

Spark GraphX

实战

[美] Michael S. Malak 著
Robin East 译
时金魁 黄光远 译

Spark GraphX in Action



Spark GraphX 实战

[美] Michael S. Malak 著
Robin East
时金魁 黄光远 译

Spark GraphX
in Action

電子工業出版社

Publishing House of Electronics Industry

北京•BEIJING

内 容 简 介

本书是一本Spark GraphX入门书籍。前5章为基础内容,即使读者对Spark、GraphX、Scala不熟悉,也能快速上手;后5章为图计算进阶,主要是图算法和机器学习算法的相关内容。专门讲图计算的书很少,本书在第2、3、4章介绍了图的基础知识、GraphX基础知识、GraphX内置的图算法。第6章到第10章,主要介绍了GraphX之外的图算法、机器学习、图工具、GraphX监控和优化、GraphX的能力增强等实用技能。第9章和第10章主要介绍性能调优和监控,主要面向生产环境,有不少可以借鉴的技巧。

本书面向对图计算感兴趣的读者,旨在帮助读者掌握Spark GraphX的相关知识及其应用。

Original English Language edition published by Manning Publications, USA. Copyright © 2016 by Manning Publications. Simplified Chinese-language edition copyright © 2017 by Publishing House of Electronics Industry. All rights reserved.

本书简体中文版专有出版权由Manning Publications 授予电子工业出版社。未经许可,不得以任何方式复制或抄袭本书的任何部分。专有出版权受法律保护。

版权贸易合同登记号 图字: 01-2016-7235

图书在版编目(CIP)数据

Spark GraphX 实战 / (美) 迈克尔·S.马拉克(Michael S.Malak), (美) 罗宾·伊斯特(Robin East)著; 时金魁, 黄光远译. —北京: 电子工业出版社, 2017.4

书名原文: Spark GraphX in Action

ISBN 978-7-121-31043-0

I. ①S… II. ①迈… ②罗… ③时… ④黄… III. ①数据处理软件 IV. ①TP274

中国版本图书馆CIP数据核字(2017)第044549号

策划编辑: 张春雨

责任编辑: 刘 舫

印 刷: 三河市良远印务有限公司

装 订: 三河市良远印务有限公司

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: 18.5 字数: 355千字

版 次: 2017年4月第1版

印 次: 2017年4月第1次印刷

定 价: 79.00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zltts@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。

译者序

2016 年夏天，明风把我推荐给了电子工业出版社计算机出版分社的张春雨（@永恒的侠少）编辑翻译这本书。当我拿到序言和第 6 章时，看了一下内容，觉得不是那么复杂难懂。虽然我的英语水平一般，但也可以强迫一下自己，可以查字典，可以找人请教嘛。关键是 Spark 发展得红红火火，使用 GraphX 的人还不那么多，图算法的小宇宙被忽略。做一点努力，希望对大家有点用。

以前读过的很多书，几乎每本书的作者都要先感谢自己的老婆和孩子，觉得真矫情。真摊到自己身上，发现没有家人的支持，是不会有时间写书和进行翻译的。所以，我在这里由衷地说一句：感谢老婆。另外，半路拉了黄光远（花名刀剑）壮丁来帮忙翻译第 7、8、10 章，刀剑对机器学习这部分内容比较熟悉，非常感谢！

这本书在图方面，详细讲解了 GraphX 的方方面面，还附带介绍了 Scala 和 Spark 的各个知识点，算是比较完整。本书涉及 Spark 的基础知识、Scala 的基础知识、图基础知识、GraphX 的内置图算法、GraphX 之外的常用图算法、基于图的机器学习算法、性能和调优，附录中还列举了 GraphX 入门的各种工具。总体来讲，这是一本 GraphX 入门书籍，深入浅出，适合广大对 GraphX 感兴趣的朋友。读完这本书，要想进一步学习，那就要多看关于图基础、图算法和图论的相关书籍，以及把 GraphX 的源代码仔细研读几遍。

另外，即使我们把 GraphX 搞得通透，没有实际应用场景也是不行的。GraphX 适用于迭代计算的图算法和关系数据场景，比如，社区发现可用于发现团伙欺诈，PageRank 发现了关系图中有影响力的对象，k-core 用于筛出关键顶点集……其实图算法除适合社交网络外，个人觉得，在互联网金融领域会比较有价值，毕竟反欺诈仅靠离线计算和规则很难及时发现团伙欺诈行为以及个体之间的关系，而互联网金融最核心的事情就是降低风险、提升贷款等业务的审核通过率，以及使通过率和风险接近一个最佳收益平衡点，即当下最赚钱的状态，这是一个技术活。

