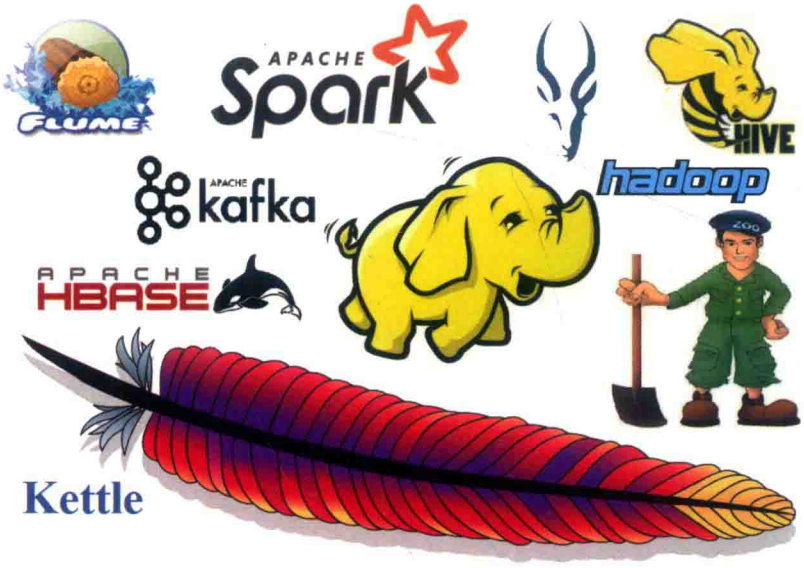
天善智能创始人 梁　勇
清控道口资本管理有限公司管理合伙人 钟志新
深圳纳实大数据技术有限公司CEO 吕　骏
斯凯奇(中国)有限公司 CIO 李宏阳
上海亦策软件科技有限公司总经理 邓强勇

联袂推荐

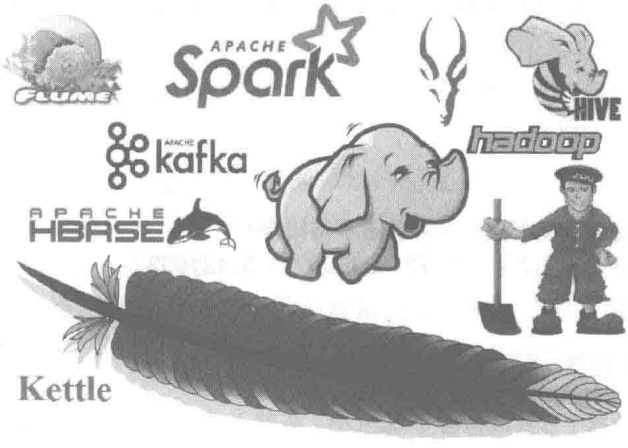
大数据技术



Cloudera Hadoop 大数据平台实战指南

宋立桓 陈建平 著

清華大學出版社



Cloudera Hadoop 大数据平台实战指南

宋立桓 陈建平 著

清华大学出版社
北京

内 容 简 介

对于入门和学习大数据技术的读者来说，大数据技术的生态圈和知识体系过于庞大，可能还没有开始学习就已经陷入众多的陌生名词和泛泛的概念中。本书的切入点明确而清晰，从 Hadoop 生态系统的明星 Cloudera 入手，逐步引出各类大数据基础和核心应用框架。

本书分为 18 章，系统介绍 Hadoop 生态系统大数据相关的知识，包括大数据概述、Cloudera Hadoop 平台的安装部署、HDFS 分布式文件系统、MapReduce 计算框架、资源管理调度框架 YARN 、Hive 数据仓库、数据迁移工具 Sqoop、分布式数据库 HBase、ZooKeeper 分布式协调服务、准实时分析系统 Impala、日志采集工具 Flume、分布式消息系统 Kafka、ETL 工具 Kettle、Spark 计算框架等内容，最后给出两个综合实操案例，以巩固前面所学的知识点。

本书既适合 Hadoop 初学者、大数据技术工程师和大数据技术爱好者自学使用，亦可作为高等院校和培训机构大数据相关课程的培训用书。

图书在版编目（CIP）数据

Cloudera Hadoop 大数据平台实战指南 / 宋立桓，陈建平著. —北京：清华大学出版社，2019
ISBN 978-7-302-51753-5

Ⅰ. ①C… Ⅱ. ①宋… ②陈… Ⅲ. ①数据处理软件Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 271428 号

责任编辑： 夏毓彦
封面设计： 王　翔
责任校对： 闫秀华
责任印制： 刘海龙

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： 三河市龙大印装有限公司
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 16　　**字　　数：** 410 千字
版　　次： 2019 年 2 月第 1 版　　**印　　次：** 2019 年 2 月第 1 次印刷
定　　价： 59.00 元

