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面试习题精选

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
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

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?

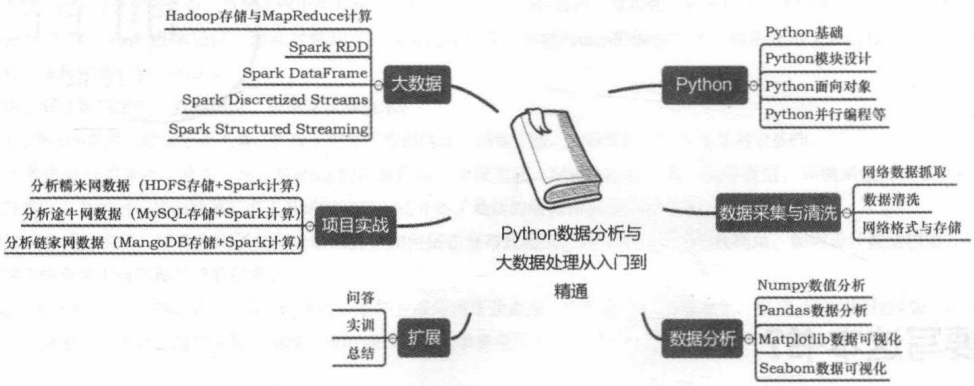
本书力求简单、实用，坚持以实例为主、理论为辅的路线。全书分五个篇章，从基础语言到大数据平台搭建和架构设计，以及最后的结论分析，覆盖了大数据项目开发阶段的整个生命周期。本书的特点如下。

（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任意二维码关注公众号，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 3 篇 数据分析与可视化

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