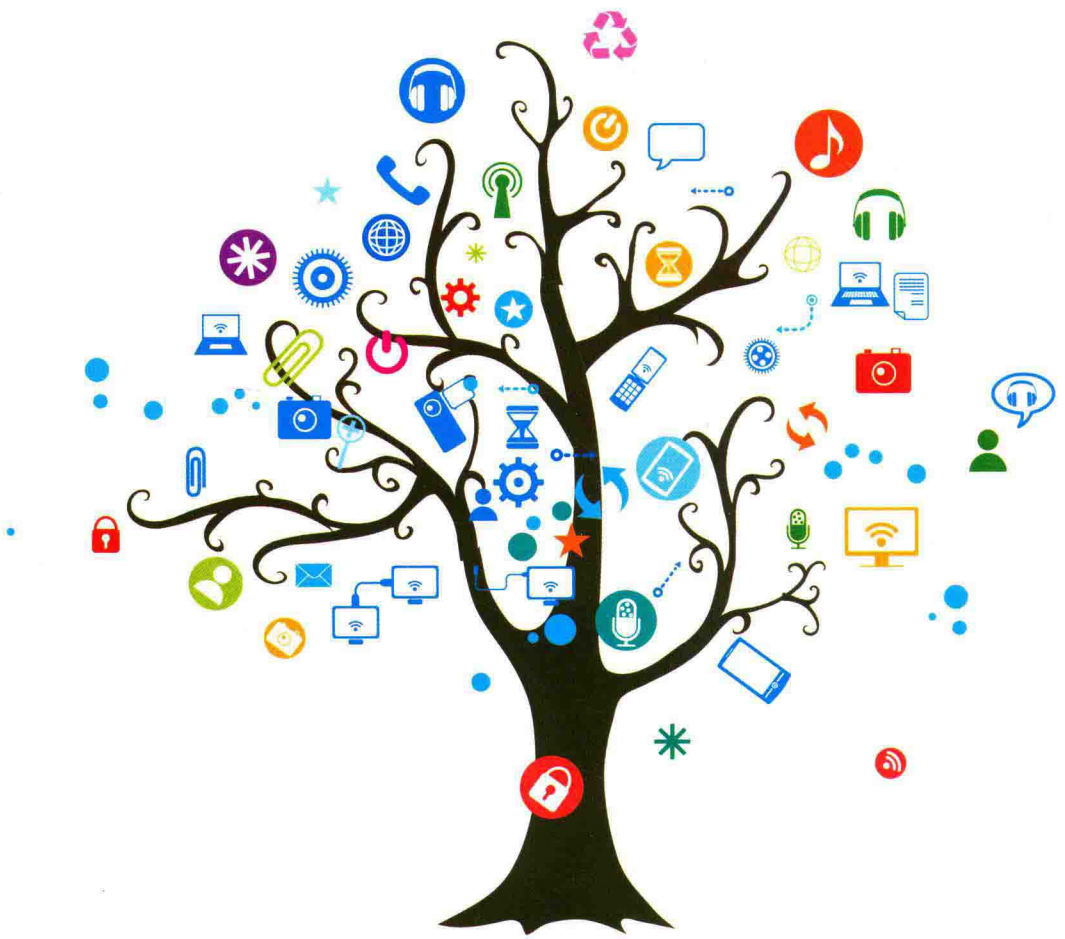
全国高校大数据教育联盟系列教材

数据科学

理论与实践

Data Science Theory and Practice

朝乐门　编著



清华大学出版社

全国高校大数据教育联盟系列教材

数据科学

理论与实践

Data Science Theory and Practice

朝乐门 编著

清华大学出版社
北京

内 容 简 介

本书重点讲解数据科学的核心理论与代表性实践，在编写过程中充分借鉴了国外著名大学设立的相关课程以及全球畅销的外文专著，而且也考虑到了国内相关课程定位与专业人才的培养需求。

全书共包括 8 个部分：数据科学的基础理论、理论基础、流程与方法、技术与工具、数据产品及开发、典型案例及实践、R 语言学习笔记与参考手册以及数据科学的重要资源。

本书的读者范围广，可以满足数据科学与大数据技术、计算机科学与技术、管理工程、工商管理、数据统计、数据分析、信息管理与信息系统等多个专业的老师、学生（含硕士生和博士生）的教学与自学需要。作者以本教材为基础，将提供 MOOC 公开课，助力培养数据科学领域的人才。

