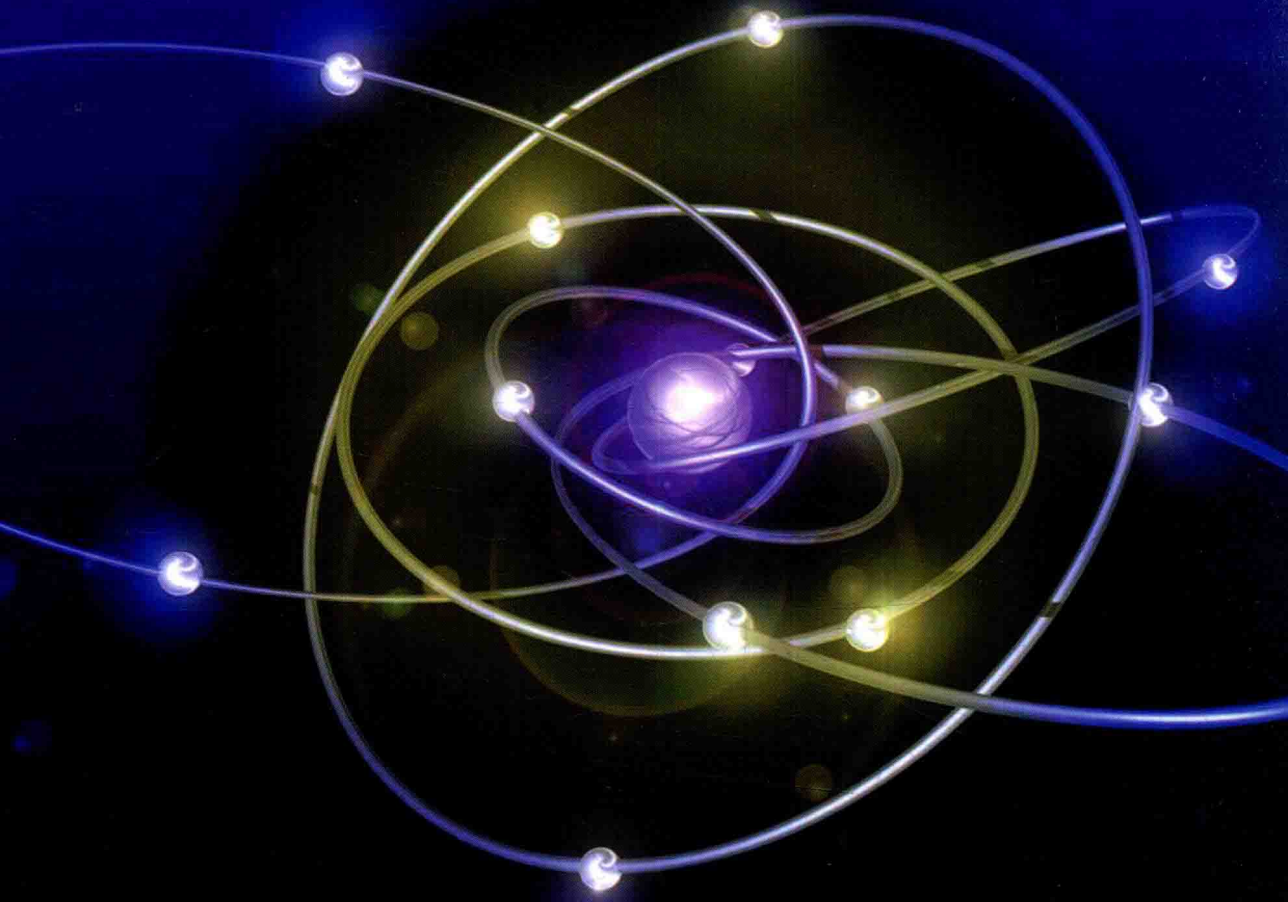
范例文件下载



Hadoop+Spark

大数据巨量分析与机器学习

整合开发实战

林大贵 著

- 大数据技术为金融财务、营销分析、商业趋势预测带来全新变革
- 详实的安装设置与程序解析，降低学习门槛
- 可单机运行/实机/虚拟机建立多台运算集群
- 提供大量实际案例详解与程序代码范例