Spark 转眼已经发展到 2.0 了，发展的速度让我们穿上飞鞋也跟不上。相对于 SQL/Streaming 等模块的快速发展，GraphX 和 Spark Streaming 的发展速度显得有些慢。但是，得益于 Spark 2.0 在性能方面显著的提升，GraphX 还是能沾些雨露的，性能也得到了相应提升。

2015 年年初在淘宝技术部，我曾反复研读 GraphX 的源码，后来社区加了 aggregateMessages() 函数，使性能提升了 15% 左右，再看其整体设计，相当精妙。我觉得 GraphX 是 Spark 中设计最出色的部分，我写了几篇分析文章，网址为 https://github.com/shijinkui/spark_study。在 Spark 的大旗下，相比其他图处理框架，GraphX 有着天然的平台优势，GraphX 在关系数据场景中会有所作为。

我近期在研究 Flink，也看好 Flink 里的 Gelly 和 FlinkML，即实时场景下的图计算和机器学习，我觉得这是一项有前途的技术。

由于本人英语水平有限，也是第一次翻译图书，为方便大家吐槽，在 GitHub 上建立了一个页面：https://github.com/pusuo/graphx_in_action，欢迎大家在这里对本书的翻译吐槽、纠错，发现问题请一定提出来，千万别手软。

时金魁

2016/10/29 于 杭州

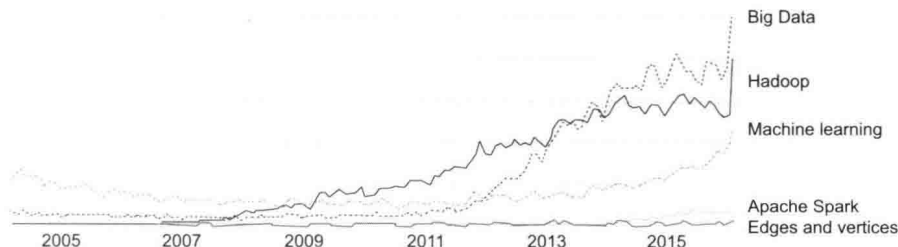
序言

图（Graph）是什么？图是由边和顶点组成的，不是由坐标轴和刻度构成的。在 Spark 中是如何使用图的？这就是本书将要回答的问题。

常常说，“图可以任何事情”，或者“有很多不同的事情可以用图来实现”。当然了，这两种说法等于什么也没说。所以在本书中我们展示了一些具体的、实际的图应用，以及探讨如何用 Spark GraphX 实现这些图应用。

本书中有许多专业术语：大数据、Hadoop、Spark、图、机器学习、Scala 和函数式编程，这些内容本书都会一一讲解。本书会涉及技术的高级部分，但不会涉及编程能力的每个方面，如 Java 编程。

下图是 Google 在趋势上的统计，展示了这些专业术语在 2016 年之前的受欢迎程度。



注意，通常用 *Spark* 和图作为规范的通用术语，而不是 *Apache Spark* 和 *Edges*

and vertices，趋势上后者已明显被取代。机器学习和图，在计算机科学中有悠久的历史，现在作为主流的大数据技术，在商业领域又引起了新的潮流。如果你在学校学习过这些技术的理论知识，那么现在准备实践一下这些技术吧。

许多我们正在或曾经工作过的公司，已经把 Spark 用在生产环境中了，尽管不一定用了 GraphX。当尝试用 GraphX 做一些图解决方案的原型时，会很方便。如果你已经有了一个 Spark 集群或者决定用云平台上的 Spark 集群（例如 Databricks 或 Amazon），那么无须重新搭建一个新的特定于图计算的集群，并且你可以在 GraphX API 中使用已有的 Spark 技能。现在越来越多的图应用为大家所熟知，从根据 Twitter 数据发掘出恐怖分子网络到根据信用卡交易数据发现欺诈行为，GraphX 已成为一个快速尝试这些图算法场景的易于使用的平台。

本书有两个明确目标：一是全面覆盖 Spark GraphX 的方方面面；二是以读者在前面提到的大数据和图计算方面没有任何专业知识为假定前提。写这本书最大的挑战是要有许多技术储备，特别是 Spark、Scala 和图；了解大量的 GraphX API 以及图的不同用法也是不小的挑战。面对这种情形，本书就需要与其他技术书籍略有不同：首先要花点时间入门，前 5 章主要讲解的就是基础内容；本书还有大量有趣的实例，可以跟着一步步练习。本书中涉及的其他技术，读者需要另做学习，本书将努力做到让读者并不需要有过多背景知识和经验，就可以浅显易懂地了解图所能解决的问题。