图书在版编目(CIP)数据

数据科学理论与实践/朝乐门编著. —北京：清华大学出版社，2017
（全国高校大数据教育联盟系列教材）
ISBN 978-7-302-48054-9

Ⅰ. ①数… Ⅱ. ①朝… Ⅲ. ①数据处理－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 208851 号

责任编辑：刘向威
封面设计：文　静
责任校对：梁　毅
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：19.5　　**字　　数**：475 千字
版　　次：2017 年 11 月第 1 版　　**印　　次**：2017 年 11 月第1次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：076962-01

编　委　会

这是一本值得
推荐的优秀教材！

陈国良

2017.5.18

翁老师的《数据科学理论与实践》是一本通俗易懂且充满智慧，读了之后颇有收获与感动的精品教材，让我觉得相见恨晚！

庞艳蓓

2017.6.20

Data Science is transforming every sphere of human endeavour. This book is an invaluable resource to anyone who wants to create the future.

Leon Katsnelson
CTO, IBM Analytics Emerging Technologies

preface 序

大数据时代的到来催生了一门新学科——数据科学，并在全球范围内引发了相关课程和专业建设的大讨论。核心教材的开发是课程设计与专业建设的关键环节。作为一门新兴学科，数据科学与大数据技术类课程亟待一批优秀教材来揭示其核心理论体系及代表性实践。为此，全国高校大数据教育联盟于2016年牵头成立数据科学与大数据技术教材专家指导委员会，并特邀中国人民大学朝乐门老师主持《数据科学理论与实践》一书的编写工作。

不负众望，朝乐门老师完成了一本极具特色的、高水平优秀教材。本书的主要特色包括以下四点。

一是坚持系统性与重点突出并重。本书并不是相关知识的简单汇编，不仅给出了数据科学的知识体系，而且还重点讲解了一些关键细节性知识和新知识，如数据产品开发、数据加工、数据故事化描述和数据连续性保障等。

二是重视理论与实践相结合。数据科学是一门实践性很强的课程，不但需要扎实的理论功底，而且还要具备丰富的实战经验。为此，本书吸收了国内外重要的研究进展与实践经验。

三是遵循教与学的规律。每章的开始和结尾之处分别给出了“如何开始学习”和“如何继续学习”，并以图表、小故事形式解读重要知识点，使得原本有所“枯燥”的课程变得“有趣”，不仅提升了教材的可读性，更重要的是培养了学生的学习信心与兴趣。

四是力争继承与创新相结合。本书不仅吸收了国内外相关领域的最新研究成果，而且结合作者自己的研究，有很强的系统性和前瞻性，体现了作者的科学态度、坚实理论功底和独到见解。

继《数据科学》一书出版以来，朝乐门老师专注于数据科学与大数据技术的研究，在数据科学理论与实践方面做出了诸多有益探索。他的认真负责、开拓进取、刻苦钻研的做事态度值得鼓励。在此，也希望更多的专家学者加入数据科学队伍之中，本书将是带您走进数据科学与大数据技术之门的金钥匙。

陈国良

2017.5.18

contents 目录

第 1 章　基础理论 …… 1
如何开始学习 …… 1
1.1　术语定义 …… 2
1.2　研究目的 …… 7
1.3　发展简史 …… 11
1.4　理论体系 …… 14
1.5　基本原则 …… 16
1.6　数据科学家 …… 23
如何继续学习 …… 28
习题 …… 32
参考文献 …… 32
第 2 章　理论基础 …… 34
如何开始学习 …… 34
2.1　数据科学的学科地位 …… 35
2.2　统计学 …… 36
2.3　机器学习 …… 42
2.4　数据可视化 …… 56
如何继续学习 …… 58
习题 …… 59
参考文献 …… 59
第 3 章　流程与方法 …… 61
如何开始学习 …… 61
3.1　基本流程 …… 62
3.2　数据加工 …… 70
3.3　数据审计 …… 79
3.4　数据分析 …… 83
3.5　数据可视化 …… 86
3.6　数据故事化 …… 96
3.7　项目管理 …… 101
如何继续学习 …… 103

