环境

本书示例是在 64 位 Linux 上使用 Scala、Java、R 和 Python 语言实现的。另外，还应在机器上安装下列内容（建议使用最新版本）：

- Spark 2.3.0（或更高版本）。
- Hadoop 3.1（或更高版本）。
- Flink 1.4。
- Java（JDK 和 JRE）1.8+。
- Scala 2.11.x（或更高版本）。
- Python 2.7+/3.4+。
- R 3.1+和 RStudio 1.0.143。
- Eclipse Mars 或 Idea IntelliJ（最新版本）。

关于操作系统，最好使用 Linux 发行版（包括 Debian、Ubuntu、Fedora、RHEL 和 CentOS）。具体来说，例如，对于 Ubuntu，建议使用完整的 14.04 （LTS） 64 位安装、VMWare player 12 或 Virtual box。此外，还可在 Windows（XP/7/8/10）或者 macOS X（10.4.7+）上运行代码。

关于硬件配置，可采用 Core i3、Core i5（推荐）～Core i7（获得最佳效果）。然而，多核处理将提供更快的数据处理以及较好的可伸缩性。另外，对于单系统模式，至少使用

8GB RAM（推荐）；单个 VM 至少使用 32GB RAM；对于集群，则至少使用 32GB RAM。足够的存储空间可运行繁重的任务（取决于将要处理的数据集大小），最好至少包含 50GB 的空闲磁盘存储空间（用于独立系统和 SQL 仓库）。

资源下载

读者可访问 http://www.packtpub.com 并通过个人账户下载示例代码文件。另外，http://www.packtpub.com/support，注册成功后，我们将以电子邮件的方式将相关文件发与读者。

读者可根据下列步骤下载代码文件：

（1）登录 www.packtpub.com 并注册我们的网站。

（2）选择 SUPPORT 选项卡。

（3）单击 Code Downloads & Errata。

（4）在 Search 文本框中输入书名并执行后续命令。

当文件下载完毕后，确保使用下列最新版本软件解压文件夹：

- Windows 系统下的 WinRAR/7-Zip。
- Mac 系统下的 Zipeg/iZip/UnRarX。
- Linux 系统下的 7-Zip/PeaZip。

另外，读者还可访问 GitHub 获取本书的代码包，对应网址为 https://github.com/PacktPublishing/Big-Data-Analytics-with-Hadoop-3。代码与 GitHub 存储库将实现同步更新。

此外，读者还可访问 https://github.com/PacktPublishing/以了解丰富的代码和视频资源。

除此之外，我们还提供了 PDF 文件，其中包含了本书所用截图/图表的彩色图像。读者访问 http://www.packtpub.com/sites/default/files/downloads/BigDataAnalyticswithHadoop3_ColorImages.pdf 进行下载。

本书约定

代码块则通过下列方式设置：

```
hdfs dfs -copyFromLocal temperatures.csv /user/normal
```

代码中的重点内容则采用黑体表示：

```
Map-Reduce Framework -- output average temperature per city name
    Map input records=35
    Map output records=33
    Map output bytes=208
    Map output materialized bytes=286
```

命令行输入或输出如下所示：

```
$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

图标表示较为重要的说明事项。

图标则表示提示信息和操作技巧。

读者反馈和客户支持

欢迎读者对本书的建议或意见予以反馈。

对此，读者可向 feedback@packtpub.com 发送邮件，并以书名作为邮件标题。若读者对本书有任何疑问，均可发送邮件至 questions@packtpub.com，我们将竭诚为您服务。

勘误表

尽管我们在最大程度上做到尽善尽美，但错误依然在所难免。如果读者发现谬误之处，无论是文字错误抑或是代码错误，还望不吝赐教。对此，读者可访问 http://www.packtpub.com/submit-errata，选取对应书籍，单击 Errata Submission Form 超链接，并输入相关问题的详细内容。

版权须知

一直以来，互联网上的版权问题从未间断，Packt 出版社对此类问题异常重视。若读者

在互联网上发现本书任意形式的副本，请告知网络地址或网站名称，我们将对此予以处理。关于盗版问题，读者可发送邮件至 copyright@packtpub.com。

若读者针对某项技术具有专家级的见解，抑或计划撰写书籍或完善某部著作的出版工作，则可访问 www.packtpub.com/authors。

问题解答

若读者对本书有任何疑问，均可发送邮件至 questions@packtpub.com，我们将竭诚为您服务。

目　录