清華大學出版社

Hadoop+Spark

大数据巨量分析与机器学习

整合开发实战

林大贵 著

清華大學出版社
北京

内 容 简 介

本书从浅显易懂的“大数据和机器学习”原理介绍和说明入手，讲述大数据和机器学习的基本概念，如：分类、分析、训练、建模、预测、机器学习（推荐引擎）、机器学习（二元分类）、机器学习（多元分类）、机器学习（回归分析）和数据可视化应用。为降低读者学习大数据技术的门槛，书中提供了丰富的上机实践操作和范例程序详解，展示了如何在单台 Windows 系统上通过 Virtual Box 虚拟机安装多台 Linux 虚拟机，如何建立 Hadoop 集群，再建立 Spark 开发环境。书中介绍搭建的上机实践平台并不限制于单台实体计算机。对于有条件的公司和学校，参照书中介绍的搭建过程，同样可以将实践平台搭建在多台实体计算机上，以便更加接近于大数据和机器学习真实的运行环境。

本书非常适合于学习大数据基础知识的初学者阅读，更适合正在学习大数据理论和技术的人员作为上机实践用的教材。

图书在版编目（CIP）数据

Hadoop + Spark 大数据巨量分析与机器学习整合开发实战 / 林大贵著. — 北京：清华大学出版社，2017
ISBN 978-7-302-45375-8

I. ①H… II. ①林… III. ①数据处理软件 IV.①TP274

中国版本图书馆 CIP 数据核字（2016）第 260890 号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：沈　露

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
印 装 者：北京鑫丰华彩印有限公司
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张**：27.75　　**字　　数**：730 千字
版　　次：2017 年 1 月第 1 版　　**印　　次**：2017 年 1 月第1 次印刷
印　　数：1～3000
定　　价：79.00 元

产品编号：069535-01

序

大数据的影响力正深入到各个领域和行业中。特别在商业、经济以及其他领域，将大量数据进行分析后，便可得到许多数据的关联性。这些关联性可用于预测商业趋势、营销研究、金融财务、疾病研究、打击犯罪等。大数据对每一个企业的决策方式将发生变革——决策方式将基于数据和分析的结果，而不是依靠经验和直觉。

信息科技（Information Technology，IT）浪潮的第一波是大型计算机，第二波是个人计算机（PC 机），第三波是网络，第四波是社交媒体，第五波则是“大数据”。每一波的信息科技浪潮都会带来工作与生活方式的改变，创造大量商机、新的产业、大量的工作机会。例如，在网络时代，创造了淘宝、百度、Google（谷歌）、Amazon（亚马逊）等大公司，以及无数.com 公司。

每一波浪潮开始时，相关人才的需求激增，从而造成相关人才的紧缺。因此对个人而言，如果能在浪潮兴起时就投入，往往成果很丰硕，并且有机会占有重要职位。例如，网络刚兴起时，每个公司都需要建立网站，但是这方面的人才当时相对不够，能掌握编写网页相关应用程序设计语言的工程师就能够获得高薪。之后，投入的人越来越多，这方面的工程师就没有当初那么吃香了。

之前的科技浪潮，也许你没有机会躬逢其盛，或是没有机会在浪潮初期进入。而目前大数据的浪潮方兴未艾，正是进入的好时机。根据 IBM 公司调查预估，大数据在 2014 年的市场规模为 71 亿美元，2015 年则达到了 180 亿美元，并将以每年增长 20%的速度持续成长。机会是给有准备的人的，学会了大数据分析的相关技能，让你有机会获得更好的薪资与职业发展前景。根据美国调查机构 Robert Half Technology 2016 年趋势报告，在美国，大数据工程师的薪水年增长 8.9%，年薪大约 13 万至 18 万美金（约合人民币 85 万元~120 万元）。因为人才短缺，企业不惜重金挖角。（搜索 Robert Half Technology 2016 就可以下载此调查报告。）

