

朱春旭◎编著

# Python

## 数据分析与 大数据处理

# 从入门到精通

45个“新手问答”

17个章节的“实训”

3个综合项目实战

50道Python面试题精选

教你轻松玩转数据分析与大数据处理



北京大学出版社  
PEKING UNIVERSITY PRESS

# Python

## 数据分析与 大数据处理

# 从入门到精通

朱春旭◎编著



北京大学出版社  
PEKING UNIVERSITY PRESS

## 内 容 提 要

本书主要讲解数据分析与大数据处理所需的技术、基础设施、核心概念、实施流程。从编程语言准备、数据采集与清洗、数据分析与可视化，到大型数据的分布式存储与分布式计算，贯穿了整个大数据项目开发流程。本书轻理论、重实践，目的是让读者快速上手。

第1篇首先介绍了Python的基本语法、面向对象开发、模块化设计等，掌握Python的编程方式。然后介绍了多线程、多进程及其相互间的通信，让读者对分布式程序有个基本的认识。

第2篇介绍了网络数据采集、数据清洗、数据存储等技术。

第3篇介绍了Python常用的数据分析工具，扩展了更多的数据清洗、插值方法，为最终的数据可视化奠定基础。

第4篇是大数据分析的重点。首先介绍了Hadoop的框架原理、调度原理，MapReduce原理与编程模型、环境搭建，接着介绍了Spark框架原理、环境搭建方式，以及如何与Hive等第三方工具进行交互，还介绍了最新的结构化流式处理技术。

第5篇通过三个项目实例，综合介绍了如何分析网页、如何搭建分布式爬虫、如何应对常见的反爬虫、如何设计数据模型、如何设计架构模型、如何在实践中综合运用前四篇涉及的技术。

本书既适合非计算机专业的编程“小白”，也适合刚毕业或即将毕业走向工作岗位的广大毕业生，以及已经有编程经验，但想转行做大数据分析的专业人士。同时，还可以作为广大职业院校、电脑培训班的教学参考用书。

## 图书在版编目(CIP)数据

Python数据分析与大数据处理从入门到精通 / 朱春旭编著. — 北京 : 北京大学出版社, 2019.11  
ISBN 978-7-301-30765-6

I. ①P… II. ①朱… III. ①软件工具—程序设计IV. ①TP311.561

中国版本图书馆CIP数据核字(2019)第204927号

书 名 Python数据分析与大数据处理从入门到精通

Python SHUJU FENXI YU DASHUJU CHULI CONG RUMEN DAO JINGTONG

著作责任者 朱春旭 编著

责任编辑 吴晓月 孙 宜

标准书号 ISBN 978-7-301-30765-6

出版发行 北京大学出版社

地 址 北京市海淀区成府路205号 100871

网 址 <http://www.pup.cn> 新浪微博: @北京大学出版社

电子信箱 [pup7@pup.cn](mailto:pup7@pup.cn)

电 话 邮购部 010-62752015 发行部 010-62750672 编辑部 010-62570390

印 刷 者 北京大学印刷厂

经 销 者 新华书店

787毫米×1092毫米 16开本 29印张 718千字

2019年11月第1版 2019年11月第1次印刷

印 数 1-4000册

定 价 89.00元

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究

举报电话：010-62752024 电子信箱：[fd@pup.pku.edu.cn](mailto:fd@pup.pku.edu.cn)

图书如有印装质量问题，请与出版部联系。电话：010-62756370

# 前言

Preface

## 为什么要写这本书？

我的一个学生来自津巴布韦，从他的家乡来中国需要乘坐18个小时的飞机，高昂的票价让人头疼。大家都知道，一般机票越是提前预订就越便宜，于是他提前一个月订了。然而，他后来发现这并不是最便宜的机票，因为有些机票在飞机起飞前几天会突然降价。

于是我的学生和他的团队决定建立一个系统，获取每架飞机起飞前一个月的票价信息，形成一个大型的数据库，利用Spark和一些算法来预测什么时候会出现最低票价，以此来帮助更多的乘客节省出行费用。