习题 …… 104
参考文献 …… 104
第 4 章 技术与工具 …… 106
如何开始学习 …… 106
4.1 技术体系 …… 107
4.2 MapReduce …… 112
4.3 Hadoop …… 119
4.4 Spark …… 129
4.5 NoSQL 与 NewSQL …… 136
4.6 R 与 Python …… 148
4.7 发展趋势 …… 149
如何继续学习 …… 153
习题 …… 154
参考文献 …… 154
第 5 章 数据产品及开发 …… 156
如何开始学习 …… 156
5.1 定义 …… 157
5.2 主要特征 …… 159
5.3 关键活动 …… 163
5.4 数据柔术 …… 164
5.5 数据能力 …… 175
5.6 数据战略 …… 183
5.7 数据治理 …… 185
如何继续学习 …… 188
习题 …… 189
参考文献 …… 189
第 6 章 典型案例及实践 …… 190
如何开始学习 …… 190
6.1 2012 年美国总统大选 …… 191
6.2 统计分析 …… 195
6.3 机器学习 …… 198
6.4 数据可视化 …… 213
6.5 SparkR 编程 …… 215
如何继续学习 …… 229
习题 …… 229
参考文献 …… 230
附录 A R 语言学习笔记与参考手册 …… 231
学习与参考指南 …… 231
1. R 变量定义方法 …… 233

2. R 语句的写法 …… 234
3. R 中的赋值语句 …… 234
4. R 的文件读写路径——当前工作目录 …… 235
5. R 的变量查找机制——搜索路径 …… 235
6. R 中查看帮助的方法 …… 237
7. R 区分大小写字母 …… 237
8. R 的注释 …… 238
9. R 的语句 …… 238
10. R 中的变量命名规范 …… 238
11. R 中的关键字/保留字 …… 239
12. R 中的默认数据类型 …… 240
13. R 中数据类型的判断与强制类型转换 …… 241
14. R 中的向量 …… 241
15. R 中的列表 …… 245
16. R 中的数据框 …… 248
17. R 中的因子类型 …… 253
18. R 中的循环语句 …… 254
19. R 中的选择语句 …… 255
20. R 中的特殊运算符 …… 256
21. R 中的快速生成数列的方法 …… 257
22. R 中自定义函数 …… 258
23. R 中常用的数学函数 …… 259
24. R 中的字符串处理函数 …… 260
25. R 中的常用统计函数 …… 262
26. R 中的随机数 …… 263
27. R 包的用法 …… 264
28. R 中的矩阵 …… 265
29. R 中的数组 …… 268
30. R 中的面向对象编程 …… 269
31. R 中的 S4 类 …… 270
32. R 中的数据可视化 …… 273
33. R 的输入输出 …… 274
34. R 中的正则表达式 …… 275
35. R 的数据集 …… 278
36. R 第三方包——R 的灵魂 …… 279
37. 基于 R 的数据加工 …… 280
附录 B 数据科学的重要资源 …… 285
1. 学术期刊 …… 285
2. 国际会议 …… 285

3. 研究机构 …… 286
4. 课程资源 …… 286
5. 硕士学位项目 …… 286
6. 专家学者 …… 287
7. 相关工具 …… 287
附录 C 术语索引 …… 288
后记 …… 293

