Broadview®
www.broadview.com.cn

任柳江◎著

全栈数据之门

天下没有免费的午餐，
没有一个框架与算法适用于所有的数据，
全栈由此而来。

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

全栈数据之门

任柳江◎著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书以数据分析领域最热的Python语言为主要线索，介绍了数据分析库numpy、Pandas与机器学习库scikit-learn，使用了可视化环境Orange 3来理解算法的一些细节。对于机器学习，既有常用算法kNN与Kmeans的应用，决策树与随机森林的实战，还涉及常用特征工程与深度学习中的自动编程器。在大数据Hadoop与Hive环境的基础之上，使用Spark的ML/MLlib库集成了前面的各部分内容，让分布式机器学习更容易。大量的工具与技能实战的介绍将各部分融合成一个全栈的数据科学内容。

本书不是从入门到精通地介绍某一种技术，可以把本书当成一本技术文集，内容定位于数据科学的全栈基础入门，全部内容来自当前业界最实用的技能，有非常基础的，也有比较深入的，有些甚至需要深入领悟才能理解。

本书适用于任何想在数据领域有所作为的人，包括学生、爱好者、在职人员与科研工作者。无论想从事数据分析、数据工程、数据挖掘或者机器学习，或许都能在书中找到一些之前没有接触过的内容。

图书在版编目（CIP）数据

全栈数据之门/任柳江著. —北京：电子工业出版社，2017.4
ISBN 978-7-121-30905-2

Ⅰ. ①全… Ⅱ. ①任… 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（2017）第022361号

策划编辑：张春雨
责任编辑：刘　舫
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：24.75　　字数：445千字
版　　次：2017年4月第1版
印　　次：2017年4月第1次印刷
定　　价：79.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

0x00 自序

慈悲为怀大数据，云中仙游戒为师。

这是自己从几年前一直沿用到现在的签名，几年之后的今天，再来体会这句话，不一样的处境，不一样的心境，却依然有着同样的追求。

曾想出世修行，渴望每日有高山流水相伴，能过着青灯古佛的生活。终因现实残酷只得入世而求存，在多少次碌碌无为中坚定了技术这条路。

技术之路，注定会一波三折。在下也经历了从安全测试、安全分析，到大数据分析，再到 APP 后端开发，直至数据分析、机器学习与深度学习之后，技术之栈才得以完全确立。技术之路漫长而曲折，需要不断修行，目前我也仅仅是入得门内，自此方有机会窥探神秘数据世界之一二而已。

少年不识愁滋味，为赋新词强说愁。而今识尽愁滋味，却道天凉好个秋。

学无止境。曾经以为学会 Linux 便够了，殊不知，这仅仅是系统的基础；后来学了 Python，以为这便是编程的全部；殊不知，Python 最强大的领域在数据科学；直到接触大数据与机器学习，才发现，原来种种际遇，都只是为数据科学而铺设的“套路”。

本书并非从入门到精通的讲解，只是想通过浅显易懂的语言让读者了解全栈数据的全貌。阅读本书时，如果其中某个知识点，让你入了门，我甚感欣慰；如果其中某节内容，让你得到了提高，我备受鼓舞。另外，入门之路千千万，用时下流行的话来说，只希望本书不会导致你“从入门到放弃”。

全栈数据，主要想尽可能多地涉及数据科学中的主题。任何复杂的技术，都

是一点点积累起来的，数据科学也不例外。如果能将本书中涉及的全栈数据技术，如 Linux、Python、SQL、Hadoop、Hive、Spark、数据挖掘、机器学习与深度学习进行系统性整合，则全栈数据之技可成也。

诗词歌赋，是诗人与词人对人生的情感寄托；技术写作，也是技术人员对技术的情感寄托。

然术业有专攻，每个人的知识都是有限的，写书的目的，并非要证明自己，而是把自己所知所想记录下来，让读者能有哪怕一小点的收获即可。

全栈并非全能，钱都不是万能的，何况技术乎？在数据领域，都懂一点，生活会更美好。

全栈是一种修行，数据技术如此，人生亦如是：

哲人的智慧，诗人的优雅，佛徒的慈悲；
开源的思想，安全的思路，数据的思维；
程序员的逻辑，测试员的严谨，分析员的远见。

阅读本书，不能让你立刻走上人生巅峰、出任 CEO……但至少可以达到以下几点：

- 使用 Linux 工具或者 MySQL 进行数据统计分析。
- 使用 Orange 进行机器学习实验。
- 使用 Python 或者 PySpark 进行项目实战。
- 使用 Hadoop 环境，如 HDP2 的集成环境，进行大数据研究。
- 使用 scikit-learn，并且可以阅读 Spark 的机器学习库文档。
- 熟练构建自己的数据科学技能。
- 从事数据领域相关的职位。

