火眼金睛大数据技术实训丛书

Hado

大数据

财务分析R与Hadoop实训

李晓龙 主编

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

火眼金睛大数据技术实训丛书

大数据

财务分析R与Hadoop实训

李晓龙 主编

经济管理出版社
ECONOMY & MANAGEMENT PUBLISHING HOUSE

图书在版编目（CIP）数据

大数据：财务分析 R 与 Hadoop 实训/李晓龙主编. —北京：经济管理出版社，2018. 1

ISBN 978 - 7 - 5096 - 4225 - 2

Ⅰ. ①大… Ⅱ. ①李… Ⅲ. ①会计分析 Ⅳ. ①F231. 2

中国版本图书馆 CIP 数据核字（2016）第 021164 号

组稿编辑：魏晨红
责任编辑：魏晨红
责任印制：黄章平
责任校对：超 凡

出版发行：经济管理出版社
（北京市海淀区北蜂窝 8 号中雅大厦 A 座 11 层 100038）
网 址：www. E - mp. com. cn
电 话：（010）51915602
印 刷：北京九州迅驰传媒文化有限公司
经 销：新华书店
开 本：720mm × 1000mm/16
印 张：16. 5
字 数：246 千字
版 次：2018 年 1 月第 1 版 2018 年 1 月第 1 次印刷
书 号：ISBN 978 - 7 - 5096 - 4225 - 2
定 价：48. 00 元

前　言

众所周知，企业的目标在于创造价值。而契约理论又告诉我们，企业是契约的结合体，涉及员工、客户、供应商和政府等众多的利益关系。因此，企业必须通过市场部门的员工发现客户的“痛点”，从而引导研发部门开发出解决客户“痛点”的方案和产品。而综合部门，则是使用各种资源为这一发现“痛点”到解决“痛点”的过程进行服务，而这一过程离不开信息系统和数据分析等工具的支持。古人云，“工欲善其事，必先利其器”。而本书也是为广大财务分析人员、金融工作者和信息工作者，基于大数据技术提供最基础的工具，以期抛砖引玉。

当今世界，计算机技术的日新月异，智能手机的日益普及，物联网的逐渐兴起，人工智能应用的突破，使得海量数据瞬间处理成为了可能。“互联网 +”、“大数据技术”以及“电子化交易”正在颠覆着一个又一个传统产业。

2016 年，达沃斯金融论坛全球金融精英关注的不再是风险控制的《巴塞尔协议》，而是聚焦于很可能给金融业带来颠覆性改变的区块链金融技术。与此同时，欧美对冲基金以及投行的自营盘都开始热衷于开发基于“大数据技术”的套利策略，其中最具代表性的包括温顿资本（Winton Capital）在牛津设立数据研究中心，以及瑞信信贷（Credit Suisse）对 HOLT 选股系统进行技术革新等。此外，据称大摩、小摩和高盛未来将可能共同组建大数据公司，从而为三者提供“证券产品参考数据”（Securities Product Reference Data）。大数据技术深受全民热捧，欧美投行几乎都设立了“量化”部门，而与此相关

的量化分析职位更是成为众人眼中的“香饽饽”。即使是一个入门级量化交易员也往往能够拥有 10 万美元起的年收入，堪称“金饭碗”。

与此同时，财会人员梦想的四大会计事务所，在进行传统审计、税务和咨询业务的情况下，在推出内部控制报告和社会责任报告等相关的整合报告的同时，也正在紧锣密鼓地推出基于大数据的自动审计业务。随着大数据技术的应用，势必改变传统的工作指导思想和工作流程。

要作为时代的弄潮儿，就必须掌握相应的弄潮技能。一位合格的量化交易员不仅需要拥有深厚的数学背景和坚实的金融理论，而且还必须掌握丰富的编程语言和熟练的建模技能。具体而言，技术层面上要求开发人员首先利用 Hadoop 或者 Spark 来搭建初步的模型框架，并通常会为其提供分布式文件系统，用以存储所有计算节点的数据，而后再要求其根据 R - Hadoop 或 Spark - Python/Scala 等程序语言来实现具体算法。大数据时代，R 被拉到了潮流尖端，作为免费的开源软件，随着加入的人数增多，R 的计算引擎、性能、各种程序包都得到了改进和升级，其中 R - Hadoop 可用于进行分布式计算和建立相关性算法模型，处理一些海量数据。

因此，本书创新地将财务分析与大数据相结合，通过介绍大数据技术在财务分析中的应用，培养学生应用大数据技术进行财务分析的技能。为了更好地适应新形势，我们推出了《大数据：财务分析 R 与 Hadoop 实训》一书，该书具有以下几个方面的特色：

第一，紧扣当前最新科技科研前沿。目前，市面上流通的有关财务分析的大数据系列教材总体较少，而基于 R - hadoop 语言进行财务分析的实践教材更是凤毛麟角。本书将大数据 R - hadoop 实践与财务分析紧密结合，抽丝剥茧，一步一步地引导读者学习大数据架构体系。