现在这个系统已经表现出了它的强大潜力，即便目前只有62%的预测准确率，也能为每次乘机节省20%左右的成本。

数据量越大，预测的准确率就会越高，整个社会的无效成本也会降低，这就是大数据的力量。

现在已经进入了大数据时代，在往数据智能时代大步迈进。在任何行业、任何场景中，都能看到大数据和人工智能的影子。比如在2017年，百度无人驾驶车上路；2018年，建设银行推出了无人银行，同年年底，支付宝的刷脸支付已经在北京全面落地。

这些蕴藏无限价值的高端技术，很多都已经开源免费，但是学习门槛之高，将大量的技术人员拒之门外。本书的目标就是降低这个门槛，让读者能用最低的成本快速进入大数据领域。

## 这本书的特点是什么？

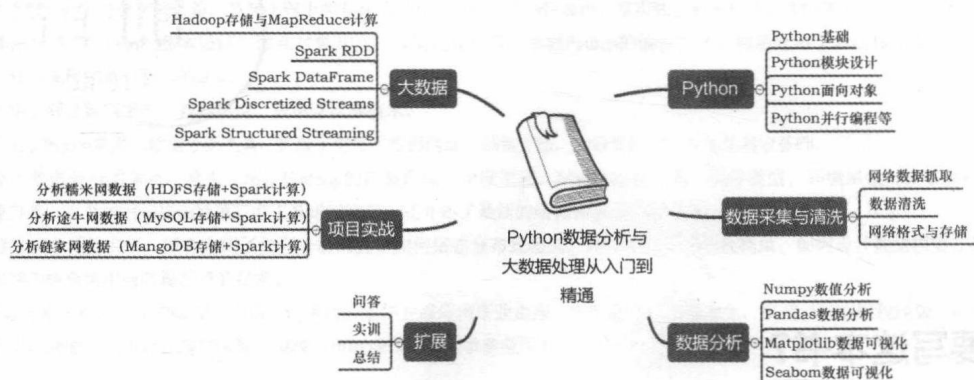
本书力求简单、实用，坚持以实例为主、理论为辅的路线。全书分五个篇章，从基础语言到大数据平台搭建和架构设计，以及最后的结论分析，覆盖了大数据项目开发阶段的整个生命周期。本书的特点如下。

(1) 没有的高深理论，每一章都是以实例为主，读者参考源码，修改实例，切换数据源，就能得到自己想要的结果。目的就是让读者看得懂、学得会、做得出。

(2) 因为专注，所以专业。Numpy、Pandas、Matplotlib、Hadoop、Spark这些组件功能非常丰富，然而本书专注于基于Python做大数据项目分析，重点描述在生产环境中实际用到的这部分技术。相比大而全的书，本书能让读者尽快上手，然后投身项目开发。

(3) 书中的问答与实训板块让读者在学完本章知识后能够尽快得到巩固，举一反三，学以致用。

## 这本书的主要内容有什么？



## 通过这本书能学到什么？

- (1) 大数据分析概念。了解大数据的特征、大数据项目开发流程、大数据分析目标如何确立。
- (2) 1种编程语言和14种工具。掌握Python编程语言，能搭建大数据分析平台、利用爬虫获取原始数据、通过编程实现大数据分析，再搭配14种工具，构建完整的大数据分析项目。
- (3) 数据采集。能够分析业务需求、明确采集目标、设计数据模型、搭建分布式爬虫、应对反爬虫、解析原始数据、制订存储方案。
- (4) 数据清洗。用正确的步骤和方法处理错误、默认数据，进行数据检查；对数据进行计算、转换、分类等加工处理。
- (5) 数据分析与可视化。了解简单的统计原理，能对数据进行常规探索及基本可视化。
- (6) 大数据分析。理解分布式存储原理、分布式计算原理及分布式资源调度原理，掌握HDFS存储数据技术及Spark大数据分析技术。
- (7) 掌握架构设计与实施。能设计不同场景下的项目架构，并做好不同业务下的数据建模。
- (8) 项目开发。熟练使用Python语言，综合运用各类组件，独立完成项目开发。