本书的主题是 Hadoop+Spark 大数据分析与机器学习。众所周知，Hadoop 是运用最多的大数据平台，然而 Spark 异军突起，与 Hadoop 兼容而且运行速度更快，各大公司也开始加入 Spark 的开发。例如，IBM 公司加入 Apache Spark 社区，打算培育百万名数据科学家。谷歌（Google）公司与微软公司也分别应用了 Spark 的功能来构建服务、发展大数据分析云与机器学习平台。这些大公司的加入，也意味着未来更多公司会采用 Hadoop+Spark

进行大数据的数据分析。

然而，目前市面上虽然很多大数据的书，但是多半偏向理论或应用层面的介绍，网络上的信息虽然很多，但是也很杂乱。本书希望能够用浅显易懂的原理介绍和说明，再加上上机实践操作、范例程序，来降低大数据技术的学习门槛，带领读者进入大数据与机器学习的领域。当然整个大数据的生态系非常庞大，需要学习的东西太多。希望读者通过本书的学习，有了基本的概念后，能比较容易踏入这个领域，以便继续深入与研究其他大数据的相关技术。

林大贵

推荐序

如同本书作者所说的，信息技术已经来到了第五波浪潮——“大数据”，在因特网、社交媒体、电子商务等交叉发展和呼应下，“网络”这个巨人已经拥有了难以计数的海量数据，有传统结构化的数据、半结构化的数据，但更多的是非结构化的数据。这些貌似杂乱无章、毫无意义的海量数据，却是一座等待发掘的巨大“金矿”。

这些海量数据中蕴含着极为丰富的人类知识库，它是一笔巨大的信息资产。这些原本很难收集整理的大数据，随着云计算时代的来临，对它们进行及时甚至是实时分析和处理并加以有效利用，就不再是“海市蜃楼”了。

与大数据相关的内容中，不外乎三个方面：大数据理论，大数据分析和处理的技术，大数据的实践应用。目前与大数据有关的出版物中，偏重于理论教学和技术介绍一类的比较多，而偏重于上机实践和自我学习的书却比较少见。因此，本书非常适合大数据学习的初学者和正在学习大数据理论和技术的人员作为上机实践用的教材。

本书从浅显易懂的“大数据和机器学习”原理介绍和说明开始，介绍大数据和机器学习——分类、分析、训练、建模、预测——机器学习（推荐引擎）、机器学习（二元分类）、机器学习（多元分类）、机器学习（回归分析）和数据可视化应用。

在本书中，不是对这些原理进行纯理论的阐述，而是提供了丰富的上机实践操作和范例程序，这样极大地降低了读者学习大数据技术的门槛，对于需要直接上机实践的学习者而言，本书更像是一本大数据学习的实践上机手册。书中首先展示了如何在单台 Windows 系统上通过 Virtual Box 虚拟机安装多台 Linux 虚拟机，而后建立 Hadoop 集群，再建立 Spark 开发环境。搭建这个上机实践的平台并不限制于单台实体计算机，主要是考虑个人读者上机实践的实际条件和环境。对于有条件的公司和学校，参照这个搭建过程，同样可以将实践平台搭建在多台实体计算机上。

在搭建好大数据上机实践的软硬件环境之后，就可以在各个章节的学习中结合本书提供的范例程序逐一设置、修改、调试和运行，从中学到大数据实践应用中核心技术的真谛——对大数据进行高效的“加工”，萃取大数据中蕴含的“智能和知识”，实现数据的“增值”，并最终将其应用于实际工作或者商业中。