致谢

感谢 Manning 出版社许多工作人员对本书出版所做的努力，特别需要感谢三个人，他们的诸多指导使本书可以较好地完成。Marina Michaels，我们的开发编辑，从一开始就指出有些章节有很大的技术问题，这些问题大都是 Spark 和图计算的新问题。Michael Roberts，我们的技术开发编辑，在本书的制作过程中，与 Marina 一样，他给出了大量的建议。Antonio Magnaghi，我们的技术校对，不但对本书的示例代码进行了严谨的核对，还编辑校对了本书的文本内容。

同时也感谢本书草稿时期给出诸多有价值建议的读者，他们是 Andy Petrella, Brent Foust, Charles Feduke, Gaurav Bhardwaj, Jason Kolter, Justin Fister, Michael Bright, Paul-Michael Sorhaindo, Rodrigo Abreu, Romi Kuntsman, Sumit Pal, Vincent Liard。

作者 Michael Malak 感谢妻子和孩子在这几个月写作期间给予的耐心支持。

作者 Robin East 感谢妻子和两个儿子，他们容忍和支持了作者长时间的写作以及在楼上偶尔消失。

关于本书

通过学习本书，希望能降低难懂的图学习门槛，了解如何在市场份额最大的分布式计算框架 Apache Spark 中开发图应用。

本书的读者对象

我们假定本书的读者并不熟悉 Spark、Scala 和图相关的知识，本书会快速学习前面提到的这些知识，会特别侧重于 Scala。在第 3 章有 Scala 的简要介绍，全书中只要出现新的 Scala 知识点都会有 Scala 小贴士做详细介绍（完整的列表见附录 D）。实际上，本书通过第 3 章、Scala 小贴士和附录对 Scala 做了较全面的介绍。

另外，虽然在大学的图论课程中数学证明很常见，但本书完全不做数学证明。本书的目标是图算法和图应用，有时会应需介绍图相关的术语。

本书使用的是 Spark/GraphX 1.6 版本。

我们假定读者在 Java 语言编程方面有一些经验，而在图方面要求不多，但通过书中插图能自然地知道这些图应用。

本书的内容组织

本书分为 3 个部分。第 1 部分有 3 章，主要介绍使用 Spark GraphX 的准备知识。

第 2 部分有 4 章，主要介绍如何使用 GraphX。第 3 部分有 3 章，主要介绍 GraphX 的进阶知识。

也可以将本书分为两部分，前 5 章为准备知识和 GraphX 的基本 API，后 5 章为 GraphX 应用。

下面是各章的内容提要。

- 第 1 章介绍了什么是大数据、Spark 和图，Spark GraphX 如何处理数据流。第 1 章是一本迷你书，篇幅不长但内容涉及面较广。
- 第 2 章简单示范了如何使用 GraphX，无须具有 GraphX 经验。
- 第 3 章介绍了 Spark、Scala 和图的基础知识。
- 第 4 章介绍了 Spark GraphX 的基础操作，以及如何使用 GraphX 的两个主要算法：Map/Reduce 和 Pregel。
- 第 5 章演示了如何使用 GraphX 的诸多内置算法。
- 第 6 章介绍了 GraphX API 之外的内容，即 20 世纪中期经典的图算法，并用 GraphX 实现了这些算法。
- 第 7 章重点讲机器学习。机器学习的内容本身就够讲一本书的，这里没有讲解太多机器学习的基础知识和经验，而是直接介绍监督学习、无监督学习和半监督学习的高级实例。
- 第 8 章展示了 GraphX 如何完成一些自定义操作，有可能会构建一个图处理库：读 RDF 文件、图的合并、图查找和计算全局聚类系数。
- 第 9 章介绍了如何监控性能以及查看正在执行的 GraphX 应用程序，如何利用缓存、checkpointing 和序列化调优做性能调优。
- 第 10 章介绍了在 GraphX 中如何使用 Scala 之外的语言（强烈建议不要这么做），以及如何使用一些工具来补充 GraphX 的不足。展示了用 GraphX 在 Apache Zeppelin 的交互式命令行 notebook 上对图进行可视化。第三方的工具 Spark JobServer 可以让 GraphX 从单纯的批处理系统转变成一个在线图数据库。最后，介绍了 Github 上的一个项目——GraphFrames（GraphX 的开发者开发的），它用 Spark SQL DataFrames 而非 RDD 提供了一种方便和高性能的图查询方式。