## 这本书的组件版本和阅读时的注意事项

### 1. 核心组件版本（出稿时最新版本）

Python: Anaconda3 Python 3.7版本

CentOS: 7.5

Hadoop: 3.1.1

Spark: 2.4.0

Hive: 3.1.1

MangoDB: 4.0.6

Redis: Windows版本

其中，Hadoop、Spark、Hive安装过程相对复杂，版本不匹配容易出错。建议读者使用与本书一致

的版本，待对大数据平台精通之后，再选择其他版本。

## 2. 注意事项

在问答题与实训板块，建议读者根据题目回顾小节内容，进行思考后动手写出答案，以强化学习效果。

## 除了书，您还能得到什么？

(1) 赠送：案例源码。提供书中相关案例的源代码，方便读者学习参考。

(2) 赠送：Python常见面试题精选（50道），旨在帮助读者在工作面试时提升过关率。习题见附录，具体答案参见下方的资源下载。

(3) 赠送：《微信高手技巧随身查》《QQ 高手技巧随身查》《手机办公 10 招就够》三本电子书，教会读者移动办公诀窍。

(4) 赠送：“5 分钟学会番茄工作法”视频教程。教会读者在职场中高效地工作，轻松应对职场那些事儿，真正让读者“不加班，只加薪”！

(5) 赠送：“10 招精通超级时间整理术”视频教程，教会读者如何整理时间、有效利用时间。无论是职场还是生活，都要学会整理时间，因为时间是人类最宝贵的财富，只有合理整理时间，充分利用时间，才能让人生价值最大化。

**温馨提示：**以上资源，请用微信扫一扫下方二维码关注公众号，输入代码pY0119Ht，获取下载地址及密码。



本书由凤凰高新教育策划，朱春旭老师编著。在本书的编写过程中，我们竭尽所能地为您呈现最好、最全的实用内容，但仍难免有疏漏和不妥之处，敬请广大读者不吝指正。

读者信箱：2751801073@qq.com

# 目录

## Contents

### 第 1 篇 Python 程序设计

#### 第 1 章 Python 入门 /3

- 1.1 Python 概述 /4
  - 1.1.1 Python 的发展历程 /4
  - 1.1.2 Python 生态的应用 /4
  - 1.1.3 Python 的前景 /5
- 1.2 搭建 Python 开发环境 /6
  - 1.2.1 独立安装 /6
  - 1.2.2 安装 Anaconda /9
- 1.3 Python 开发工具介绍 /11
- 1.4 Python 软件包的管理 /13
  - 1.4.1 搜索软件 /13
  - 1.4.2 安装软件 /13
  - 1.4.3 卸载软件 /14
  - 1.4.4 更新软件 /14
  - 1.4.5 显示已安装软件包 /14
- 1.5 实训：编写“Hello World” /15
- 本章小结 /16

#### 第 2 章 Python 基础 /17

- 2.1 变量 /18
  - 2.1.1 什么是变量 /18
  - 2.1.2 变量与类型 /18
  - 2.1.3 变量赋值 /19
  - 2.1.4 动态类型 /20

- 2.1.5 内存管理 /20
- 2.1.6 垃圾回收 /22
- 2.1.7 Python 代码执行过程 /23
- 2.2 标识符 /24
  - 2.2.1 有效的标识符 /24
  - 2.2.2 特殊标识符 /24
  - 2.2.3 关键字 /25
  - 2.2.4 内建模块 /26
- 2.3 代码组织 /26
  - 2.3.1 缩进 /26
  - 2.3.2 代码注释 /27
  - 2.3.3 多行语句 /27
- 2.4 输入与输出 /28
  - 2.4.1 输入 /28
  - 2.4.2 输出 /28
  - 2.4.3 一个完整的示例程序 /29
- 2.5 运算符与优先级 /30
- 2.6 新手问答 /30
- 2.7 实训：设计一个简易计算器 /31
- 本章小结 /31