contents 目录

图 1-1 DIKW 金字塔 …… 2
图 1-2 数据与数值的区别 …… 3
图 1-3 大数据的本质 …… 4
图 1-4 大数据的特征 …… 5
图 1-5 DIKUW 模型及应用 …… 8
图 1-6 数据洞见 …… 8
图 1-7 业务数据化与数据业务化 …… 9
图 1-8 常用驱动方式 …… 9
图 1-9 数据的层次性 …… 9
图 1-10 大数据生态系统 …… 10
图 1-11 Gartner 技术成熟度曲线 …… 11
图 1-12 数据科学的萌芽期(1974—2009 年) …… 11
图 1-13 数据科学的快速发展期(2010—2013 年) …… 12
图 1-14 数据科学的逐渐成熟期(2014 年至今) …… 13
图 1-15 数据科学的理论体系 …… 14
图 1-16 数据科学的主要内容 …… 15
图 1-17 数据科学的“三世界原则” …… 16
图 1-18 数据科学的“三个要素”及“3C 精神” …… 17
图 1-19 数据范式与知识范式的区别 …… 18
图 1-20 数据管理范式的变化 …… 18
图 1-21 大数据的资产属性 …… 19
图 1-22 常用驱动方式 …… 20
图 1-23 CAPTCHA 项目 …… 21
图 1-24 ReCAPTCHA 项目 …… 21
图 1-25 数据与算法之间的关系 …… 22
图 1-26 BellKors Pragmatic Chaos 团队获得 Netflix 奖 …… 23
图 1-27 Netflix 奖公测结果 …… 23
图 1-28 RStudio 中编辑 Markdown 的窗口 …… 25
图 1-29 数据科学家团队 …… 28
图 1-30 学习数据科学的四则原则 …… 29

图 2-1 数据科学的理论基础 …… 35
图 2-2 统计方法的分类(目的与思路视角) …… 37
图 2-3 统计学中的数据推断 …… 37
图 2-4 数据统计方法的类型(方法论视角) …… 38
图 2-5 数据统计基本方法 …… 38
图 2-6 元分析与基本分析 …… 39
图 2-7 GFT 预测与美国疾病控制中心数据的对比 …… 40
图 2-8 GFT 估计与实际数据的误差(2013 年 2 月) …… 40
图 2-9 大数据时代的思维模式的转变 …… 41
图 2-10 西洋双陆棋 …… 42
图 2-11 自动驾驶 …… 42
图 2-12 机器学习的基本思路 …… 42
图 2-13 机器学习的三要素 …… 44
图 2-14 机器学习的类型 …… 45
图 2-15 KNN 算法的基本步骤 …… 46
图 2-16 决策树示例——识别鸟类 …… 48
图 2-17 感知器示例 …… 50
图 2-18 前向神经网络 …… 50
图 2-19 归纳学习与分析学习 …… 52
图 2-20 增强学习 …… 53
图 2-21 IBM Watson …… 54
图 2-22 Pepper 机器人 …… 54
图 2-23 机器学习及其应用 …… 55
图 2-24 Anscombe 四组数据的可视化 …… 57
图 2-25 John Snow 的鬼地图(Ghost Map) …… 58
图 3-1 数据科学的基本流程 …… 62
图 3-2 量化自我 …… 63
图 3-3 整齐数据与干净数据的区别 …… 64
图 3-4 整齐数据示意图 …… 64
图 3-5 残差 …… 67
图 3-6 数据分析的类型 …… 68
图 3-7 Analytics 1.0～3.0 …… 69
图 3-8 数据加工方法 …… 70
图 3-9 数据审计与数据清洗 …… 71
图 3-10 缺失数据的处理步骤 …… 72
图 3-11 冗余数据的处理方法 …… 72
图 3-12 数据分箱处理的步骤与类型 …… 73
图 3-13 均值平滑与边界值平滑 …… 74
图 3-14 内容集成 …… 76