第二，全书内容体系编排合理。按照“大数据与财务分析基础知识—财务分析与 R - hdoop 语言—财务分析实训”的思路组织全书，既方便读者（特别是初学者）在了解财务分析和大数据基础架构的概念和技术，又能帮助读者快速使用 R - Hadoop 工具进行大数据挖掘和分析。

第三，考虑了不同群体的阅读偏好和水平。本书涉及面广，既介绍了大数据基本技术，也介绍了当前大数据最流行的架构 Hadoop 生态系统和当前最流行的开源数据分析工具 R 语言，并通过多次大数据实践将 R 语言和 Hadoop 架构相结合，详细地解析了 R - Hadoop 在财务分析领域中的应用。

本书适合准备从事金融风险管理、财务管理、证券投资、投资银行、风险投资、产业投资或公司财务工作的同学修读。通过对该课程的学习，同学可以对大数据时代下的财务分析有一定了解，更重要的是可以掌握 R 语言工具进行财务建模和分析，并在此基础上撰写研究报告和科研论文，为未来深造或工作奠定坚实的理论和实践基础。限于编者的能力和时间，本书难免存在纰漏或不足之处，欢迎读者批评指正。

在此，我非常感谢参与本书编写的所有参与单位和人员，感谢广东外语外贸大学、经济管理出版社、暨南大学、广州铭诚计算机技术有限公司、深圳国泰安教育技术股份公司的支持，感谢本书编辑魏晨红的辛勤校对，也感谢陈荣腾、钟福海、胡少柔、林小龙、杨金壶、梁敏耀、李丽、陈鹏、张金秀、黄维付、陈贤斌和李婷婷等同事对本书底稿的编写和修改。

李晓龙
广东外语外贸大学
2015.9

目　录

第一部分　大数据与财务分析

第三部分 财务分析实训

第一部分
大数据与财务分析

金融的实质就是“把适量的钱投到适合的位置”（Put the “Right” Money in the “Right” Place），从而以适度的金额购买适当的产品（Inorder to get the “Right” Amount for the “Right” Price）。财务分析意味着，投资不是赌博而是博弈，理性的投资者应该学会运用投资策略来实现财富增值。

如何才能将模糊、抽象的策略变成具体可信的数字呢？那么就应该使用大数据技术，将投资策略通过数学模型和计算机代码数量化，从而基于数据分析和动态模拟而合理预测其投资行为的未来走势。具体而言，财务分析人员应该使用大数据技术识别风险指标，构建定价模型结果或者交易策略，同时根据实际情况略微修改参数，最终实现自己的资产配置及投资组合。随着大数据的发展，众多金融公司和金融交易场所已变成了信息技术（IT）或大数据技术（DT）人员的集结地。例如花旗、摩根大通及瑞士信贷等在内的众多欧美顶尖投行，都在不计血本地培养自己的IT团队，并命其专门从事产品模型研发，从而得以跻身于“得模型者得天下”的金融大战之中。这些拥有专属开发任务的IT团队往往被称为量化团队，专门从事量化投资分析以及衍生品定价策略。

除金融市场的参与者都欲借“量化技术”的东风大展拳脚外，众多欧美金融监管机构也针对金融技术的兴起顺势推出了相关的监管政策。英国《金融时报》曾撰写过《英国监察机构检测保险公司对

于大数据的使用》一文。文章指出，英国金融市场行为监管局（FCA）已正式发表声明，表示会继续监视金融技术开发以及金融技术对于公司和投资者的影响，如开展一项专门针对“保险公司大数据使用现状”的市场调查，从而更为精准、有效地打击预防以金融技术为核心的新型金融犯罪行为。这充分说明，在“互联网 +”时代中，“大数据技术”已成为财务分析必不可少的一项技术。本部分内容将介绍大数据基础概念、财务分析基础知识和大数据基础技术。

第 1 章　大数据基础概念

大数据时代早已到来。《大数据时代》的作者维克托·迈尔·舍恩伯格说："世界的本质就是数据，大数据将开始一次重大的时代转型。"其实早在 1980 年，美国著名未来学者托夫勒便在《第三次浪潮》一书中提出"数据就是财富"的观点，将大数据热情地赞颂为"第三次浪潮的华彩乐章"。作为云计算领域的重要延伸，大数据正在引领信息革命进入新的时代。2001 年，全球最具权威的 IT 研究与顾问咨询公司 Gartner 提出大数据面临"4 个 V"的挑战，《自然杂志》（2008）推出《大数据》专刊来全方位地介绍大数据问题，美国总统奥巴马（2012）将数据定义为"未来的新石油"。2013 年，Gartner 在一篇报告中指出，64% 的受访企业都表示他们正在或是即将进行大数据工作。信息技术、计算机技术和互联网技术的迅速发展，使得人类社会各类数据呈爆炸性增长，对这些复杂大数据的有效管理，已成为当前社会的热点问题。

1.1　大数据基础概念

大数据（Big Data），或称巨量资料，是指所涉及的资料规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策的资讯。大数据一般指 10TB（1TB = 1024GB）规模以上的数据量，其基本特征可以用"4 个 V"来总结：