O'REILLY®

Broadview®
www.broadview.com.cn

大数据猩球

海量数据处理实践指南

Big Data for Chimps

[美] Philip Kromer Russell Jurney 著
唐李洋 译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

O'REILLY®

大数据猩球：海量数据处理实践指南

Big Data for Chimps

[美] Philip Kromer Russell Jurney 著
唐李洋 译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书以实用的、可操作的视角解释了大数据。采用黑猩猩和大象的隐喻，基于棒球统计数据集，使用 Apache Hadoop 和 Pig 等工具展示了如何处理大规模数据。此外，通过处理真实数据、解决现实问题，作者还以实例的形式总结了一些实践分析模式，为有创造力的分析人员提供了最强大、最有价值的方法。

本书特别适合那些需要大数据工具箱来解决实际问题的人们。

版权贸易合同登记号 图字：01-2015-7886

图书在版编目（CIP）数据

大数据猩球：海量数据处理实践指南 /（美）菲利普 · 克罗默（Philip Kromer），（美）拉塞尔 · 贾米（Russell Jurney）著；唐李洋译. — 北京：电子工业出版社，2016.8

书名原文：Big Data for Chimps
ISBN 978-7-121-29418-1

Ⅰ. ①大… Ⅱ. ①菲… ②拉… ③唐… Ⅲ. ①数据处理 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字（2016）第 165693 号

责任编辑：张春雨
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：13.25　字数：313.76 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
定　　价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、在线服务、杂志、调查研究和会议等方式传播创新的知识。自 1978 年开始，O'Reilly 一直都是发展前沿的见证者和推动者。超级极客正在开创未来，我们关注着真正重要的技术趋势，通过放大那些“微弱的信号”来刺激社会对新科技的采用。作为技术社区中活跃的参与者，O'Reilly 的发展充满着对创新的倡导、创造和发扬光大。

作为出版商，O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织开放源代码峰会，以至于开源软件运动以此命名；通过创立了 Make 杂志成为 DIY 革命的主要先锋；公司一如既往地通过用各种方式和渠道连接人们和他们所需要的信息。O'Reilly 的会议和峰会聚集了超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择、O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过印刷书籍、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是一位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

前言

《大数据猩球：海量数据处理实践指南》以实用、可操作的视角解释了大数据，以经过检验的最佳实践为中心，向读者展示了 Hadoop 的实战智慧。

读者将对大数据形成有用的、概念性的认识。数据就是洞察力，关键是理解大数据的可扩展性（scalability）：即无限规模的数据取决于相异的枢轴点（pivot point）。我们会教你如何运用这些枢轴点进行数据操作。

最后，本书提供了真实数据和实际问题的具体示例，将概念和实际应用相结合。

本书梗概

《大数据猩球：海量数据处理实践指南》讲述了如何使用简单、有趣、精致的工具，解决大规模数据处理中的重要问题。

从超大规模的事件流中发现模式是一件重要而且困难的事情。大部分时候，地震是不会发生的——但是模式能够根据平静时期的数据提前预测是否会发生地震。如何在数以亿计的事件中逐个对比数万亿个连续事件，从而发现极少数事关紧要的事件呢？一旦找到了这些模式，如何实时地做出响应？

我们选用大家都能够理解的案例，而且它们具有普适性，能够适用于其他问题解决的场景。我们的目的是向读者提供：

- 大规模思考的能力——使读者深刻理解如何将一个问题分解为有效的数据转换（data transformation），以及集群中的数据流动如何影响这些转换。
- 用详细的示例代码在场景中展现如何使用 Hadoop 解决有意思的问题。

- 关于有效软件开发的建议和最佳实践。

本书的全部示例都采用真实数据，用来描述很多问题领域中的模式，包括：

- 创建统计概要。
- 识别数据中的模式和组。
- 批量查找、过滤和移动记录。

本书强调简洁性和趣味性，特别吸引初学者，但同样适合有经验的人。你会发现本书为有创造力的分析人员提供了最强大、最有价值的方法。我们的座右铭是“机器人是廉价的，而人是重要的”：编写可读的、可扩展的代码，然后再确定是否需要一个较小的集群。本书的代码改编自 Infochimps 和 Data Syndrome 解决企业级业务问题的程序，这些简单的高级转换能够满足我们的需求。

