异步图书
www.epubit.com.cn

“十三五”
国家重点图书出版规划项目

Official Study Kit for
Wrox Certified Big Data Analyst Program

大数据分析师权威教程

大数据分析与预测建模

Wrox 国际 IT 认证项目组／编　姚军／译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Official Study Kit for
Wrox Certified Big Data Analyst Program

大数据分析师权威教程

大数据分析与预测建模

Wrox 国际 IT 认证项目组／编　姚军／译

人 民 邮 电 出 版 社
北　京

图书在版编目（CIP）数据

大数据分析师权威教程. 大数据分析与预测建模 / Wrox国际IT认证项目组编 ; 姚军译. -- 北京 : 人民邮电出版社, 2017.11
书名原文: Official Study Kit for Wrox Certified Big Data Analyst Program
ISBN 978-7-115-46366-1

Ⅰ. ①大… Ⅱ. ①W… ②姚… Ⅲ. ①数据处理－教材
Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第202599号

内 容 提 要

“大数据”已连年入选IT领域的热点话题，人们每天都会通过互联网、移动设备等生产大量数据。如何从海量数据中洞悉出隐藏其后的见解是当今社会各领域人士极为关注的话题。本系列图书以“大数据分析师”应掌握的IT技术为主线，共分两卷，以7个模块（第1卷包括4个模块，第2卷包括3个模块）分别介绍大数据入门，分析和R编程入门，使用R进行数据分析，用R进行高级分析，机器学习的概念，社交媒体、移动分析和可视化，大数据分析的行业应用等核心内容，全面且详尽地涵盖了大数据分析的各个领域。

本书为第1卷，首先提供大数据的概览，介绍大数据概念及其在商业中的应用、处理大数据的技术、Hadoop生态系统和MapReduce的相关内容，然后介绍如何理解分析、分析方法与工具，重点讲解流行分析工具R，介绍如何将数据集导入R和从R导出数据、在R中如何操纵和处理数据，最后详细介绍R中的函数和包、R的描述性统计、R中的图形分析、R中的假设检验、R中的线性回归、非线性回归、聚类分析、决策树、R和Hadoop的集成及Hive，通过这些实战内容，使读者掌握R语言在数据分析中的全面应用。通过本书，读者能对大数据概念、重要性及其应用有全面的了解，熟悉各种大数据分析工具。

本书适用于想成为大数据分析师的人员以及所有对大数据分析感兴趣的技术人员和决策者阅读。

◆ 编　　Wrox国际IT认证项目组
译　　姚　军
责任编辑　杨海玲
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：32
字数：715千字　　2017年11月第1版
印数：1－2 400册　　2017年11月北京第1次印刷

著作权合同登记号　图字：01-2015-2395号

定价：108.00元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

版权声明

前 言

欢迎阅读《大数据分析师权威教程》和《大数据开发者权威教程》!

信息技术蓬勃发展，每天都有新产品问世，同时不断地形成新的趋势。这种不断的变化使得信息技术和软件专业人员、开发人员、科学家以及投资者都不敢怠慢，并引发了新的职业机会和有趣的工作。然而，竞争是激烈的，与最新的技术和趋势保持同步是永恒的要求。对于专业人士来说，在全球 IT 行业中，入行、生存和成长都变得日益复杂。

想在 IT 这样一个充满活力的行业中高效地学习，就必须做到：

- 对核心技术概念和设计通则有很好的理解；
- 具备适应各种平台和应用的敏捷性；
- 对当前和即将到来的行业趋势和标准有充分的认识。

鉴于以上几点，我们很高兴地为大家介绍《大数据分析师权威教程》（两卷）和《大数据开发者权威教程》（两卷）系列。

这两个系列共 4 本书，旨在培育新一代年轻 IT 专业人士——他们能够灵活地在多个平台之间切换，并能胜任核心职位。这两个系列是在对技术、IT 市场需求以及当今就业培训方面的全球行业标准进行了广泛并严格的调研之后才开发出来的。这些计划的构思目标是成为理想的就业能力培训项目，为那些有志于在国际 IT 行业取得事业成功的人提供服务。这一系列目前已经包含了一些最为热门的 IT 领域中的认证项目，如大数据、云、移动和网络应用程序、网络安全、数据库和网络、计算机操作、软件测试等。根据我们的全球质量标准加以调整之后，这些项目还能帮助你识别和评估职业机会，并为符合全球最著名企业的招聘流程做好最佳的准备。

