



火眼金睛大数据技术实训丛书

R Hadoop

大数据

财务分析R与Hadoop实训

李晓龙 主编



经济管理出版社

ECONOMIC MANAGEMENT PRESS



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

科学技术

图书馆藏

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

· 版权所有 翻印必究 ·

凡购本社图书, 如有印装错误, 由本社读者服务部负责调换。

联系地址: 北京阜外月坛北小街 2 号

电话: (010) 68022974 邮编: 100836

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科技。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 - hadoop 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1章 大数据基础概念	3
1.1 大数据基础概念	3
1.2 大数据缘起特征	4
1.2.1 数据产生由企业内部向企业外部扩展	4
1.2.2 数据产生由 Web1.0 向 Web2.0、 由互联网向移动互联网扩展	5
1.2.3 数据产生由计算机/互联网 (IT) 向物联网 (IOT) 扩展	5
1.3 大数据应用案例	6
1.3.1 余额宝的业务背景	6
1.3.2 余额宝的系统建设	7
1.4 大数据人才培养	9
第2章 财经数据分析基础	11
2.1 会计基础概念	11
2.1.1 会计基本职能	12
2.1.2 会计基本目标	12
2.1.3 会计核心要素	12



2.1.4	会计记账方法	15
2.1.5	会计核算原则	15
2.2	资产负债表	16
2.2.1	资产负债表概述	16
2.2.2	企业资产分析	17
2.2.3	企业负债分析	18
2.2.4	企业权益分析	18
2.2.5	资产负债分析其他事项	19
2.3	损益表	21
2.3.1	收入分析	21
2.3.2	支出分析	22
2.3.3	影响事项	22
2.3.4	市盈率分析	23
2.4	现金流量表	24
2.4.1	现金流量表内容	24
2.4.2	现金流量表计算	25
2.4.3	自由现金流	26
第3章	大数据基础技术	28
3.1	数据采集和清洗	28
3.2	数据库和数据存储	29
3.2.1	信息世界	30
3.2.2	数据世界	30
3.2.3	实体—联系方法 (Entity - Relationship Approach)	30
3.2.4	数据模型	31
3.2.5	SQL 的发展	31
3.3	数据挖掘和数据分析	35
3.3.1	聚类分析：如基于历史的 MBR 分析、 遗传算法	36
3.3.2	关联分析：如购物篮分析	36

3.3.3 分类分析：如决策树、判别分析	36
3.4 大数据分析 with R	37

第二部分 财务分析 R 与 Hadoop 语言

第 4 章 R 语言简介	41
4.1 R 语言概述	41
4.1.1 R 语言的优势	41
4.1.2 其他常用统计软件	42
4.2 在 Windows 下获取和安装 R 软件	43
4.2.1 RGui 界面	43
4.2.2 RStudio 界面	46
4.2.3 R 语言的帮助 (Help)	47
4.3 在 Linux 上搭建 R 环境	49
4.3.1 搭建前的准备工作	49
4.3.2 下载与解压 (以 R-2.15.3 为例)	49
4.3.3 编译	50
4.3.4 安装	50
4.4 R 包 (Packages)	51
4.4.1 如何寻找相关的 Packages	51
4.4.2 安装 Packages	54
4.4.3 调用 Packages	59
4.5 工作目录和工作空间	61
4.5.1 获取和设定工作目录	61
4.5.2 工作空间的保存	62
4.5.3 系统设置	64
第 5 章 R 语言基本操作	65
5.1 赋值和运算	65
5.1.1 赋值	65



5.1.2	简单运算	66
5.2	数据结构	67
5.2.1	向量的建立	68
5.2.2	矩阵的建立	71
5.2.3	数组的建立	73
5.2.4	数据框的建立	74
5.3	导入和导出数据	79
5.3.1	导入数据	79
5.3.2	导出数据	81
5.3.3	Excel 数据的导入	81
5.3.4	导入导出数据时需要注意的问题	82
5.4	数据的管理	84
5.4.1	数据排序 (order/sort)	84
5.4.2	数据集的合并 (insert)	86
5.4.3	剔除变量 (delete)	91
5.4.4	数据集提取 (select)	92
5.4.5	subset () 函数	94
5.4.6	缺失值的处理	94
5.5	常用函数	97
第 6 章	Hadoop 简介	100
6.1	Hadoop	100
6.1.1	Hadoop 概述	100
6.1.2	Hadoop 的功能和特点	100
6.1.3	Hadoop 的发展与现状	101
6.1.4	Hadoop 的核心架构	102
6.2	Hadoop 的数据管理	104
6.2.1	HDFS	104
6.2.2	HBase	107
6.2.3	Hive	110
6.3	ZooKeeper 原理	113