#### 第 3 章 数据类型与流程控制 /32

- 3.1 数字类型 /33
  - 3.1.1 数字对象的创建、修改与删除 /33

- 3.1.2 整型 /34
  - 3.1.3 浮点型 /35
  - 3.1.4 复数型 /35
  - 3.1.5 运算符 /36
  - 3.2 字符串类型 /37
    - 3.2.1 字符串对象的创建、修改和删除 /37
    - 3.2.2 格式化 /38
    - 3.2.3 字符串模板 /39
    - 3.2.4 转义字符 /39
  - 3.3 集合类型 /40
    - 3.3.1 列表 /40
    - 3.3.2 元组 /42
    - 3.3.3 字典 /43
  - 3.4 流程控制语句 /45
    - 3.4.1 循环 /45
    - 3.4.2 条件 /46
  - 3.5 新手问答 /47
  - 3.6 实训：设计算法，输出乘法表 /49
  - 本章小结 /50
- ## 第4章 函数、模块、包 /51
- 4.1 自定义函数 /52
    - 4.1.1 创建函数 /52
    - 4.1.2 调用函数 /52
    - 4.1.3 函数解包 /53
    - 4.1.4 递归函数 /54
  - 4.2 函数参数 /55
    - 4.2.1 位置参数 /55
    - 4.2.2 可选参数 /56
    - 4.2.3 可变参数与关键字参数 /57
  - 4.3 函数式编程 /58
    - 4.3.1 高阶函数 /58
    - 4.3.2 装饰器 /61
    - 4.3.3 偏函数 /62
    - 4.3.4 变量作用域 /62
    - 4.3.5 闭包 /63
  - 4.4 模块与包 /63
    - 4.4.1 导入模块 /63
    - 4.4.2 包 /65
  - 4.5 新手问答 /65
  - 4.6 实训：设计算法，对列表进行排序 /67
  - 本章小结 /68
- ## 第5章 面向对象的程序设计 /69
- 5.1 面向对象 /70
    - 5.1.1 面向对象思想 /70
    - 5.1.2 类和对象 /70
  - 5.2 自定义类 /71
    - 5.2.1 创建类语法 /72
    - 5.2.2 创建可实例化类 /72
    - 5.2.3 创建抽象类 /72
  - 5.3 属性 /73
    - 5.3.1 类属性 /73
    - 5.3.2 实例属性 /74
    - 5.3.3 动态属性 /76
    - 5.3.4 特性 /76
  - 5.4 方法 /79
    - 5.4.1 实例方法 /79
    - 5.4.2 静态方法 /80
    - 5.4.3 类方法 /81
    - 5.4.4 抽象方法 /81
    - 5.4.5 动态方法 /82
    - 5.4.6 适用场景 /83
  - 5.5 类的继承 /83
    - 5.5.1 继承 /83
    - 5.5.2 多态 /85
  - 5.6 可调用对象 /86
    - 5.6.1 创建可调用对象 /86
    - 5.6.2 有状态的函数 /86
  - 5.7 不可变对象 /87
    - 5.7.1 可变对象 /87

- 5.7.2 不可变对象 /88
- 5.8 新手问答 /88
- 5.9 实训：设计算法，构造一棵二叉树 /90
- 本章小结 /92

## 第6章 高级主题 /93

- 6.1 生成器 /94
  - 6.1.1 创建生成器 /94
  - 6.1.2 yield 关键字 /95
  - 6.1.3 将值传到生成器 /95
- 6.2 迭代器 /96

- 6.3 异步处理 /97
  - 6.3.1 多线程 /97
  - 6.3.2 多进程 /99
  - 6.3.3 协程 /101
- 6.4 错误、调试 /103
  - 6.4.1 异常处理 /103
  - 6.4.2 调试源码 /107
- 6.5 新手问答 /108
- 6.6 实训：使用多进程技术统计数据并  
汇总 /109
- 本章小结 /110