这两个系列是学习和培训资源的知识库，为在重要领域和信息技术行业中培养厂商中立和平台独立的专业能力而设立。这些资源有效地利用了创新的学习手段和以成果为导向的学习工具，培养富有抱负的 IT 专业人士。同时也为开设大数据分析师和大数据开发者相关培训课程的讲师提供了全面综合的教学和指导方案。

《大数据分析师权威教程》系列概览

大数据可能是今天的科技行业中最受欢迎的流行语之一。全世界的企业都已经意识到了可用的大量数据的价值，并尽最大努力来管理和分析数据、发挥其作用，以建立战略和发展竞争优势。与此同时，这项技术的出现，导致了各种**新的**和**增强的工作角色**的演变。

《大数据分析师权威教程》系列的目标是培养新一代的国际化全能大数据分析师，使他们精通数据挖掘、数据操纵和数据分析方面的基本及高级分析技术，熟悉大数据平台以及业务和行业

需求，能够高效地参与大数据分析项目。

本系列旨在：

- 使参与者熟悉整个数据分析的生命期；
- 通过众多案例分析，使参与者熟悉大数据在不同相关行业中的角色和用途；
- 提供基本及高级大数据分析以及可视化的完整技术诀窍，帮助他们分析数据、创建统计模型和提供业务洞察力；
- 最后包含一个完整的项目，使参与者能够实施分析生命周期。

学习者的必备条件

要阅读这个系列图书，读者必须具备以下基础知识：

- 统计学基础知识，包括主要趋势和平均值计量、分散度计量、概率；
- 基本图表、直方图和散点图的创建；
- 基本熟悉数据库、表和字段，包括电子表格与计算。

建议的学习时间

《大数据分析师权威教程》由 7 **个学习模块**组成（第 1 卷包括 4 个模块，第 2 卷包括 3 个模块）。

根据参与者的技能水平，可以选择任何数量的模块以积累特定领域的技能，每个模块的学习目标会在后面列出。

对于**入门级的参与者**来说，建议选择全部 7 个模块，以便为成为合格大数据分析师做好充足的准备。**专业人士**或者已经拥有某些必备技能的参与者则可以选择能够帮助自己加强特定领域技能的模块。

每个模块可能占用大约 10 小时的学习时间，因此完整的学习时间大约是 70 小时。

模块清单

《大数据分析师权威教程》第 1 卷的 4 个模块的具体名称和学习目标如表 1 所示。

表 1

模块编号	模块名称	模块目标
模块 1	大数据入门	• 了解大数据的角色和重要性 • 讨论大数据在各行各业中的使用和应用 • 讨论大数据相关的主要技术 • 解释 Hadoop 生态系统中各种组件的角色 • 解释 MapReduce 的基础概念和它在 Hadoop 生态系统中的作用

续表

模块编号	模块名称	模块目标
模块 2	分析和 R 编程入门	• 讨论高级分析的重要性 • 介绍分析方法和工具的发展 • 讨论各种分析工具的特性 • 用 R 语言开发脚本 • 用 R 语言中的各种附加编辑器执行脚本 • 用 R 语言执行读写操作 • 用 R 语言操纵数据
模块 3	使用 R 语言进行数据分析	• 使用 R 脚本和函数 • 使用 R 函数环境和方法 • 执行数据样本总结步骤 • 使用积累的统计数据和汇总表 • 用 R 创建列表、矩阵和数据帧 • 使用 R 中的循环和条件执行 • 安装 RHadoop 和创建用户定义函数 • 用 R 实现图表分析 • 用 R 进行假设检验
模块 4	用 R 语言进行高级分析	• 描述线性回归分析及其应用 • 在 R 语言中应用线性回归分析的知识 • 从应用角度理解非线性回归 • 在 R 语言中应用非线性回归分析 • 解释聚类分析技术 • 用 R 实现聚类分析 • 探索用于构建决策树的基本概念 • 用 R 构建决策树 • 将 R 与 Hadoop 集成，以进行统计分析

《大数据分析师权威教程》第 2 卷的 3 个模块的具体名称和学习目标如表 2 所示。

表 2

模块编号	模块名称	模块目标
模块 1	机器学习的概念	• 讨论机器学习在技术上和商业上的应用 • 理解图模型的用途 • 用 R 实现图模型 • 理解贝叶斯网络表示法及其解读 • 用贝叶斯网络解决预测问题 • 探索人工神经网络及其结构和学习规则 • 阐述人工神经网络的训练 • 用 R 实现神经网络 • 用因子分析和主成分分析实现降维 • 从给定的预测因素列表识别最大影响因子/维度 • 解释支持向量机 • 用 R 语言实现支持向量机