图 3-15 结构集成 …… 76
图 3-16 数据脱敏处理 …… 77
图 3-17 数据连续性的定义及重要性 …… 81
图 3-18 可视化审计 …… 82
图 3-19 Gartner 分析学价值扶梯模型 …… 83
图 3-20 冰激凌的销量与谋杀案的发生数量 …… 84
图 3-21 数据分析的类型 …… 86
图 3-22 拿破仑进军俄国惨败而归的历史事件的可视化 …… 87
图 3-23 可视化分析学的相关学科 …… 87
图 3-24 可视化分析学模型 …… 88
图 3-25 数据可视化的方法体系 …… 89
图 3-26 视觉图形元素与视觉通道 …… 89
图 3-27 雷达图示例 …… 89
图 3-28 齐美尔连带 …… 89
图 3-29 视觉隐喻的示例——美国政府机构的设置 …… 90
图 3-30 地铁路线图的创始人 Henry Beck …… 91
图 3-31 Henry Beck 的伦敦地铁线路图 …… 91
图 3-32 视觉突出的示例 …… 92
图 3-33 完图法则的示例 …… 92
图 3-34 视觉通道的选择与展示 …… 94
图 3-35 视觉通道的精确度对比 …… 94
图 3-36 视觉通道的可辨认性——某公司产品销售示意图 …… 95
图 3-37 视觉通道的可分离性差 …… 95
图 3-38 上下文导致视觉假象 1 …… 96
图 3-39 上下文导致视觉假象 2 …… 96
图 3-40 对亮度和颜色的相对判断容易造成视觉假象的示例 …… 96
图 3-41 数据可视化与数据故事化描述 …… 97
图 3-42 数据的故事化描述及故事的展现 …… 98
图 3-43 项目管理的主要内容 …… 101
图 3-44 数据科学项目的基本流程 …… 102
图 4-1 大数据产业全景图 …… 107
图 4-2 基础设施 …… 108
图 4-3 分析工具 …… 109
图 4-4 企业应用 …… 109
图 4-5 行业应用 …… 110
图 4-6 跨平台基础设施和分析工具 …… 110
图 4-7 开源系统 …… 111
图 4-8 数据源与 Apps …… 111
图 4-9 数据资源 …… 111

图 4-10 MapReduce 执行过程 …… 113
图 4-11 MapReduce 对中间数据的处理 …… 116
图 4-12 以 MapReduce 为核心和以 YARN 为核心的软件栈对比 …… 118
图 4-13 下一代 MapReduce 框架 …… 118
图 4-14 Apache Hadoop 官网 …… 119
图 4-15 Apache Hadoop 生态系统 …… 119
图 4-16 Hadoop MapReduce 数据处理流程 …… 121
图 4-17 Apache Hive 官网 …… 122
图 4-18 Apache Pig 官网 …… 123
图 4-19 Apache Mahout 官网 …… 124
图 4-20 Apache HBase 官网 …… 125
图 4-21 HBase 与 Hadoop 项目 …… 125
图 4-22 HBase 逻辑模型 …… 126
图 4-23 Apache Zookeeper 官网 …… 127
图 4-24 Apache Flume 官网 …… 127
图 4-25 Apache Sqoop 官网 …… 128
图 4-26 Spark 技术架构 …… 130
图 4-27 Spark 执行流程 …… 130
图 4-28 Spark Scheduler …… 134
图 4-29 传统关系数据库的优点与缺点 …… 136
图 4-30 关系数据库与 NoSQL 数据库之间的关系 …… 138
图 4-31 NoSQL 数据分布的两个基本途径 …… 139
图 4-32 分片处理 …… 140
图 4-33 主从复制 …… 141
图 4-34 对等复制 …… 142
图 4-35 数据不一致性 …… 142
图 4-36 CAP 理论 …… 143
图 4-37 Memcached 官网 …… 146
图 4-38 一致性散列的分配方式 …… 147
图 4-39 服务器增加时的变化 …… 147
图 4-40 计算模式的演变 …… 150
图 4-41 数据管理的新变化 …… 152
图 5-1 数据产品开发中的数据与数据柔术 …… 157
图 5-2 数据产品开发与数据柔术 …… 160
图 5-3 数据产品的多样性 …… 160
图 5-4 数据产品的层次性 …… 161
图 5-5 Google 全球商机洞察(Global Market Finder) …… 162
图 5-6 数据产品链 …… 162
图 5-7 传统产品开发与数据产品开发的区别 …… 163