另外，本书包含 4 个附录。附录 A 介绍了 Spark 的安装方法，附录 B 简要介绍

了 Gephi 可视化软件, 附录 C 包含关于 GraphX 的在线资料以及如何跟上社区最新进展, 附录 D 中列出了本书中的 Scala 小贴士。

如果你在 Spark、Scala 或图方面是新手, 通过前 5 章的阅读, 能力可以得到提升。然后, 就可以选择性地阅读后面 5 章的内容了。

如果你对 Spark、Scala 和图的知识已经比较精通, 但对 GraphX 还不熟悉, 那么可以跳过前 3 章甚至前 5 章的内容直接阅读后面的内容。

关于本书中的代码

本书中的源代码可以在博文视点官方网站上下载, 地址为 [https:// www.broadview.com.cn](https://www.broadview.com.cn)。

这本书中的大部分代码是可以在交互式的 Spark Shell 中执行的。从技术上来讲, Scala 扩展是一个误称, 因为这些文件不能用 scalac 编译器进行编译。

要想让这些实例代码方便地编译和执行, 可以用 Maven 的 pom.xml 或 SBT 的 .sbt 文件完成。

本书中的源代码实例, 有带编号的列表, 也有普通的文本, 这两类源代码都用了等宽字体以便与普通的正文区分开来。

一般情况下, 源码都会被格式化, 我们添加了换行符, 也修改了缩进以便适应本书的页面大小, 甚至在代码清单中添加了续行符 (↵)。另外, 如果正文中对代码有解释, 源码中的注释会被删掉。代码注释会附带许多列表, 以突出重要的概念。

本书的示例代码也可以从如上的出版社网站上下载。

关于作者



Michael Malak 一直从事软件开发工作, 自 2013 年年初以来他一直用 Spark 为财富 200 强的公司做开发工作, 经常进行演示和分享, 特别是在科罗拉多州他住的丹佛 / 博尔德地区。他的个人技术博客的地址是 <http://technicaltidbit.com>。



Robin East 在一些大型企业曾担任过 15 年以上的顾问，在金融、政府、医疗保健和公共事业领域提供大数据和智能解决方案。他是 Worldpay 的数据科学家，帮助公司实现把数据用于核心业务上。可以在这里看到他在 Spark、GraphX 和机器学习方面的作品：<https://mlspeed.wordpress.com>。

配套服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），即可享受以下服务。

下载资源：本书所提供的示例代码及资源文件均可在“下载资源”处下载。

提交勘误：你对书中内容的修改意见可在“提交勘误”处提交，若被采纳，将获赠博文视点社区积分（在购买电子书时，积分可用来抵扣相应金额）。

与我们交流：在页面下方“读者评论”处留下你的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/31043>

二维码：



关于封面插图

本书封面插图的标题为“来自克罗地亚达尔马提亚的希贝尼克的人”。插图是由 Nikola Arsenovic 取自 19 世纪中叶克罗地亚传统服装专辑画册，2003 年由克罗地亚斯普利特的民族博物馆出版。插图是从斯普利特民族博物馆的一个热心的图书管理员那里得到的，博物馆坐落于中世纪的罗马镇中心，大约公元 304 年皇帝戴克里先的宫殿所在的废墟。这本书中有来自克罗地亚不同地区的精细彩色的人物插图，描绘了当地特色服装和日常生活。

希贝尼克是克罗地亚的一个历史名城，位于达尔马提亚中部，克尔卡河在那里流入亚得里亚海。希贝尼克是克宁县的政治、教育、工业和旅游中心，是达尔马提亚历史地区的第三大城市。它是亚得里亚海岸线上最古老的克罗地亚城镇。

服装风格和生活方式在过去 200 年中发生了很大变化，当地特有的多样性逐渐消失。现在很难分辨不同大陆的居民，更不用说相距只有几英里的不同的小村庄或城镇。也许我们已经将文化多样性变为更多样化的个人生活，特别是为了适应更多样化和快节奏的技术生活。Manning 出版社通过在封面上呈现两个世纪之前的丰富的区域文化生活来庆祝今天计算机科技所体现的创造性和主动性，并将读者带回到来自旧书和收藏品中的插图所显示的生活中。