产品编号：080579-01

推荐序一

从 2013 年起大数据的概念在国内逐步普及，经过短短几年的时间，相关的技术就在各行各业有了深入的使用和发展，并且越来越多的企业开始重视对大数据项目的规划和建设。大数据的项目建设是以 IT 信息化部门为主导、企业各部门紧密配合、企业高层驱动的一个持续的过程，其中大数据技术的相关人才尤为重要。

2013 年，我与几位志同道合的深耕于数据领域的朋友一起成立了天善智能。至今，天善智能已经成为国内最大的大数据、商业智能 BI、人工智能 AI 的垂直社区之一，来自于百度、阿里巴巴、腾讯、微软、IBM、京东等国内一众知名公司的数据专家也积极地活跃在我们的社区。这些专家广泛地参与天善智能各类线上线下有关大数据技术的布道活动，他们用自己专业的知识、精湛的技术分享极大地点燃了广大大数据技术爱好者的热情，共同推动了大数据技术在国内的普及和发展。

在天善智能的成长过程中，我也有幸结识了很多来自各行各业大数据技术圈的朋友，其中就包括本书的两位作者宋立桓和陈建平。宋立桓老师以前在微软工作，由于我以前也是微软技术的程序开发者，包括之后在工作中使用到微软商业智能 BI 技术，因此让我们有了更多交流的话题。之后，宋老师去腾讯就职，从大数据到云计算，从一个很深的数据底层走向另一个更深的架构底层，这是一个很好的提升和发展。建平来自传统的行业，在传统行业的数据升级打怪过程中不断将数据运用到了一个很高的高度。

在我们和很多技术专家合作各类线上线下沙龙分享的过程中，大家都意识到了一个问题——大数据知识体系过于庞大，零零散散的知识体系终归需要有一个载体，而这个载体既可以是文字的沉淀，也可以是专业课程的沉淀。很惊喜的是这两位志同道合的朋友在精心酝酿了很长一段时间之后，终于开始行动并将过往经验一一成文。

Hadoop 走过这么多年，整个生态体系越来越庞大，作为 Hadoop 最有影响力的数据管理软件服务提供商之一的 Cloudera 无疑是一颗耀眼的明星。两位作者从这个切入点开始循序渐进地将 Hadoop 生态系统中核心的技术、框架、应用一一展开，构成一个完整的知识体系框架，不多不少入门正好。本书案例简洁清晰，不少料不拖沓，可以帮助大家快速学习掌握大数据相关的核心知识点，希望此书能够成为广大大数据技术学习爱好者的手边参考书。

最后，回归到整个大行业，我们仍然要意识到：许多传统企业在从业务信息化到数据信息化的过程并不会一帆风顺。一方面来自于传统业务与大数据结合的场景目前依旧需要实践的检验，

存在一个比较长的建设和提炼周期，需要企业在人力、物力、财力有持续的投入和保障。另一方面，每个企业的 IT 基础、数据基础、技术积累程度不同，对于选择适合自身的大数据方案也并不是那么容易。对于技术人员来说，有些问题我们可能无力解决，我们能够做到的就是不断地夯实大数据技术、用技术驱动传统业务、挖掘业务增长点，让大数据真正地为企业创造业务价值。

天善智能创始人 梁勇

推荐序二

最近几年，Hadoop 作为最基础和最流行的大数据技术平台，被广泛地应用。作为一个开源技术平台，Hadoop 平台发展非常迅猛，在此基础上，也发展出很多的商业版本和分支。Cloudera 作为最早开源平台的贡献者，迅速成为这个技术的领导者，无论是社区版还是商业版都被大量的客户广泛采用。

亦策软件是一家国内领先的专注于大数据整体解决方案的高科技企业，为客户提供大数据分析平台端到端的解决方案。在和客户交流的过程中，我们发现很多用户和技术人员在学习、使用和运维 Cloudera Hadoop 平台的时候都面临一个严峻的挑战，就是中文的教材和资料非常少，相关的课程和培训又非常昂贵，所以只能靠自己摸索或者学习英文的资料。这给入门和掌握 Cloudera Hadoop 平台造成很大的不便，降低了效率。听闻宋立桓老师要出版《Cloudera Hadoop 大数据平台实战指南》这本书，我很高兴受邀为这本书写序。这本书梳理了大数据基础和核心技术框架、注重实操、案例简洁，很适合广大大数据技术爱好者参考学习。我想本书的出版一定能帮助大家解决学习资料的燃眉之急。

感谢宋立桓老师为大数据技术的推广做出贡献，并真诚希望广大读者通过本书能缩短相关技术从入门到精通的过程，最后祝本书大卖！

上海亦策软件科技有限公司总经理 邓强勇

推荐序三

2011 年秋，我在西雅图参加微软的技术大会，在之前的一年，EMC 刚刚提出了 Big Data 和 Data Lake 的理念，我和同事们疯狂地讨论着这个也许会颠覆软件业的新理念。我们意识到，大量的数据和非结构化的社会网络信息会成为这个全新时代的重要资源高地。我们讨论着 2005 年诞生的 Hadoop，2007 年 Linus Torvalds 于 10 天之内开发 GitHub 的传奇，2008 年中本聪发表比特币论文，2010 年张小龙开发了微信，那是伟大的事件频繁发生的 5 年。