图 5-8 D J Patil …… 164
图 5-9 UI(User Interface)设计方案与设计思维 …… 165
图 5-10 Goolge 搜索的用户体验 …… 166
图 5-11 人与计算机图像内容识别能力的不同 …… 166
图 5-12 Amazon MTurk 平台 …… 168
图 5-13 HIT 生命周期 …… 168
图 5-14 基于人与计算机的数据处理成本曲线 …… 169
图 5-15 亚马逊数据产品：其他商家(Other Sellers) …… 170
图 5-16 LinkedIn 的数据产品——你可能认识的人们(People You May Know) …… 170
图 5-17 LinkedIn 的数据产品——你的观众是谁 …… 171
图 5-18 逆向交互定律 …… 172
图 5-19 LinkedIn 数据产品——职位推荐 …… 174
图 5-20 LinkedIn 的数据产品——帮助你的朋友找到工作 …… 174
图 5-21 Facebook 的良好用户体验 …… 175
图 5-22 DMM 基本思路 …… 175
图 5-23 CMM 基本思想 …… 177
图 5-24 CMM 成熟度等级 …… 177
图 5-25 DMM 关键过程域 …… 178
图 5-26 DMM 层级划分及描述 …… 180
图 5-27 IDEAL 模型 …… 182
图 5-28 机构数据管理能力成熟度评估结果的可视化 …… 182
图 5-29 数据战略与目标的区别 …… 183
图 5-30 数据战略的目标 …… 183
图 5-31 数据战略的侧重点 …… 184
图 5-32 数据战略的影响因素 …… 184
图 5-33 数据管理与数据治理的区别 …… 185
图 5-34 IBM 提出的企业数据管理的范畴 …… 186
图 5-35 数据治理的 PDCA 模型 …… 187
图 5-36 DGI 数据治理框架 …… 188
图 6-1 奥巴马 2012 年总统竞选芝加哥总部 …… 191
图 6-2 George Clooney …… 192
图 6-3 Sarah Jessica Parker …… 192
图 6-4 奥巴马及快速捐赠计划 …… 193
图 6-5 奥巴马通过 Reddit 与选民互动 …… 195
图 6-6 女性体重与身高的线性回归分析 …… 198
图 6-7 K-Means 算法的基本步骤 …… 208
图 6-8 工资数据的可视化 …… 215
图 6-9 起飞延误时间 …… 227
图 6-10 落地延误时间 …… 227
图附 A-1 如何参考附录 A …… 232

contents 表目录

表 1-1 结构化数据、非结构化数据与半结构化数据的区别与联系 …… 3
表 1-2 某数据科学家的画像(Profile) …… 27
表 2-1 参数估计与假设检验 …… 37
表 2-2 机器学习的相关学科 …… 45
表 2-3 已知 6 部电影的类型及片中出现的接吻和打斗次数 …… 47
表 2-4 已知电影与未知电影的距离 …… 47
表 2-5 分析学习和归纳学习的比较 …… 52
表 2-6 Anscombe 四组数据(Anscombe's Quartet) …… 56
表 3-1 测试数据 A …… 65
表 3-2 测试数据 B …… 65
表 3-3 测试数据 C …… 65
表 3-4 Pew 论坛部分人员信仰与收入数据统计(整齐化处理之前) …… 65
表 3-5 Pew 论坛部分人员信仰与收入数据统计(整齐化处理之后) …… 66
表 3-6 探索性统计中常用的集中趋势统计量 …… 67
表 3-7 探索性统计中常用的离散程度统计量 …… 67
表 3-8 探索性统计中常用的数据分布统计量 …… 67
表 3-9 数据变换的类型 …… 74
表 3-10 十进制首位数字的出现概率 …… 80
表 3-11 数据分析中常见错误 …… 85
表 3-12 数据类型及所支持的操作类型 …… 93
表 3-13 数据类型与视觉通道的对应关系图 …… 93
表 3-14 数据故事化中的"应该"与"不应该" …… 99
表 3-15 数据科学项目中的主要角色及其任务 …… 101
表 4-1 RDD 常用的 Transformation …… 132
表 4-2 RDD 常用的 Actions …… 132
表 4-3 RDD 的存储级别 …… 133
表 4-4 R 与 Spark 数据类型的映射关系 …… 135
表 4-5 典型云数据库产品 …… 138
表 4-6 NoSQL 中常用的数据模型 …… 139
表 4-7 R 与 Python 对比 …… 149

表 4-8 云计算的层次性 …… 151
表 5-1 数据转换与数据加工的区别 …… 158
表 5-2 谷歌十大产品与服务 …… 158
表 5-3 数据管理成熟度模型的过程域分类 …… 179
表 6-1 数据集 Women …… 195
表 6-2 Protein 数据集 …… 206
表 6-3 Salaries 数据集 …… 213
表 6-4 Spark 版本差异性 …… 220
表 6-5 Spark 与 R 的数据类型对比 …… 220
表 6-6 SparkR 与 sparklyr 比较 …… 228