目录

序言	XI
致谢	XIII
关于本书	XIV
关于封面插图	XVIII

第1部分 Spark和图

1 两项重要的技术：Spark和图	3
1.1 Spark：超越Hadoop MapReduce	4
1.1.1 模糊的大数据定义	6
1.1.2 Hadoop：Spark 之前的世界	6
1.1.3 Spark：内存中的 MapReduce 处理	7
1.2 图：挖掘关系中的含义	9
1.2.1 图的应用	11
1.2.2 图数据的类型	12
1.2.3 普通的关系型数据库在图方面的不足	14

1.3	把快如闪电的图处理放到一起: Spark GraphX	14
1.3.1	图的属性: 增加丰富性	15
1.3.2	图的分区: 当图变为大数据集时	17
1.3.3	GraphX 允许选择: 图并行还是数据并行	19
1.3.4	GraphX 支持的各种数据处理方式	19
1.3.5	GraphX 与其他图系统	21
1.3.6	图存储: 分布式文件存储与图数据库	23
1.4	小结	23
2	GraphX快速入门	24
2.1	准备开始并准备数据	24
2.2	用Spark Shell做GraphX交互式查询	26
2.3	PageRank算法示例	29
2.4	小结	31
3	基础知识	32
3.1	Scala——Spark的原生编程语言	33
3.1.1	Scala 的理念: 简洁和表现力	33
3.1.2	函数式编程	34
3.1.3	类型推断	38
3.1.4	类的声明	39
3.1.5	map 和 reduce	41
3.1.6	一切皆是“函数”	42
3.1.7	与 Java 的互操作性	44
3.2	Spark	44
3.2.1	分布式内存数据: RDD	44
3.2.2	延迟求值	47
3.2.3	集群要求和术语解释	49
3.2.4	序列化	50
3.2.5	常用的 RDD 操作	50
3.2.6	Spark 和 SBT 初步	54
3.3	图术语解释	55

3.3.1	基础.....	55
3.3.2	RDF 图和属性图.....	58
3.3.3	邻接矩阵.....	59
3.3.4	图查询系统.....	59
3.4	小结	60

第2部分 连接顶点

4	GraphX 基础.....	65
4.1	顶点对象与边对象	65
4.2	mapping操作	71
4.2.1	简单的图转换.....	71
4.2.2	Map/Reduce	73
4.2.3	迭代的 Map/Reduce	77
4.3	序列化/反序列化	79
4.3.1	读 / 写二进制格式的数据.....	79
4.3.2	JSON 格式	81
4.3.3	Gephi 可视化软件的 GEXF 格式.....	85
4.4	图生成	86
4.4.1	确定的图.....	86
4.4.2	随机图.....	88
4.5	Pregel API.....	90
4.6	小结	96
5	内置图算法.....	97
5.1	找出重要的图节点：网页排名	98
5.1.1	PageRank 算法解释.....	98
5.1.2	在 GraphX 中使用 PageRank.....	99
5.1.3	个性化的 PageRank.....	102
5.2	衡量连通性：三角形数	103
5.2.1	三角形关系的用法.....	103
5.2.2	Slashdot 朋友和反对者的用户关系示例	104

5.3	查找最少的跳跃：最短路径	106
5.4	找到孤岛人群：连通组件	107
5.4.1	预测社交圈子	108
5.5	受欢迎的回馈：增强连通组件	114
5.6	社区发现算法：标签传播	115
5.7	小结	117
6	其他有用的图算法	118
6.1	你自己的GPS：有权值的最短路径	119
6.2	旅行推销员问题：贪心算法	124
6.3	路径规划工具：最小生成树	127
6.3.1	基于 Word2Vec 的推导分类法和最小生成树	131
6.4	小结	135
7	机器学习	136
7.1	监督、无监督、半监督学习	137
7.2	影片推荐：SVDPlusPlus	139
7.2.1	公式解释	146
7.3	在MLlib中使用GraphX	146
7.3.1	主题聚类：隐含狄利克雷分布	147
7.3.2	垃圾信息检测：LogisticRegressionWithSGD	156
7.3.3	使用幂迭代聚类进行图像分割（计算机视觉）	160
7.4	穷人（简化版）的训练数据：基于图的半监督学习	165
7.4.1	K 近邻图构建	168
7.4.2	半监督学习标签传播算法	175
7.5	小结	180

第3部分 更多内容

8	缺失的算法	183
8.1	缺失的基本图操作	184
8.1.1	通用意义上的子图	184