续表

模块编号	模块名称	模块目标
模块 2	社交媒体、移动分析和可视化	• 应用可用于大数据实现的解决方案设计过程 • 分析业务环境中社交媒体所承担的角色 • 实施社交媒体分析 • 执行基本移动分析 • 讨论数据可视化及其重要性 • 使用表格进行数据可视化 • 有效地准备求职面试
模块 3	大数据分析的行业应用	• 理解保险业中的数据分析应用 • 理解金融机构中数据分析的实施 • 理解电信行业中的分析工具 • 实施在线客户细分中的分析

学习方法和特色

本书开发了一套独特的学习方法，这种专门设计的方法不仅以最大限度地学习大数据概念为目标，还注重对真实专业环境下应用这些概念的全面理解。

本书的独特方法和丰富特性简单介绍如下。

- 涵盖了大数据分析师必备的所有**大数据和 Hadoop 基础组件及相关组件的基本知识**，使学习者有可能在一个系列书中获得对所有相关知识、新兴技术和平台的了解。
- 在与大数据分析师关系最为密切的**描述性和预测性分析技术**上培养全面、结构化的技能，逐步理解**各种技术在 R 语言上的实施**（R 语言是最通用、使用最广泛的统计软件之一）。
- **基于场景的学习方法**，通过多种有代表性的现实场景的使用和案例研究，将 IT 基础知识融入现实环境，鼓励学习者积极、全面地学习和研究，实现体验式教学。
- 强调目标明确、**基于成果的学习**。每一讲都以“本讲目标”开始，该目标会进一步关联整个教程的更广泛的目标。
- **简明、循序渐进的**编程和编码**指导**，清晰地解释每行代码的基本原理。
- 强调**高效、实用的过程和技术**，帮助学习者深入理解巧妙、合乎道德的专业方法及其对业务的影响。

学习工具

下列学习工具将确保学习者高效地使用本教程。

- **模块目标**：列出某一讲所属模块的目标。
- **本讲目标**：列出与模块目标对应的本讲目标。
- **预备知识**：说明对特定部分或者整体概念的理解有特定作用的预备知识点。

- **交叉参考**：将整个模块中学到的相关概念联系起来，启发参与者理解分析中的不同功能、职责和挑战，确保任何概念都不是孤立地学习的。
- **总体情况**：不断提醒参与者，某个主题为什么是相关的，在行业中如何应用，从而为学习提供实践维度。
- **快速提示**：提供明智、高效地运用概念的简便技巧。
- **与现实生活的联系**：提供简短的案例分析和剪报，阐述概念在现实世界中的适用性。
- **技术材料**：提供加强技术诀窍理解的技巧和信息。
- **定义**：定义重要概念或者术语。
- **附加知识**：提供相关的附加信息。
- **知识检测点**：提出互动式课堂讨论的问题，强化每一讲之后的学习。
- **练习**：在每一讲结束时提出以知识为基础的实践问题，评估理解情况。
- **测试你的能力**：提供基于应用的实践问题。
- **备忘单**：提供本讲涵盖的重要步骤及过程的快速参考。

关键的大数据技术术语

大数据是一个非常年轻的行业，新的技术和术语每周都会出现。这种快节奏的环境是由开源社区、新兴技术公司以及 IBM、Oracle、SAP、SAS 和 Teradata 这样的业界巨人推动的。不用说，建立一个持久的权威术语表是很难的。鉴于这样的风险，我们在这里只提供一个小型的大数据词汇表，如表 3 所示。

表 3

术　语	定　义
算法	用来分析数据的数学方法。一般情况下，是一段计算过程；计算一个功能的指令列表；在软件中，这样一个过程以编程语言来实际实现
分析	一组用于查询和梳理平台数据的分析工具和计算能力
装置	专为特定活动集建立的一组优化硬件和软件
Avro	一个可编码 Hadoop 文件模式的数据序列化系统，特别擅长于数据解析，是 Apache Hadoop 项目的一部分
批处理	在后台运行、不与人发生交互的作业或进程
大数据	大数据事实上的标准定义是超越了传统的 3 个维度（数据量、多样性、速度）限制的数据。这 3 个维度的结合使得数据的摄取、处理和呈现更加复杂
Big Insights	IBM 的具有企业级增值组件的 Hadoop 商业发行版
Cassandra	由 Apache 软件基金会管理的开源列式数据库
Clojure	基于 LISP(从 20 世纪 50 年代起的人工智能编程语言事实标准)的动态编程语言，读作"closure"。通常用于并行数据处理
云	用以指代任何计算机运作的软件、硬件或服务资源的通用术语。它作为一种服务通过网络传送

