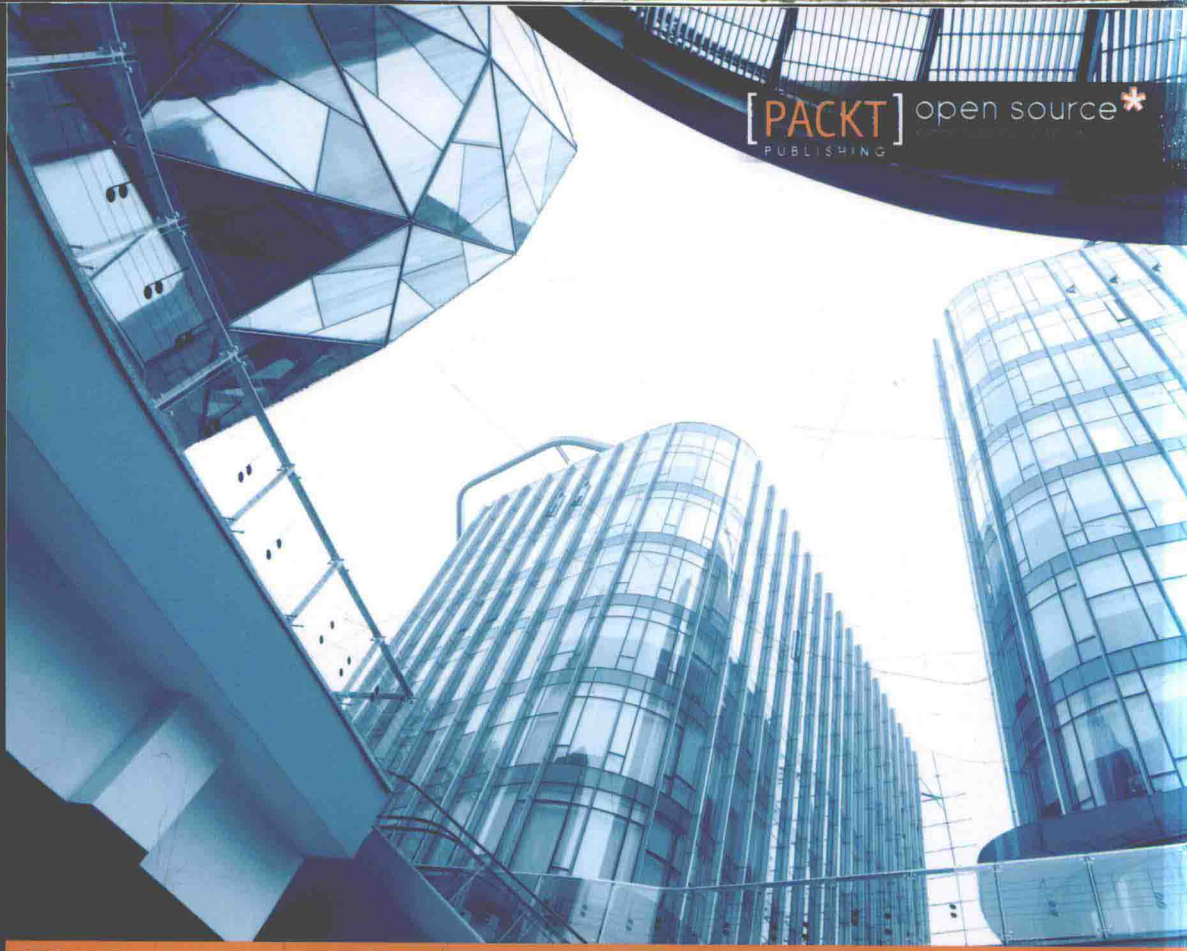

Hadoop for Finance Essentials

Hadoop金融大数据分析

利用大数据为金融机构提供超强的洞察力、分析与商业智能思想。

[美] Rajiv Tiwari 著
王小宁 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

Hadoop for Finance Essentials

Hadoop金融大数据分析

[美] Rajiv Tiwari 著
王小宁 译

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

内容简介

随着数据的增长以及企业每天处理越来越多的数据，Hadoop 作为一个数据平台已经变得很流行。金融行业想要最小化风险和最大化收益，Hadoop 作为一个主宰大数据市场的工具，在其中起着很大的作用。