我与本书作者宋立桓老师是在那时候认识的，我们谈了很多关于未来数据应用的场景，我们一致认为，数据的价值将会在未来 10 年内超过那些陈旧的系统，会有更新一代的应用基于显而易见、唾手可得的数据而诞生。我们将不再依赖复杂的流程和权限，会让每个社会实体都有自知之明和洞见之能。

从那以后，宋立桓老师一直致力于探索数据的价值，以及如何实现数据的价值，从数据分析到数据业务探索，从数据整合到数据共享，他研究得非常系统和完整。《Cloudera Hadoop 大数据平台实战指南》是他继《人人都是数据分析师：微软 Power BI 实践指南》之后的又一力作，通俗易懂，概念清晰，是对大数据架构和相关大数据系统普及的好教程。在本书中，从大数据概念到原理，从理论到实战，从部署到操作，无一不凝聚着他严谨的学习态度和实践精神，是很好的打开大数据宝藏的一把钥匙，也是近年来在该领域不可多得的学习材料。

希望你们和我一起加入本书的阅览与实战的旅程中。

感谢宋立桓老师的作序邀请，预祝有更多推动大数据行业的真知灼见早日发表。

深圳纳实大数据技术有限公司 CEO　吕骏

前　言

大数据这个词也许几年前你听着还有点陌生，但我相信你现在听到 Hadoop 这个词时会觉得“熟悉”！你会发现身边从事 Hadoop 开发或者正在学习 Hadoop 的人越来越多。

最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡，麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。互联网技术发展到现今阶段，大量日常、工作等事务产生的数据比以前有了爆炸式的增长，以前的传统数据处理技术已经无法胜任，需求催生技术——一套用来处理海量数据的软件技术框架 Hadoop 应运而生！

我本人一直从事云计算、大数据方面的咨询和培训工作。大数据产业高速发展促使 Hadoop 人才的需求井喷式增长，但 Hadoop 大数据工程师培养数量远远无法满足市场的需求。为了不被淹没在大数据技术的浪潮中，我们只有坚持学习，通过增加知识来实现对自我价值的挖掘和体现。

关于本书

Hadoop 的发行版除了社区的 Apache Hadoop 外，Cloudera、Hortonworks、华为等公司都提供了自己的商业版本。因为企业通常使用的是 Hadoop 商业版本，所以本书实操的运行环境采用 Cloudera 的 CDH。本书定位是大数据从入门到应用的简明系统教程，特色是理论联系实践、实战实用为主、内容全面系统、讲解深入浅出，是大数据技术爱好者入门的最佳图书。

本书分为 18 章（宋立桓老师撰写第 1~12 章、陈建平撰写第 13~18 章），分别从大数据概述、Cloudera Hadoop 平台的安装部署、大数据 Hadoop 组件三方面进行介绍，内容包括 HDFS 分布式文件系统、MapReduce 计算框架、资源管理框架 YARN 、Hive 数据仓库、数据迁移工具 Sqoop、分布式数据库 HBase、ZooKeeper 分布式协调服务、准实时分析系统 Impala、日志采集工具 Flume、分布式消息系统 Kafka、ETL 工具 Kettle、Spark 计算框架等知识，最后用两个综合实操案例把所有知识点串起来。

本书使用的操作环境是 Hadoop 商业发行版的 Cloudera Express（Express 是免费版本，企业版需付费）。全书秉承“实践为主、理论够用”的原则，将演示实验融入各个知识点讲解中。

本书另提供丰富的案例源文件和大数据工具软件下载，供读者亲自操作练习，在作者博客 http://blog.51cto.com/lihuansong 中有下载地址。

学习本书之前，希望大家应该具有如下基础：有一定计算机网络基础知识，熟悉常用 Linux 操作命令，对 Java 语言和数据库理论有基本的了解。

资源下载与技术支持

本书提供详细的案例资源文件，在作者博客置顶文章中提供下载地址，便于读者动手实践：**http://blog.51cto.com/lihuansong/2317021**

欢迎读者来信互动，宋立桓的邮箱是 songlihuan@hotmail.com，陈建平的邮箱是 daxia1520@163.com。

致谢

感谢我的妻子，她是我完成此书的坚强后盾。

感谢我的朋友和同事，他们让我学会知识的增值和变现。

感谢清华大学出版社的编辑夏毓彦和其他工作人员帮助我出版了这本有意义的著作。

阿基米德有一句名言：“给我一个支点，我就能撬起地球。”谨以此书献给那些为大数据与商业智能分析铺路的人，让更多的人享受到大数据时代到来的红利。

宋立桓

云计算架构师、大数据咨询顾问

2018 年 11 月

目　录