## 第2篇 数据采集与数据清洗

### 第7章 网络数据采集 /113

- 7.1 HTTP 请求概述 /114
  - 7.1.1 HTTP 请求过程 /114
  - 7.1.2 HTTP 请求语义 /114
- 7.2 XPath 网页解析 /114
  - 7.2.1 网页解析工具 /115
  - 7.2.2 HTML 页面概述 /115
  - 7.2.3 XPath 语法 /116
- 7.3 Scrapy 数据采集入门 /119
  - 7.3.1 框架简介 /119
  - 7.3.2 框架安装 /120
  - 7.3.3 创建项目 /121
  - 7.3.4 创建爬虫 /122
  - 7.3.5 爬取网页 /122
  - 7.3.6 提取数据 /122
  - 7.3.7 数据存储 /124
  - 7.3.8 常用命令 /125
- 7.4 Scrapy 应对反爬虫程序 /126
  - 7.4.1 反爬虫简介 /127
  - 7.4.2 Scrapy 应对反爬虫 /127
- 7.5 CrawlSpider 类 /131

- 7.5.1 核心概念 /131
- 7.5.2 爬取网络数据 /132
- 7.6 分布式爬虫 /132
  - 7.6.1 分布式爬虫架构 /132
  - 7.6.2 使用 scrapy\_redis 构建分布式  
爬虫 /133
- 7.7 新手问答 /136
- 7.8 实训：构建百度云音乐爬虫 /136
- 本章小结 /139

### 第8章 数据清洗 /140

- 8.1 数据清洗的意义 /141
- 8.2 数据清洗的内容 /141
- 8.3 数据格式与存储类型 /142
  - 8.3.1 Excel 数据 /142
  - 8.3.2 XML 数据 /143
  - 8.3.3 JSON 数据 /144
  - 8.3.4 CSV 数据 /144
- 8.4 数据清洗的步骤 /145
  - 8.4.1 找出噪声数据 /145
  - 8.4.2 清洗数据 /145
  - 8.4.3 保存数据 /146

- 8.5 数据清洗的工具 /147
  - 8.5.1 使用 Excel 清洗数据 /147
  - 8.5.2 使用文本编辑器清洗数据 /149
  - 8.5.3 使用 Tabula 清洗数据 /149
- 8.6 新手问答 /151
- 8.7 实训：清洗百度云音乐数据并储存到 CSV /151
- 本章小结 /152

## 第 3 篇 数据分析与可视化

### 第 9 章 NumPy 数值计算 /155

- 9.1 NumPy 基础 /156
  - 9.1.1 数组属性 /156
  - 9.1.2 数据类型 /157
  - 9.1.3 创建数组 /158
  - 9.1.4 基本操作 /161
  - 9.1.5 索引、切片和迭代 /162
- 9.2 形状操作 /164
  - 9.2.1 更改形状 /164
  - 9.2.2 数组堆叠 /165
  - 9.2.3 矩阵拆分 /166
- 9.3 副本、浅拷贝和深拷贝 /166
  - 9.3.1 副本 /166
  - 9.3.2 浅拷贝 /167
  - 9.3.3 深拷贝 /167
- 9.4 高级索引 /168
  - 9.4.1 通过数组索引 /168
  - 9.4.2 通过布尔索引 /170
  - 9.4.3 通过 ix() 函数索引 /170
- 9.5 排序统计 /171
  - 9.5.1 排序 /171
  - 9.5.2 统计 /173
- 9.6 新手问答 /173
- 9.7 实训：销售额统计 /174
- 本章小结 /175

