Spark快速大数据分析 (图灵程序设计丛书)

作者：[美] 卡劳（Holden Karau）[美] 肯维尼斯科（Andy Konwinski）[美] 温德尔（Patrick Wendell）[加] 扎哈里亚（Matei Zaharia）

版权信息

书名：Spark快速大数据分析

作者：[美] Holden Karau Andy Konwinski,Patrick Wendell [加] Matei Zaharia

译者：王道远

ISBN：978-7-115-40309-4

版权声明

O'Reilly Media, Inc. 介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

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）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多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。无论是通过书籍出版、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*

推荐序

近年来大数据逐渐升温，经常有人问起大数据为何重要。我们处在一个数据爆炸的时代，大量涌现的智能手机、平板、可穿戴设备及物联网设备每时每刻都在产生新的数据。当今世界，有 90% 的数据是在过去短短两年内产生的。到 2020 年，将有 500 多亿台的互联设备产生 Zeta 字节级的数据。带来革命性改变的并非海量数据本身，而是我们如何利用这些数据。大数据解决方案的强大在于它们可以快速处理大规模、复杂的数据集，可以比传统方法更快、更好地生成洞见。

一套大数据解决方案通常包含多个重要组件，从存储、计算和网络等硬件层，到数据处理引擎，再到利用改良的统计和计算算法、数据可视化来获得商业洞见的分析层。这中间，数据处理引擎起到了十分重要的作用。毫不夸张地说，数据处理引擎之于大数据就像 CPU 之于计算机，或大脑之于人类。

早在 2009 年，Matei Zaharia 在加州大学伯克利分校的 AMPLab 进行博士研究时创立了 Spark 大数据处理和计算框架。不同于传统的数据处理框架，Spark 基于内存的基本类型（primitive）为一些应用程序带来了 100 倍的性能提升。Spark 允许用户程序将数据加载到集群内存中用于反复查询，非常适用于大数据和机器学习，日益成为最广泛采用的大数据模块之一。包括 Cloudera 和 MapR 在内的大数据发行版也在发布时添加了 Spark。

目前，Spark 正在促使 Hadoop 和大数据生态系统发生演变，以更好地支持端到端的大数据分析需求，例如：Spark 已经超越 Spark 核心，发展到了 Spark streaming、SQL、MLlib、GraphX、SparkR 等模块。学习 Spark 和它的各个内部构件不仅有助于改善大数据处理速度，还能帮助开发者和数据科学家更轻松地创建分析应用。从企业、医疗、交通到零售业，Spark 这样的大数据解决方案正以前所未见的力量推进着商业洞见的形成，带来更多更好的洞见以加速决策制定。

在过去几年中，我的部门有机会与本书的作者合作，向 Apache Spark 社区贡献成果，并在英特尔架构上优化各种大数据和 Spark 应用。《Spark 快速大数据分析》的出版为开发者和数据科学家提供了丰富的 Spark 知识。更重要的是，这本书不是简单地教开发者如何使用 Spark，而是更深入介绍了 Spark 的内部构成，并通过各种实例展示了如何优化大数据应用。我向大家推荐这本书，或更具体点，推荐这本书里提倡的优化方法和思路，相信它们能帮助你创建出更好的大数据应用。

英特尔软件服务事业部全球大数据技术中心总经理 马子雅

2015 年 7 月于加州圣克拉拉

Big data is getting hot in recent years. Quite often, folks ask why big data is a big deal. We are in the era of data explosion, with the emergence of smart phones,tablets, wearables, IoT devices, etc. Ninety percent of the data in the world today was generated in just the past two years. By 2020, we will see >50B devices connected and Zeta byte data created. It is not the quantity of the data that is revolutionary. It is that we can now do something with it that's revolutionary. The power of big data solutions is they can process large and complex data sets very fast, generate better and faster insights than conventional methods.

A big data solution suite can consist of several critical components, from the hardware layer like storage, compute and network, to data processing engine, to analytics layer where business insights are generated using improved statistical & computational algorithms and data visualization. Among all, the data processing engine is one most critical player. It is not overstating that the data processing engine for big data is like CPU for a computer or brain for a human being.

Spark was initially started for the purpose of creating a big data processing and computing framework, when Matei Zaharia was doing his Ph.D. research at UC Berkeley AMPLab in 2009. Different from the traditional data processing framework, Spark's in-memory primitives provide performance up to 100 times faster for certain applications. By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well-suited for big data and machine learning use cases. Spark is becoming one best adopted among all big data modules. Big Data Distributions like Cloudera, MapR now all include Spark into their distributions.

Spark is now evolving the Hadoop and big data ecosystem to better support the end-to-end big data analytics needs, e.g. Spark grew beyond Spark core to Spark streaming, SQL, MLlib, GraphX, SparkR, etc. Learning Spark and its internals will not just help improve the processing speed for big data, but also help developers and data scientists create analytics applications with more ease. With big data solutions like Spark, we expect to see significant improvement with business insights which will help expedite the decision making—like we've never seen before, from enterprise, healthcare, transportation, and retail.

Over the years, my organization had the opportunities to work with authors of this book, contribute to Apache Spark, and optimize various Big Data and Spark application on Intel Architecture. The publication of Learning Spark offers developers and data scientists'extensive knowledge on Spark. Moreover, Learning Spark does not simply try to tell the developers how to use Spark, it also addresses the internals and shows various examples of how to improve your big data applications. I recommend Learning Spark—that this book, and, more specifically, the method it espouses, will change your big data application for the better.

Ziya Ma, General Manager of the global Big Data Technologies organization,

SSG STO, Intel Corp.

Santa Clara, California, July 2015