本书介绍了大数据和 Hadoop 的基础知识，让读者掌握项目管理、欺诈检测等 TOP 大数据金融项目，其中不仅包含行业参考和代码模板，同时包括实现中使用的多个 Hadoop 组件。

读完本书，读者会理解一些行业领先的架构模式、大数据管理经验、窍门和大数据最佳实践方案，以便基于 Hadoop 成功地开发出适合自己的解决方案。

版权贸易合同登记号 图字：01-2015-6646

图书在版编目（CIP）数据

Hadoop 金融大数据分析 /（美）拉吉夫 • 蒂瓦里（Rajiv Tiwari）著；王小宁译. —北京：电子工业出版社，2017.5

书名原文：Hadoop for Finance Essentials

ISBN 978-7-121-31051-5

Ⅰ. ① H… Ⅱ. ①拉… ②王… Ⅲ. ①金融－数据处理软件 Ⅳ. ① F830.49

中国版本图书馆 CIP 数据核字（2017）第 044538 号

策划编辑：高洪霞
责任编辑：徐津平
特约编辑：赵树刚
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
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：720×1000　1/16　印张：10.75　字数：172 千字
版　　次：2017 年 5 月第 1 版
印　　次：2017 年 5 月第 1 次印刷
定　　价：59.00 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。

译者序

从2013年暑假接触Hadoop到现在已有3年，我清楚地记得第一个伪分布式弄了近10天才跑出来第一个WordCount，期间太多的Bug已经把我搞得神魂颠倒，好在最后“成功”了。至此，我与Hadoop结下了不解之缘。刚开始用中国人民大学数据挖掘中心的十几台机器搭建了第一个Hadoop集群，而后发展成两台服务器各包括20台机器的集群。Hadoop的版本也从1.2.0发展到2.6.0，随后帮助中国人民大学统计与调查中心搭建了自己的Hadoop集群。

“巧妇难为无米之炊”，再优秀的工具没有数据也只能是一个摆设，好在我们在做项目的过程中不时地有新的数据加入，也为我们进一步的学习和研究打下了基础。我们集群的组件也从单纯的Hadoop增加到Hive、HBase、Mahout和Spark。这几个组件都是比较流行的，我们在使用过程中也体会到了这些组件优于传统数据分析工具的特点。随着数据采集量的增多，也使得很多公司为我们提供了一些可进行分布式计算的平台环境，充分利用这些资源，会为我们的研究和工作锦上添花。

感谢电子工业出版社的编辑给了我一次这么好的机会，也希望本书能为金融行业的同仁带来一定的收获。金融行业的数据可以说是最有价值的数据，其数据量大、价值高，从这些数据中提取价值是提升业务收入的一个重要手段。面对日益增长的数据量，传统的数据分析工具已经很难满足这些需求，新的开源工具可为我们解决这些问题。文中列举了很多现实中的例子及实现方案，为我们进一步挖掘数据的价值提供了一种思路。鉴于译者水平有限，有些术语及语句可能理解有误，欢迎读者发邮件和我联系：sdwangxiaoning@foxmail.com。

王小宁

2016年

前　言

数据正以惊人的速度增加，而公司要么疲于应付，要么急于利用这些数据进行分析。Hadoop 是一个优秀的开源框架，可以应付这些大数据问题。

在过去的几年里，我一直在金融部门使用 Hadoop，但在使用的过程中，一直没有发现有关 Hadoop 在金融应用中的任何案例资源或书籍。我遇到的关于 Hadoop、Hive 或一些 MapReduce 模式的书籍大都是用各种各样的方式统计单词数量或分析 Twitter 信息。

我写这本书旨在解释 Hadoop 和其他相关产品在处理金融案例大数据中的基本应用。在书中，介绍了很多案例并提供了一个非常实用的方法。

这本书包含什么

第 1 章，大数据回顾。本章包含大数据概览、前景和技术演变，也介绍了 Hadoop 架构的基本知识、组成部分和分布式框架。如果你之前已经了解 Hadoop，这一章可以忽略。