大数据与云计算的关系密不可分，涉及众多关键技术，如分布式处理、分布式数据库和云存储、虚拟化技术等，本书并未在这些方面深入讲解，因为它们不是本书的重点，建议需要深入学习这方面内容的读者去寻找相关出版物，结合本书的实践来丰富和完善自己的大数据知识体系。

资深架构师 赵军

2016年7月

本书章节与范例程序介绍

本书特色

本书的特色是提供了大量上机实践操作与范例程序。

- **上机实践操作**

一般人可能会认为大数据需要很多台机器的环境才能学习，但是通过本书介绍使用 Virtual Box 虚拟机的方法，就能在自家的计算机上演练建立 Hadoop 集群，并且建立 Spark 开发环境。同时，上机实践操作介绍了 Hadoop MapReduce 与 HDFS 的基本概念，以及 Spark RDD 与 MapReduce 的基本概念。

- **范例程序**

以实际范例程序来学习程序设计是最有效率的学习方式。因此本书使用实际的数据集，配合范例程序代码来介绍各种机器学习的算法，并示范如何获取数据、训练数据、建立模型、预测结果，由浅入深地介绍 Spark 机器学习。

本书章节内容及上机实践操作与范例程序介绍

- **基本概念**

章节名称	说明
第 1 章大数据与机器学习	介绍大数据、Hadoop、HDFS、MapReduce、Spark、机器学习

- **Hadoop 的安装**

章节名称	说明
第 2 章 Virtual Box 虚拟机软件的安装	上机实践操作 安装 Virtual Box 虚拟机，让你可以在 Windows 系统上安装多台 Linux 虚拟机
第 3 章 Ubuntu Linux 操作系统的安装	上机实践操作 安装 Ubuntu Linux 操作系统

（续表）

章节名称	说明
第 4 章 Hadoop Single Node Cluster 的安装	上机实践操作 安装单台机器的 Hadoop Single Node Cluster
第 5 章 Hadoop Multi Node Cluster 的安装	上机实践操作 安装多台机器的 Hadoop Multi Node Cluster

- **Hadoop 的基本功能**

章节名称	说明
第 6 章 Hadoop HDFS 命令	上机实践操作 示范如何使用 HDFS 命令
第 7 章 Hadoop MapReduce	介绍 Hadoop MapReduce 的原理 WordCount.java 范例程序 示范使用 Hadoop MapReduce 计算文章内的每一个单词出现的次数

- **Spark 的基本功能**

章节名称	说明
第 8 章 Spark 的安装与介绍	上机实践操作 Spark 安装与 spark-shell 交互界面在不同环境中的运行示范
第 9 章 Spark RDD	上机实践操作 介绍 Spark 最基本的功能 RDD（Resilient Distributed Dataset，弹性分布式数据集）的基本运算
第 10 章 Spark 的集成开发环境	上机实践操作 安装集成开发环境（IDE） WordCount.scala 范例程序 示范使用 Spark MapReduce 计算文章内的每一个单词出现的次数

- **机器学习（推荐引擎）**

章节名称	说明
第 11 章创建推荐引擎	介绍如何使用 Spark MLlib 以 MovieLens 数据集建立电影的推荐引擎（Recommendation Engine） Recommend.scala 范例程序 示范如何获取数据、训练模型、推荐用户或电影，建立电影的推荐系统 AlsEvalution.scala 范例程序 示范如何调试推荐引擎参数，找出最佳的参数组合