### 第 10 章 Matplotlib 可视化 /176

- 10.1 图形的基本要素 /177
- 10.2 绘图基础 /177
  - 10.2.1 入门示例 /178
  - 10.2.2 图形样式 /179
  - 10.2.3 使用 NumPy 数组 /180
  - 10.2.4 使用关键字参数绘图 /180
  - 10.2.5 分组绘图 /181
  - 10.2.6 线条属性 /182
  - 10.2.7 画布与子图 /183
  - 10.2.8 添加文本 /185
- 10.3 设置样式 /186
  - 10.3.1 样式表 /186
  - 10.3.2 临时引入样式 /187
  - 10.3.3 rc 参数 /188
- 10.4 图形样例 /189
  - 10.4.1 线图 /189
  - 10.4.2 直方图 /192
  - 10.4.3 条形图 /193
  - 10.4.4 饼状图 /195
  - 10.4.5 散点图 /196
  - 10.4.6 箱线图 /197
  - 10.4.7 极坐标图 /197
  - 10.4.8 折线图 /198
- 10.5 新手问答 /198
- 10.6 实训：营业数据可视化 /199
- 本章小结 /201

**第 11 章 Pandas 统计分析 /202**

- 11.1 Pandas 数据结构 /203
  - 11.1.1 数据结构 /203
  - 11.1.2 序列 /203
  - 11.1.3 数据帧 /206
  - 11.1.4 访问数据 /208
- 11.2 基础功能 /210
  - 11.2.1 描述性统计 /210
  - 11.2.2 索引重置 /212
  - 11.2.3 数据排序 /213
  - 11.2.4 数据遍历 /214
  - 11.2.5 自定义函数 /216
- 11.3 统计分析 /217
  - 11.3.1 统计基础 /217
  - 11.3.2 聚合统计 /220
  - 11.3.3 分组统计 /222
  - 11.3.4 连接合并 /225
- 11.4 时间数据 /229
  - 11.4.1 创建与转换 /229
  - 11.4.2 时间运算 /230
- 11.5 数据整理 /231
  - 11.5.1 数据清洗 /231
  - 11.5.2 稀疏数据 /233
- 11.6 高级功能 /234
  - 11.6.1 多维度分析 /234
  - 11.6.2 选取数据 /235

- 11.7 读写 MySQL 数据库 /236
- 11.8 新手问答 /237
- 11.9 实训：成绩分析 /237
- 本章小结 /239

**第 12 章 Seaborn 可视化 /240**

- 12.1 Seaborn 概述 /241
- 12.2 可视化数据关系 /242
  - 12.2.1 使用散点图观察数据分布 /242
  - 12.2.2 使用线图观察数据趋势 /245
  - 12.2.3 在同一个画布上显示更多关系 /246
- 12.3 根据数据分类绘图 /246
  - 12.3.1 分类散点图 /247
  - 12.3.2 在类别内部观察整体分布 /248
  - 12.3.3 在类别内部观察集中趋势 /249
- 12.4 单变量与双变量 /251
  - 12.4.1 绘制单变量分布 /251
  - 12.4.2 绘制双变量分布 /252
  - 12.4.3 数据集中的成对关系 /255
- 12.5 线性关系 /256
  - 12.5.1 线性回归函数 /256
  - 12.5.2 绘制多项式回归 /257
- 12.6 新手问答 /258
- 12.7 实训：成绩分析可视化 /258
- 本章小结 /260

**第 4 篇 大数据存储与快速分析篇****第 13 章 Hadoop 数据存储与基本操作 /263**

- 13.1 Hadoop 概述 /264
  - 13.1.1 什么是大数据 /264
  - 13.1.2 大数据的特征 /264
  - 13.1.3 处理大数据遇到的困难 /264
  - 13.1.4 Hadoop 入门 /265

- 13.1.5 Hadoop 简介 /265
- 13.1.6 Hadoop 的生态 /267
- 13.2 Hadoop 数据存储与任务调度原理 /268
  - 13.2.1 HDFS 的体系架构与文件读写流程 /268
  - 13.2.2 YARN 的结构与资源调度过程 /271