本书是一本无固定主题的技术文集合体，围绕“数据”这个主线，进行了大量的展开，从不同的侧面去靠近全栈数据技能，去靠近数据科学这个大主题。因内容宽泛，且作者水平有限，不足之处甚多，若读者发现书中的问题，还望不吝指正。可以通过我的微信公众号 yunjie-talk 反馈问题，我将不胜感激。

最后，本书得以成册出版，必须要感谢电子工业出版社计算机出版分社的张春雨老师，伯乐张老师于杂乱文字中，发现了闪光之处，促成了本书的问世。世人皆说本书体裁太乱，无章法可言，唯张伯乐以无招胜有招接下，众皆信服。

本书在写作过程中，得益于爱妻梁玉霞女士的大力支持，常于深夜端茶倒水，

询问进度，并且照顾家庭与小孩，让我可以抽出大量时间来书写，感激之情在心，在此道谢。与此同时，也感谢全力支持我写作的父母，他们帮忙照顾小孩与生活，对我学业、事业与写作的支持，让我感恩。

另外，本书在写作过程中，得到好友司旭鹏的很多支持与建议，在初稿审校过程中，得到好友尹高峰、卢西、彭玺锦的很多建议与修改，在此一并感谢。因为你们的付出，让本书质量得到了提升，非常感谢。

在写作本书的约一年时间之内，还得到了其他很多朋友、同事的大量建议，在此虽不一一提名，但必须要感谢你们的支持。

云戒

2016.11.11 于成都

前言　自强不息，厚德载物

本书共有 8 个章节的内容，涉及数据科学中的相关基础知识与内容，但内容的编排并非完全从易到难，有部分章节的内容，是需要用到其他章节的知识的。

相对来说，第 1、2、3 章，内容比较单一，涉及基础的 Linux、Python 与 Hadoop 知识。如果对这三章中的某些知识不熟悉，建议先阅读。第 4 章比较特殊，其内容也是数据科学中比较重要的，不仅需要前 3 章的知识，也需要部分 Spark 的知识，因为 Spark 的特殊性，单独放到机器学习之后了。

第 5、6 章，涉及数据科学中最重要的主题：机器学习与算法，介绍了机器学习的常用环境、概念、方法以及几个典型的算法应用。这两章是本书的难点，如果不熟悉，必须单独攻克。

第 7 章，Spark 本身就是一个全栈框架，无论是在分布式计算还是在机器学习领域，都大有用处。因此最好有前面章节的基础知识，方能更好地理解本章的内容，尤其是 MLlib/ML 库，必须有机器学习算法的知识。

最后一章，反而是最简单的，因为基本不涉及技术细节，但对整个数据科学的理解，以及技术积累都是非常重要的。

本书章节的编排，在一定程度上参考了知识的由易到难，另外一个方面，也参考了《易经》的乾坤两个卦象。不需要读者熟悉乾坤两个卦象，下面会将其中乾坤两个卦的爻辞进行粗略的解释。不需要读者完全理解，只求有个概念上的认识即可。

全栈数据，其中的数据既指数据技术，也指业务数据。只有将技术应用到业务中，才能在实际的生产环境中发挥作用。

介绍卦象的时候，是由低层次向高层次渐渐提升的，对应的技能与业务也一样。

从 Linux 走向数据科学，就是一个技能提升的过程，类似的，从数据采集到数据应用也是对业务从入门到应用的一个过程。

01 全栈技能，自强不息

乾坤为一体，况且一阴一阳之谓道。世界上不可能只有男人而没有女人，正如不可能只有女神而没有女汉子一样。因此，和纯阳卦乾卦相对的便是纯阴卦坤卦了。

乾卦，正是天行健，君子以自强不息。下面先用乾卦来说全栈技术。

1.1 Linux，潜龙勿用

初九爻，爻辞为：**潜龙勿用**。

乾卦每一爻都代表了一条龙，初九爻为潜龙。潜字很有意思，是叫你要藏起来，不要露面，那能干什么呢？

刚接触 Linux，觉得 Linux 非常自由。此时刚开始入门，兴趣最重要，这一阶段，需要打好各种系统与命令行的基础。

要做好数据科学，不学 Linux 不行，不论你喜欢与否。后续的 Python、Spark、Hadoop、数据挖掘等都需要用到大量基础的 Linux 知识。勿用的意思是，时候不到就不要用，当准备好了，就要用。