6.3.1 ZooKeeper 的基本概念	113
6.3.2 ZooKeeper 在 Hadoop 及 HBase 中的具体作用	114
6.4 Hadoop 大数据处理的意义	114

第三部分 财务分析实训

第 7 章 财经大数据探索性分析	125
7.1 探索性分析常用函数	126
7.1.1 数学函数	126
7.1.2 统计函数	127
7.1.3 分类分组函数	130
7.1.4 概率函数	133
7.1.5 日期函数	134
7.1.6 极端值处理	137
7.2 盈利能力分析	138
7.2.1 查询盈利能力数据	138
7.2.2 导入盈利能力数据	138
7.2.3 合并盈利能力数据	139
7.2.4 计算所有公司的盈利能力	140
7.2.5 计算万科的盈利能力	141
7.2.6 计算房地产的盈利能力	141
7.2.7 计算竞争对手的盈利能力	142
7.2.8 导出盈利能力 roa 数据	143
7.3 偿债能力分析	143
7.3.1 查询偿债能力数据	143
7.3.2 导入偿债能力数据	143
7.3.3 合并偿债能力数据	145
7.3.4 计算所有公司的偿债能力	145
7.3.5 计算万科的偿债能力	146
7.3.6 计算房地产的偿债能力	146



7.3.7	计算竞争对手的偿债能力	147
7.3.8	导出偿债能力 Lev 数据	148
7.4	经营能力分析	148
7.4.1	查询经营能力数据	148
7.4.2	导入经营能力数据	148
7.4.3	合并经营能力数据	149
7.4.4	计算所有公司的经营能力	150
7.4.5	万科经营能力分析	151
7.4.6	计算房地产的经营能力	151
7.4.7	导出经营能力 yingshou 数据	152
7.5	风险水平分析	152
7.5.1	查询风险水平数据	153
7.5.2	导入风险水平数据	153
7.5.3	合并风险水平数据	154
7.5.4	计算所有公司的风险水平	154
7.5.5	计算万科的风险水平	155
7.5.6	计算房地产的风险水平	156
7.5.7	导出风险水平 DFL 数据	157
第 8 章	财经大数据可视化分析	158
8.1	指标可视化概述	158
8.1.1	图形的最基础构成	158
8.1.2	绘图函数分类	159
8.2	绘图参数	161
8.2.1	高级绘图参数	161
8.2.2	低级绘图参数	165
8.2.3	点符号、线条与颜色	168
8.2.4	标题、坐标轴与图例	172
8.2.5	一幅图多个图表	175
8.3	高级绘图函数	178
8.3.1	通用二维图	178

8.3.2	饼图	179
8.3.3	箱线图	181
8.3.4	条形图	185
8.3.5	直方图	188
8.3.6	核密度图	190
8.4	低级绘图函数	193
8.5	ggplot2	195
8.5.1	ggplot2 简介	195
8.5.2	几个基本概念	196
8.5.3	图层控制与直方图	197
8.5.4	位置调整与条形图	199
8.5.5	散点图	203
第 9 章	财经大数据相关回归分析	206
9.1	回归模型基本原理	206
9.2	普通最小二乘估计 (OLS)	207
9.3	极大似然估计 (MLE)	207
9.4	线性回归模型应用分析	207
9.4.1	查询下载指数数据	209
9.4.2	导入指数数据	209
9.4.3	合并数据	210
9.4.4	缺失值处理	210
9.4.5	探索性统计分析	211
9.4.6	导入更新的数据	213
9.4.7	数据的回归模型	217
第 10 章	财经大数据时间序列分析	220
10.1	金融时间序列及其特征	220
10.2	ARMA 模型	223
10.2.1	ARMA 模型简介	223
10.2.2	ARMA 模型定阶	224



10.2.3	ARMA 模型拟合	226
10.3	异方差时间序列模型	227
10.3.1	GARCH 模型简介	227
10.3.2	GARCH 模型拟合	228
10.4	多项式回归	232
10.5	分位数回归	236
10.5.1	分位数回归拟合	236
10.5.2	分位数回归与 VaR	239
附 录	242
附表 1	2012 年我国各地区科技发展状况	242
附表 2	2015 年 1 月 4 日至 12 月 30 日上证 综合收盘价格指数	243
参考文献	247
后 记	249

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（In order 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on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科技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科技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：