- **机器学习（二元分类）**

章节名称	说明
第 12 章 StumbleUpon 数据集	StumbleUpon 数据集属于二元分类问题，可以根据网页的特征预测哪些网页是暂时性的或是可以长久存在的
第 13 章决策树二元分类	RunDecisionTreeBinary.scala 范例程序 示范如何使用决策树二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性的或可以长久存在的，并且找出最佳的参数组合，提高预测准确度
第 14 章逻辑回归二元分类	RunLogisticRegressionWithSGDBinary.scala 范例程序 示范如何使用决策树二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性的或是可以长久存在的，并且找出最佳的参数组合，提高预测准确度
第 15 章支持向量机 SVM 二元分类	RunSVMWithSGDBinary.scala 范例程序 示范如何使用支持向量机 SVM 二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性的或是可以长久存在的，并且找出最佳的参数组合，提高预测准确度
第 16 章朴素贝叶斯二元分类	RunNaiveBayesBinary.scala 范例程序 示范如何使用朴素贝叶斯（Naïve-Bayes）二元分类分析 StumbleUpon 数据集，预测哪些网页是暂时性的或是可以长久存在的，并且找出最佳的参数组合，提高预测准确度

- **机器学习（多元分类）**

章节名称	说明
第 17 章决策树多元分类	RunDecisionTreeMulti.scala 范例程序 示范如何使用决策树多元分类分析 Covtype 数据集（森林覆盖植被），根据不同的土地条件可以预测该地的植被，并且找出最佳的参数组合，提高预测准确度

- 机器学习（回归分析）

章节名称	说明
第 18 章决策树回归分析	RunDecisionTreeRegression.scala 范例程序 示范介绍决策树回归分析，分析 Bike Sharing 数据集。根据天气和假日条件，可以预测每一小时租借的数量，并且找出最佳的参数组合，提高预测准确度

- **数据可视化**

章节名称	说明
第 19 章使用 Apache Zeppelin 数据可视化	上机实践操作 安装 Zeppelin 并使用 ml-100k 数据集，示范使用 Spark SQL 进行数据分析与数据可视化

本书上机实践操作命令的整理

本书从第 2 章到第 10 章，我们使用了很多 Linux、spark-shell、SparkSQL 等命令。不过很多命令都很长，只要有一个字母打错就无法运行，这样会增加学习的挫折感。因此我们在博客文章中整理了各个章节使用的命令。可参考网页：http://www.weibo.com/hadoopsparkbook。

安装或练习命令时，你可以复制博客文章中的命令，然后粘贴到“终端”程序中，这样既可以节省打字的时间，也不用担心打错字母（如果无法在 VirtualBox 虚拟机的 Ubuntu“终端”程序中执行复制/粘贴操作，可参考第 3.9 节的说明，设置好 VirtualBox 的共享剪贴板）。

本书范例程序的下载与安装

以实际范例程序来学习程序设计是最有效率学习的方式。因此本书使用实际的数据集，配合范例程序来介绍各种算法，并示范如何获取数据、训练数据、建立模型、预测结果，由浅入深地介绍 Hadoop 与 Spark 机器学习。

你可以到下面的网址下载本书的范例程序：

http://pan.baidu.com/s/1qYMtjNQ（注意数字及字母大小写，如果下载有问题，请电子邮件联系 booksaga@126.com，邮件主题为“Hadoop + Spark 大数据巨量分析程序”）

但是，在这些范例程序安装和使用之前，必须建立整个开发环境。建议按照本书的顺序阅读第 2 到 10 章，并且真正建立整个开发环境。本书范例程序的安装说明将在第 10.16 节中介绍。

读者服务与社区交流

在网络时代，我们希望购买本书的读者不只是获得本书的内容，还能通过网络社区获得更多的信息。

- **本书的博客**

网址：http://blog.sina.com.cn/hadoopsparkbook。

我们还建立了本书的博客，将一些需要排列整齐，系统化的信息放在博客文章中。如果有新的博客文章，内容包括：

（1）本书上机实践操作命令的整理。
（2）本书内容或程序代码的勘误。
（3）分享最新的 Hadoop 或 Spark 信息。

- **本书的微博**

网址：http://www.weibo.com/hadoopsparkbook。

我们建立了本书的 Facebook 粉丝团，欢迎读者们加入。粉丝团会不定期贴文，分享最新的 Hadoop 或 Spark 信息，你也可以提问并参与交流。

目 录