- 13.2.3 MapReduce 执行过程 /271
- 13.3 Hadoop 基础环境搭建 /273
  - 13.3.1 安装虚拟机 /273
  - 13.3.2 安装 Linux 系统和客户端工具 /276
  - 13.3.3 安装 Hadoop /289
  - 13.3.4 安装 SSH /291
  - 13.3.5 安装 JAVA /292
- 13.4 Hadoop 部署模式 /294
  - 13.4.1 Hadoop 单机部署 /294
  - 13.4.2 Hadoop 伪分布式部署 /295
- 13.5 Hadoop 常用操作命令 /298
  - 13.5.1 系统管理 /298
  - 13.5.2 文件管理 /299
- 13.6 新手问答 /300
- 13.7 实训：动手搭建 Hadoop 集群环境 /301
- 本章小结 /309

## 第 14 章 Spark 入门 /310

- 14.1 Spark 概述 /311
  - 14.1.1 Spark 简介 /311
  - 14.1.2 Spark 特点 /311
  - 14.1.3 Spark 生态 /312
- 14.2 Spark 核心原理 /312
  - 14.2.1 重要概念 /313
  - 14.2.2 架构设计 /313
  - 14.2.3 运行流程 /314
- 14.3 Spark 基础环境搭建 /315
  - 14.3.1 安装 Python 3 /315
  - 14.3.2 安装 Spark /316
- 14.4 Spark 运行模式 /317
  - 14.4.1 Local 模式 /317
  - 14.4.2 Standalone 模式 /318
  - 14.4.3 Mesos 模式 /320
  - 14.4.4 YARN 模式 /320
- 14.5 新手问答 /321
- 14.6 实训：动手搭建 Spark 集群 /322

本章小结 /323

## 第 15 章 Spark RDD 编程 /324

- 15.1 RDD 设计原理 /325
  - 15.1.1 RDD 常用操作 /325
  - 15.1.2 RDD 依赖关系 /326
  - 15.1.3 Stage 概述 /327
- 15.2 RDD 编程 /328
  - 15.2.1 准备工作 /328
  - 15.2.2 读取外部数据源创建 RDD /329
  - 15.2.3 使用数组创建 RDD /330
  - 15.2.4 转换操作 /330
  - 15.2.5 行动操作 /333
- 15.3 键值对 RDD /335
  - 15.3.1 读取外部文件创建 Pair RDD /336
  - 15.3.2 使用数组创建 Pair RDD /336
  - 15.3.3 常用的键值对转换操作 /337
- 15.4 文件读写 /340
  - 15.4.1 读取 HDFS 并保存到本地 /340
  - 15.4.2 读取 HDFS 并保存到 HDFS /341
  - 15.4.3 读取本地文件并保存到 HDFS /341
- 15.5 编程进阶 /342
  - 15.5.1 分区 /342
  - 15.5.2 持久化 /343
  - 15.5.3 共享变量 /345
- 15.6 新手问答 /347
- 15.7 实训：统计海鲜销售情况 /348
- 本章小结 /350

## 第 16 章 Spark SQL 编程 /351

- 16.1 Spark SQL 概述 /352
  - 16.1.1 Spark SQL 简介 /352
  - 16.1.2 Spark SQL 特点 /353
  - 16.1.3 Spark SQL CLI 工具 /353

16.2	创建 DataFrame 对象	/360
16.2.1	读取文本文件	/360
16.2.2	读取 MySQL	/361
16.2.3	读取 Hive	/362
16.2.4	将 RDD 转换为 DataFrame	/362
16.3	DataFrame 常用 API	/364
16.3.1	显示数据	/364
16.3.2	查询数据	/365
16.3.3	统计数据	/368
16.3.4	执行 SQL 语句	/369
16.4	保存 DataFrame	/370
16.4.1	保存到 json 文件	/370
16.4.2	保存到 MySQL	/371
16.4.3	保存到 Hive	/371
16.5	新手问答	/372
16.6	实训：统计手机销售情况	/373
	本章小结	/375