Linux 只是一个基础系统，必须要结合实际的业务来更好地应用，“勿用”到最后就是为了要用，此时就进入了第二阶段。

1.2 Python，见龙在田

九二爻，爻辞为：**见龙在田，利见大人**。

见（xiàn）通“现”，意为展现，要能在工作中快速解决问题，学习 Python，就可以让你快速展现出业绩来。由 Python 入门数据分析与数据挖掘，这算是程序员入行数据挖掘最好的方法了，Python 目前最火的领域，就是在数据科学领域。况且 Python 语法优美，第三方库庞大且高效，专门用于快速解决问题。

Python 是一条巨龙，一旦展现在田野，必须拿出一点实力来表现自己，对上面的大人有利。从 Linux 过来，将 Python 应用到业务中，快速地解决了问题，业

绩自然来了，其利也自现了。

在数据科学中，学了 Python 的开发与数据分析，这还只是基础，还有另外一个工程领域的技能——大数据，需要学习。此时，进入九三爻。

1.3　大数据，终日乾乾

所有的事件都会进入第三阶段，即九三爻，爻辞为：**君子终日乾乾，夕惕若，厉无咎**。

君子，指有志气、有抱负的人。整天都很努力地学习，是会被人嫉妒的，所以言行都要小心。

工作中，必须要兢兢业业，努力做好工作。太阳下山之后，更需要警惕，时刻告诫自己，三天不学习，可能就赶不上曾经比自己差的人了。必须要利用业余时间学习新技术，而以 Hadoop 为代表的大数据，正是近年来火得一塌糊涂的技术，况且网上的学习资料已经非常多了。

这是一门在学校很难学到的技术，因此必须自己利用业余时间，每天厉兵秣马，努力让自己的数据科学技能更加完善。

1.4　机器学习，或跃在渊

有些人和事，是没有第四阶段的，因为他们可能一辈子就留在第三阶段。进入第四阶段，即九四爻：**或跃在渊，无咎**。

学习了前面三种技术（Linux、Python 和大数据）后，要想再深入，此时必须往数据挖掘与机器学习方面发展。数据挖掘与数据分析还是有一定区别的，数据挖掘更偏向于使用算法解决问题，而分析更多是偏统计与业务（包括运营、产品）层面的。

尤其是机器学习，要看得懂 scikit-learn 的文档，必须辅助以相应的理论基础，涉及数学、统计学、计算机等多个领域。

一部分人在此阶段向机器学习方向跳跃，可是没有坚定的信念，最后就真应了那句“从入门到放弃”的话了。

机器学习，因其特殊的学科领域，由程序员转过来的占很大一部分，要想跃上去，必须坚定信念，不忘初心。对码农而言，其难，就难在理论。下足功夫，坚持苦修，一定会有所成就。

1.5 Spark，飞龙在天

过了九四阶段，到九五阶段也容易。而九五阶段也是最舒服的阶段，我们常说的九五至尊便是这个阶段。爻辞为：**飞龙在天，利见大人。**

有了前面的基础，再学习 Spark 技术，相对就很容易了。再面对 Spark 中的机器学习库 ML 或 MLlib，也就不会束手无策了。Spark 可以说是前面技能的集大成者，需要 HDFS 文件的支持，能天生支持 Hive 的数据，能使用 Python 的 API 接口，以前在 Python 中能用的那些库，在 Spark 中，通通都能用，而且效率更高。

技能到了，境界也差不了多少，浑身散发的数据气场也很强。此阶段也可以认为是第二阶段的见龙眼中的大人了，行事也需要有大人的风范才行。

1.6 数据科学，亢龙有悔

做技术的人，其实比较希望停留在九五阶段，可却经常事与愿违啊！而且物极必反的道理相信你也懂，事物的发展是不会停止的。过了九五阶段，还有九六阶段。九六爻爻辞为：**亢龙有悔。**