第 2 章，金融服务中的大数据。本章将延伸到站在一个金融机构的角度去看大数据。主要介绍大数据在金融部门的演进故事，在项目落地时的一些挑战，以及利用相关工具和技术处理金融案例的应用。

第 3 章，在云端使用 Hadoop。本章包含大数据在云端使用的概览，以及基于端到端数据处理的样本投资组合风险模拟项目。

第 4 章，使用 Hadoop 进行数据迁移。本章讨论了将历史数据从传统数据源迁到 Hadoop 上的几种常用项目。

第 5 章，入门。本章包含了一个非常大的企业数据平台的实施项目，以支持各种风险和监管要求。

第 6 章，变得有经验。本章给出了实时分析的概览和检测欺诈交易的样本项目。

第 7 章，深入扩展 Hadoop 的企业级应用。本章包含的主题扩展到 Hadoop 在公司中的使用，如企业数据湖、Lambda 架构和数据管理。还介绍了更多基本的财务案例与简短的解决方案。

第 8 章，Hadoop 的快速增长。本章讨论了 Hadoop 分布式架构的升级周期，并用最佳实践和标准完成此书。

阅读这本书你需要哪些基础知识

因为 Hadoop 是一个数据处理和分析的技术框架，因此在数据库、项目和分析工具上有一些经验对读者会有帮助。

这本书是一个入门指南，包含了大量外部引用的大数据产品。因此，如果在任何时候需要深入了解 Hadoop，我们鼓励读者参考书中提到的外部资源。

哪些人适合读这本书

本书主要面向致力于使用 Hadoop 的金融部门工作人员，包含数据项目开发人员、分析师、架构师和管理人员。

它也有助于来自其他行业最近转换或想将业务领域转向金融部门的技术专业人士。

这本书是一本初学者指南，涵盖了使用 Hadoop 作为金融案例主题的大部分内容，并非真正意味着深入了解 Hadoop 或提供现成代码。

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

- **提交勘误**：您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（在您购买电子书时，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】处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/31051

二维码：

作者简介

Rajiv Tiwari 是一位有着超过 15 年经验的自由大数据架构师，他的研究方向包括大数据、数据分析、数据管理、数据架构、数据清洗 / 数据整合、数据仓库，以及银行和其他金融组织中的数据智能等。

他毕业于瓦拉纳西印度理工学院（IIT）电子工程专业，在英国工作了 10 年有余，大部分时间居住在英国金融城——伦敦。从 2010 年起，Rajiv 就开始使用 Hadoop，当时银行部门使用 Hadoop 的还很少。他目前正在帮助 1 级投资银行（Tier 1 Investment Bank）在 Hadoop 平台上实施一个大型风险分析项目。

如果想联系 Rajiv，则可以通过他的网站 http://www.bigdatacloud.net 或推特 @ bigdataoncloud。

我一直认为当作家把自己的书献给他们的妻子、合作伙伴或孩子时有点俗气，但是近几个月来，让我明白了为什么一个家庭的支持对写一本书那么重要。

考虑到我目前在投资银行每天工作时间很长，且很难抽出时间来写这本书，所以，我一直在深夜和周末写这本书。我要感谢我的妻子 Seema，她几乎帮我照料一切能分散我写作注意力的东西；还有我的儿子 Rivaan。[1]

1 楷体字部分是引用作者的原话。

审稿人简介

Harshit Bakliwal 是一位印度领先的 IT 公司的 Hadoop 开发者。他有 6 年左右的工作经验和超过 3 年的大数据 /Hadoop 经验。他从 2010 年开始使用 Hadoop，当时 Hadoop 刚刚在科技界崭露头角，并没有太多的在线帮助。从那时起，他继续用自己的方式学习这门语言及其他高水平的语言，如 Pig、Hive、Sqoop、Oozie 和 HBase。现如今他能处理 4 ～ 5 个集群（每个集群大约有 200 个节点）上 PB 级的数据。