## 第 17 章 Spark 流式计算编程 /376

17.1	流计算简介	/377
17.1.1	流式处理背景	/377
17.1.2	常用流计算框架	/378
17.2	Discretized Stream	/379
17.2.1	快速入门	/379
17.2.2	数据源	/382
17.3	Structured Streaming	/385
17.3.1	快速入门	/385
17.3.2	编程模型	/386
17.3.3	流式 DataFrame 源	/389
17.3.4	集成 Kafka 和窗口聚合	/390
17.3.5	查询输出	/395
17.4	新手问答	/397
17.5	实训：实时统计贷款金额	/397
	本章小结	/398

## 第 5 篇 项目实战篇

### 第 18 章 分析电商网站销售数据 /401

18.1	目标分析	/402
18.1.1	分析主页面	/402
18.1.2	分析商家商品列表	/403
18.1.3	分析商品详情页面	/405
18.2	数据采集	/405
18.2.1	模型设计	/406
18.2.2	架构设计	/406
18.2.3	采集数据	/407
18.3	数据分析	/411
18.3.1	将数据上传到 HDFS	/411
18.3.2	筛选口碑最好的十户商家	/412
18.3.3	筛选人均消费最高的十户商家	/412
18.3.4	筛选卖得最好的十个商品	/413

18.3.5	筛选卖得最贵的十户商家	/414
--------	-------------	------

18.3.6	分析口碑和销量的关系	/415
--------	------------	------

本章小结 /416

### 第 19 章 分析旅游网站数据 /417

19.1	目标分析	/418
19.1.1	分析主页面	/418
19.1.2	分析游记详情页面	/420
19.2	数据采集	/420
19.2.1	模型设计	/421
19.2.2	架构设计	/421
19.2.3	采集数据	/421
19.3	数据分析	/425
19.3.1	分析“驴友”普遍去了哪些地方	/425
19.3.2	分析“驴友”出行特点	/426

19.3.3 推测“驴友”都喜欢在哪个  
季节出行 /427

19.3.4 推测未来的热门景点 /428

本章小结 /429

## 第 20 章 分析在售二手房数据 /430

20.1 目标分析 /431

20.1.1 分析主页面 /431

20.1.2 分析房源详情页面 /432

20.2 数据采集 /434

20.2.1 模型设计 /434

20.2.2 架构设计 /435

20.2.3 安装 MongoDB /435

20.2.4 采集数据 /437

20.3 数据分析 /440

20.3.1 给 Spark 配置 MongoDB  
驱动 /440

20.3.2 筛选靠近地铁的房源 /440

20.3.3 分析各区域在售房源占比 /441

20.3.4 分析在售房源的户型 /442

20.3.5 分析房龄和平米单价的  
关系 /443

20.3.6 分析在售房源小区的热度 /445

本章小结 /446

附录：Python 常见面试题精选 /447

主要参考文献 /450

# 第 1 篇

# Python程序设计

Python 是一种解释性、面向对象和跨平台的高级编程语言。经过多年的发展，其功能越来越丰富，运行越来越稳定，性能越来越好。同时，活跃的技术社区开发了各种各样的高级工具，如 django、flask、jieba、nltk 等，为 Python 的广泛应用提供了强大的支持。

本篇先介绍 Python 的发展历程和发展前景，建立起读者对 Python 的基本认知；然后介绍 Python 的基本语法、逻辑控制、模块化设计、面向对象设计，让读者掌握用 Python 编程的技能；最后在高级主题部分，介绍生成器、迭代器、多线程、多进程、协程，以此来优化 Python 代码和提高程序性能。



## 第1章

# Python入门



### 本章导读

本章主要介绍Python的历史背景与发展现状、应用场景、环境搭建、常用开发工具、软件包的安装与卸载、编码规范。学习Python需要先打好基础，对它有一个初步的了解，才能进一步往上攀登。



### 知识要点

通过对本章内容的学习，读者能掌握以下内容。

- ◆ Python的历史与发展状况
- ◆ Python的应用范围
- ◆ Python的开发环境搭建
- ◆ Python的软件包管理
- ◆ Python的基本编码结构