有的人感觉自己领悟了整个数据科学的技能与本质，其实也还只是一些皮毛，若此时停止不前，甚至转向管理，那么技术也就基本荒废了。

正确的做法应该是，回头再去加深领悟各个阶段的技术与技能，将其更好地应用到业务中去。回想曾经犯的错，曾经走过的弯路，反思悔悟，以期技术上能有一个更透彻的理解。

此时，你才掌握和领悟了数据科学。

1.7 群龙无首

最后，还有最重要的一点，乾卦最厉害之处在于还有个用九：**见群龙无首，吉。**就是你看见所有的龙，却不知哪一条是首领时，你才领悟到乾卦的真谛。

有些问题能用 Linux 脚本工具解决，那岂不更好吗？何必写出一段低效的代码来做呢？常识有时能比专业技能更好地解决问题，并且不要受限于工具与技术，SQL 能解决好的问题，用 Python 不一定高效。scikit-learn 能解决好的问题，Spark 未必就能做得同样好。

每个工具与技术都有自己擅长的领域，不同环境，不同需求，用不同的工具

解决。这才是群龙无首的思想，这才是全栈技术存在的意义。

说完乾卦与全栈技术，下面说坤卦与全栈业务。

02 全栈业务，厚德载物

学过物理的人都知道世界充满了电场和磁场。了解过佛学的人，都知道世界充满了念力场与信息场，通过信息场，可以与更高一级的文明进行沟通。

将技能应用到业务，可形成特殊的“数据”场。

2.1 数据采集，坚冰至

坤卦的初六爻，爻辞为：**履霜，坚冰至**。脚踩到霜了，坚冰就会来了。因此，要见微知著，对未来要有更多的准备与预料，虽然事情现在还处在萌芽期，但未来也许会更坏。

数据采集，就是要善于从细节处采集数据，于可能出问题的地方采集数据，不要放过一些看起来不太重要的数据。

在做数据分析的时候，也要非常重视一些小的、可疑的情况，因为往往在这些小问题中，可能会发现大的、有用的信息，从而挖掘出其背后的问题。

2.2 数据清洗，直方大

六二爻的爻辞：**直方大，不习无不利**。做人的基本原则，就是要正直，方正，大气。

对数据处理，同样如此。应用到数据清洗环节，更好理解。把一切不符合规范的数据进行整理，把一切可能的异常数据拿出来分析，或者修正它，或者舍弃它。

另外，不要轻易相信看到的表象数据，数据本身不会欺骗人，可是有些表象数据会迷惑人。对待数据要诚实，不要将人的各种坏习性带到数据中去。

在做分析、挖掘的时候，一定要保持清晰的思路，不要绕到数据里面去。保持一颗初心不变，方可顺利走出数据的泥沼。

2.3 数据处理，无成有终

六三爻，爻辞为：**含章可贞，或从王事，无成有终**。

含，隐藏，低调；章：才华，能力。含蓄地处事，保持美好的德行。或，或许，疑惑。低调且德行好的人，随时都要带着给王室或者皇帝办事的心态去做，最后，就算是没有成功或者没有成就，至少也得把事情办完。这就是贞，也就是忠心，正所谓：地势坤，君子以厚德载物。使用大数据技术，就是要像给帝王做事一样，只要你有能力写出好的代码，Hadoop 或者 Spark 这个工头，会找一堆小伙伴勤勤恳恳地完成任务。

进行数据分析时也需要有坚定的心志，遇到问题后，要坚持不懈地去寻找思路，即使最终没有彻底解决问题，也要把事情做完，这是一种品德，也是一种职业操守。

2.4 深入挖掘，括囊

六四爻，其爻辞为：**括囊，无咎无誉**。括，收紧；囊，口袋。把口袋的口收紧，踏实努力地将工作做好。

口袋的口虽然很小，但可以将肚子发展得非常大，你看 Google 的入口那么小，可是其内部之大，远远超过一般人的想象。要更好地利用数据，还必须对数据进行深入挖掘。

机器学习，是人类集体智慧的结晶，用一定的数学方法，即可以让机器自动学习并做预测。机器学习就是对前面收集与整理的数据，进行深入挖掘，让机器自动发现一些规律，以便为业务所用。