很多章节都配有练习。如果你是初学者，我们强烈建议你每一章都至少完成一个练习。在面前摆本书看，不如边看书边写代码学得更深入。本书官网上有一些简单的解决方案和结果数据集。

本书适合谁

我们希望你至少熟悉一种编程语言，并不一定非要是 Python 或 Pig。熟悉 SQL 会有些帮助，但这不是必需的。如果有商务智能方面的数据工作经历或分析背景，会很有帮助。

更重要的是，你应该有一个需要大数据工具箱来解决问题的实际项目——这个问题要求在多个机器之间横向扩展（scale out）。如果你没有这样的项目，但又确实很想学习大数据工具箱，看一下第 3 章，我们采用棒球数据。这是一个探索起来很有趣的大型数据集。

本书不适合谁

本书不是《Hadoop 权威指南》（*Hadoop: The Definitive Guide*，已出版），而更像是《Hadoop 固执指南》（*Hadoop: A Highly Opinionated Guide*）。本书唯一提到裸 Hadoop API 的地方就是，“大多数情况下，不要使用它”。我们推荐以某种空间不高效的格式存储数据，还有很多时候我们鼓励以小部分的性能损失换取程序员更多的愉悦。本书不厌其烦地强调编写可扩展的代码，却只字不提编写高性能的代码，因为获取成倍加速比的最佳途径是使用双倍数量的机器。

这是因为，对大部分人来说，集群的成本远远低于数据科学家使用它的机会成本。如果数据不仅大，还很巨大（比如 100TB），而且我们期望在生产线上不断地运行作业，那

就需要考虑其他权衡了。但是，即使是 PB 级规模，仍然要按照我们介绍的方式来开发。

本书涉及 Hadoop 的提供和部署问题，以及一些重要的设置。但是并没有真正介绍任何高级算法、操作或调优问题。

本书没有包括的内容

目前我们不讨论 Hive。对于熟悉 Hive 的人，Pig 脚本能够天然地翻译成 Hive。

本书讲的是互联网上没有的东西。我们不准备花时间介绍基础教程和核心文档。另外，我们也不会涉及以下内容：

- Hadoop 的安装或维护。
- 其他类 MapReduce 的平台（Disco、Spark 等），或其他框架（Wukong、Scalding、Cascading）。
- 有时候我们用到了 Unix 测试工具包（cut/wc/etc），但只是作为工具临时用一下。我们并不会深入讲述这些东西，有其他 O'Reilly 书籍详细介绍这些实用工具。

理论：黑猩猩和大象

从第 2 章开始，你会看到黑猩猩和大象公司（Chimpanzee and Elephant Company）热情的员工们。大象记性好（内存很大），易于进行大规模迁移。通过大象类比组装数据，有助于理解移动超大量数据的易点和难点。黑猩猩聪明，但是一次只能考虑一件事情。它们展示了如何在单个关注点下实现简单的转换，以及如何在不占用更多空间的情况下分析 PB 级的数据。

黑猩猩和大象结合起来，共同隐喻了如何处理大规模数据。

实战：Hadoop

Doug Cutting 说，Hadoop 是“大数据操作系统的内核”。Hadoop 是最主流的批处理方案，既有商用企业支持，也拥有庞大的开源社区，能够在每一个平台和云上运行——短期内这种形势并不会改变。

本书中的代码无须改动即可在你的笔记本电脑或企业级 Hadoop 集群上运行。我们使用 docker 提供一个虚拟 Hadoop 集群，你可以在自己的笔记本上运行。

示例代码

使用 Git 签出（check out）本书的源代码：

```
git clone https://github.com/bd4c/big_data_for_chimps-code
```

运行上述命令以后，示例代码在 `examples/ch_XX` 目录下。

关于Python和MrJob

我们选择 Python 有两个原因。第一，作为一种高级语言（除了 Python，还有 Scala、R 等），Python 既拥有完美的 Hadoop 框架又具备广泛的支持。更重要的是，Python 是一种可读性很强的语言。本书提供的示例代码能够清晰地映射到其他高级语言，而且我们推荐的方法在任何语言中都是可用的。