Dr.Daniel Fasel 是 Scigility 公司的创始人和 CEO。Scigility 公司为瑞士和欧洲其他国家的大规模信息系统和大数据技术提供解决方案。它的专业团队在大数据技术上有超过 7 年的极强的学术背景和实际知识经验。

他是瑞士电信（瑞士第一大电信运营商）商业智能团队的第一位数据科学家，并在就职期间实现了 NoSQL 技术在瑞士电信公司的探索性分析技术。在注重科学数据和 NoSQL 技术之前，他是合同和客户域（瑞士电信数据仓库的核心组件）的商业智能工程师。他还担任商业情报架构师和 Oracle Hyperion Essbase 立方体管理员。

他在瑞士福里堡大学（University of Fribourg）获得经济学博士学位。他写了一篇关于模糊数据仓库的文章，让他获得了最高的成绩。除了他的博士研究，他一直担任福里堡大学信息学系的系统工程师和系统管理员团队的领导。2009 年（当时大数据还不是一个流行词），他安装和维护了分布式计算集群和 NoSQL 技术。他还经常在大数据和数据仓库领域出版英语或德语的书籍与文章。

Mark Reddy 是软件工程师和分布式系统爱好者。他从爱尔兰的高威梅奥理工学院（Galway-Mayo Institute of Technology）荣誉毕业后，曾在 Hewlett-Packard 和 Avaeon Solutions 公司任职。他目前在 Boxever 工作，这是一家专

注于旅游行业大数据和预测分析的爱尔兰初创企业。他使用 Hadoop、Spark、Cassandra、ZooKeeper、Storm、Kafka 等工具设计并实现了大规模分布式的解决方案，这些系统处理的数据达 TB 级。他喜欢利用他的知识和经验为开源项目做贡献，并对行业热点话题进行公开演讲。

当他不写代码的时候，他喜欢公开演讲或写博客（http://markreddy.ie/），他也喜欢旅游、健身，以及发推特随想 @ markreddy。

目　录

第 1 章

大数据回顾

任何一个组织或个人都有一个可用来使用或分析的数字足迹。简单来说，大数据分析指的是对数量不断增长的数据的使用。

在本章中，我们将在下面几个标题的帮助下回顾一下大数据和 Hadoop：

- 大数据是什么。
- 大数据技术的演进。
- 大数据的发展前景。
- 与大数据相关的职业。
- Hadoop 架构。
- Hadoop 生态圈简介。
- Hadoop 版本。

大数据是什么

不同的咨询公司和 IT 供应商对大数据给出了不同的定义。以下是两种典型的定义。

第一种定义是“数量如此之多以至于无法用传统的数据处理工具和应用来处理的数据被称为大数据”。

第二种定义是著名的 3V 定义，也被公认为最专业的定义，即“大量（Volume）、多样（Variety）、高速（Velocity）是与大数据相关的三个属性或维度。大量指的是数据的量很大，多样指的是数据的类型很多，高速指的是数据处理的速度很快”。

数据量

关于数据量的大小一直存在争议，究竟什么样的数据才能被归为大数据呢？不幸的是，没有一个定义好的规则来对其进行分类。对于一家处理 GB（$1GB=2^{30}B$）级数据的小公司来说，TB（$1TB=2^{10}GB$）级的数据可能被认为是大数据。对于处理 TB 级数据的大公司来说，PB（$1PB=2^{10}TB$）级的数据则被认为是大数据。但无论在什么情况下，我们所说的大数据至少是 TB 级的。

对于个人和组织来说，数据正以指数级的速率在增长。没人想要抛弃数据，尤其是现在硬盘的价格一直在下降。除了想要存储无尽的数据，企业也需要分析它们。一般来说，数据是以不同的形式存储的，其中大量的交易数据被称为结构化数据，图像和音频等数据被称为非结构化数据。

数据速度

直到大约 5 年前，公司常常提取（**Exact**）、转换（**Transform**）和加载（**Load**）（简称 ETL）日常产生的批量数据，导入数据仓库或数据集市，然后使用这些平台上的商业智能分析工具对数据进行分析和处理。如今，有更多的数据源如邮件、社交媒体和交易等为我们提供数据，然而对日常批量数据的处理