整个乾坤两卦，追求的目标都是无咎，而不是大吉大利。做好了括囊，也仅仅是无咎无誉。

要让内部发展得更大，自然可以将大数据与机器学习结合起来，这也是 Spark 的使命。将数据技术与业务很好地结合起来，在内部构建了一个强大的数据平台，还要能很好地支持业务系统。

2.5 产出数据，黄裳元吉

六五爻：**黄裳，元吉**。黄，黄色；裳，衣裳。看过金庸小说的人，一定知道黄裳是《九阴真经》的作者吧，他读遍了道家的书籍，终于创出江湖人人争夺的《九阴真经》。

黄色表示尊贵，古代只有皇帝的龙袍才能使用黄色。穿上黄色的衣裳，说明

你的努力有了回报。细心一点，你也许会发现，Hadoop 的图标是一头黄色的大象，而 Spark 的图标是黄色的星星。这两门技术，都是数据工程中重要的框架，只有将数据科学的理论更好地与工程方法结合，并将其与业务结合起来，才可能产出好的数据。

有了好的数据产出，更要将数据可视化出来，给数据穿上“黄色”的衣服，让人们重视，让大家理解。

2.6　数据应用，龙战于野

上六爻：**龙战于野，其血玄黄**。

一切数据，都需要与产品结合起来，否则产出的数据再好，对产品却不一定有效果。数据与产品有冲突，以产品为准。

物极必反，过分追求技术，过分追求数据，不顾及产品，或许就会从吉变为凶了。数据与产品战，其结果必然很血腥。

但换种好的说法，利用强大的数据平台，产出优质的数据，与竞争对手进行对战，必将可以成就你。

2.7　利永贞

坤卦的用六：**利永贞**。长久保持应有的贞操，构建好强大的数据平台，将全栈的数据技术与全栈的业务场景进行结合。这些事情做好了，对业务与产品都是长久的利益，公司和数据科学人员都要重视。

投入是有回报的，学习全栈技术有极大的回报，投入建设强大的数据平台也会有相应的回报。

在数据科学的工作过程中，将全栈技术应用到全栈业务中，乾坤一体，才是目的。

03 本书约定

通过前面的介绍，相信你对本书的章节编排有了一个全面的了解，即可开始阅读了。在阅读过程中，因为一些技术细节，还需要注意相应的环境与版本信息。

3.1 版本信息

```
操作系统：Ubuntu 14.04
Python：Anaconda 3(包含 Python 3.5)
Hadoop 发行版本：HDP 2.4.0
Spark 版本：1.6.2
scikit-learn 版本：0.17.1
```

3.2 Python 说明

因为 Python 2 与 Python 3 的特殊性，本书中所有的示例代码均使用 Python 3.5 来实现。如果需要用于 Python 2.7 的版本，请注意两个细节：

1. 在代码前面添加：
`from __future__ import print_function`，来使用 python 3 的 print 函数风格。
2. 需要自己处理中文编码的问题。

3.3 代码提示符

```
MySQL 提示符：
mysql>

Hive 提示符：
hive>

Bash 提示符：
$

Bash root 权限提示符：
$ sudo

Python 提示符：
>>>

Scala 提示符：
scala>

HBase Shell 提示符：
hbase>
```

说明：因为并没有必须使用 root 账号的时候，如果需要 root 权限的操作，通常都是在命令前面加 sudo。因此，配置一个无密码的 sudo 是很有必要的。

3.4　代码的注释

```
SQL 的注释符：  --  （后面必须带一个空格）
Shell 与 Python 的注释符号：#
Scala 的注释：//
```

3.5　交互式环境

本书主要使用了两个交互式环境，一个是 Python 代码使用的 Jupyter 的 Notebook，请参考“Anaconda，IPython”这节的说明。如果 Python 代码中最后一行只有一个变量，没有加 print 函数，是直接使用了 Jupyter 的优化显示效果，Jupyter 会默认对最后一个变量进行优化显示。

另外一个是 Spark-SQL 和 PySpark 使用的交互式 Zeppelin，请参考“Zeppelin，一统江湖”一节。

目　录