具体来说，我们选择 Python 语言框架 MrJob。这是一个广泛使用的开源框架。

其他有益读物

- 《Pig 编程》：作者 Alan Gates，全面介绍 Pig Latin 语言及 Pig 工具。强烈推荐此书。
- 《Hadoop 权威指南》：作者 Tom White，必备书籍。不要试图一下子完全学会——Hadoop 最强大的地方正是它最简单的地方——但是在你的应用程序上线之前常常需要参考这本书。
- 《Hadoop 管理手册》：作者 Eric Sammer——最好有其他人替你看这本书，不过运行 Hadoop 集群的人，最终还是需要这本指南来配置和管理大型生产线集群。

读者反馈

请联系我们！如果有问题、建议或意见，请通过问题跟踪（issue tracker：*http://bit.ly/bd4c_issues*）分享。如果想直接联系我们，请发送邮件至 *flip@infochimps.com* 和 *russell.journey@gmail.com*（作者）——欢迎抄送至 *meghan@oreilly.com*（我们从未不耐烦的编辑）。还可以通过 Twitter 联系我们：

- Flip Kromer (@mrflip)
- Russell Jurney (@rjurney)

印刷约定

本书使用的印刷约定如下：

斜体字

表示新术语、URL、邮件地址、文件名及文件扩展名。

等宽字体

用于程序代码，以及在段落中表示变量名、函数名、数据库、数据类型、环境变量、语句及关键字等程序元素。

等宽粗体

表示由用户键入的命令或文本。

等宽斜体

表示应当替换成用户自定义值或上下文值的文本。

这个图标表示技巧或建议。

这个图标表示一般注解。

这个图标表示警告或注意。

中文版书中切口以“  ”表示原书页码，便于读者与原英文版图书对照阅读，本书的索引中所列的页码为原英文版页码。

使用示例代码

补充材料（示例代码、练习等）的下载地址为 *http://github.com/bd4c/big_data_for_chimps-code*。

本书旨在帮助你完成工作。一般来说，如果书中提供了示例代码，可以在你的程序和文

档中直接使用，而不需要联系我们获得使用许可，除非你需要大规模仿造（reproduce）这些代码。例如，使用本书的几个代码片段写个程序，不需要获得许可；而出售或分发O'Reilly图书附带光盘中的示例代码，则需要许可。引用本书及书中的示例代码回答某个问题，不需要许可；而在产品文档中大量使用书中的示例代码，则需要许可。

我们感激您注名出处，但不强制要求。注名出处通常包括书名、作者、出版社和ISBN。例如，"《大数据猩球：海量数据处理实践指南》，Philip Kromer和Russel Jurney，O'Reilly，2015，978-1-491-92394-8"。

如果你觉得你对示例代码的使用方式可能需要获得许可，请随时联系我们*permissions@oreilly.com*。

Safari® 在线书库

Safari在线图书（http://safaribooksonline.com）是应需而变的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发者、Web设计师、商务人士和创意精英都可以将Safari在线图书作为他们的调研、解决问题、学习和认证的主要资料来源。

Safari在线图书对于组织团体、政府机构和个人提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sam、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGarw-Hill、Jones & Bartlett、Course Technology及其他数十家出版社的上千种图书、培训视频和正式出版前的书稿。要了解更多关于Safari Books Online的信息，请访问我们的网站。

联系我们

关于本书的建议，可以与下面的出版社联系。

美国：

O'Reilly Media, Inc.

1005 Gravenstein Highway North

Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）

奥莱利技术咨询（北京）有限公司

我们将关于本书的勘误表、示例以及其他信息列在本书的网页上，地址是：*http://bit.ly/developing-web-components*。

如要评论本书或咨询关于本书的技术问题，请发邮件到 *bookquestions@oreilly.com*。

了解更多关于本书、课程、会议和新闻的信息，请访问我们的网站：*http://www.oreilly.com*。

我们的 Facebook：*http://facebook.com/oreilly*

我们的 Twitter：*http://twitter.com/oreillymedia*

我们的 YouTube 视频：*http://www.youtube.com/oreillymedia*

目录

第二部分　战术：分析模式