续表

术　语	定　义
Cloudera	Hadoop 的第一个商业分销商。Cloudera 提供了 Hadoop 发行版的企业级增值组件
列式数据库	按列进行的数据存储与优化。使用基于列的数据，对于一些分析处理特别有用
复杂事件处理(CEP)	对实时发生事件进行分析并采取措施的过程
数据挖掘	利用机器学习，从数据中发现模式、趋势和关系的过程
分布式处理	在多个 CPU 上的程序执行
Dremel	一个可扩展、交互式、点对点分析查询系统，有能力在数秒内对数万亿行的表进行聚合查询
Flume	一种从 Web 服务器、应用服务器、移动设备等目标抓取数据填充 Hadoop 的框架
网格	松散耦合的服务器通过网络连接起来，并行处理工作负载
Hadapt	一家提供 Hadoop 相关插件的商业供应商，这个插件可以通过高速连接器在 HDFS 和关系型表中移动数据
Hadoop	一个开源项目框架，可以在计算机集群（网格）中存储大量的非结构化数据（HDFS）并在其中对其进行处理（MapReduce）
HANA	来自 SAP 的内存处理计算平台，为大容量事务和实时分析而设计
HBase	一种分布式、列式存储的 NoSQL 数据库
HDFS	Hadoop 文件系统，是 Hadoop 的存储机制
Hive	一种 Hadoop 的类 SQL 查询语言
Norton	具有企业级增值工作组件的 Hadoop 商业发行版
HPC	高性能计算。通俗地说，就是为高速浮点处理、内存磁盘并行化而设计的设备
HAStreaming	为 Hadoop 提供实时 CEP（复杂事件处理）的 Hadoop 商业插件
机器学习	从经验数据中学习，然后利用这些经验教训去预测未来新数据的结果的算法技术
Mahout	为 Hadoop 创建可伸缩机器学习算法库的 Apache 项目，主要用 MapReduce 实现
MapR	具有企业级增值组件的 Hadoop 商业发行版
MapReduce	一种 Hadoop 计算批处理框架，其中的作业大部分用 Java 编写。作业将较大的问题分解为较小的部分，并将工作负载分布到网格中，使多个作业能够同时进行（mapper）。主作业（reducer）收集所有中间结果并将其组合起来
大规模并行处理（MPP）	能协调并行程序执行的系统（操作系统、处理器和内存）
MPP 装置	带有处理器、内存、磁盘和软件，能够并行处理工作负载的集成平台
MPP 数据库	一种已为 MPP 环境优化的数据库
MongoDB	一种用 C++编写的可扩展、高性能的开源 NoSQL 数据库
NoSQL 数据库	一个用以描述数据库的术语。这种数据库不使用 SQL 作为数据库的数据主检索，且可以是任意类型。NoSQL 拥有有限的传统功能，并为可扩展性和高性能检索及添加而设计。通常情况下，NoSQL 数据库利用键值对存储数据，能够很好地处理在本质上不相关的数据
Oozie	一个工作流处理系统，允许用户定义一系列用各种语言（如 MapReduce、Pig 和 Hive）编写的作业
Pig	一种使用查询语言（Pig Latin）的分布式处理框架，用以执行数据转换。目前，Pig Latin 程序被转换为 MapReduce 作业，在 Hadoop 上运行

续表

术　语	定　义
R	一种开源的语言和环境，用以统计计算和图形化
实时	今天，通俗地说，它被定义为即时处理。实时处理起源于 20 世纪 50 年代，当时多任务处理机提供了为更高优先级任务的执行而“中断”一个任务的能力。这些类型的机器为空间计划、军事应用和多种商业控制系统提供了动力
关系型数据库	按照行和列存储和优化数据
Scording	使用预测模型，预测新数据的未来结果
半结构化数据	依靠可用的格式描述符，把非结构化的数据放入结构中
Spark	内存分析计算处理的高性能处理框架，通常被用来做实时查询
SQL（结构化查询语言）	关系型数据库中，存储、访问和操作数据的语言
Sqoop	一种命令行工具，具有把单个表或整个数据库导入 Hadoop 文件中等能力
Sorm	分布式、容错、实时分析处理的开源框架
结构化数据	有预设定数据格式的数据
非结构化数据	无预先设定结构的数据
Whirr	一套用于运行云服务的库
YARN	Apache Hadoop 的下一代计算框架，除了 MapReduce 之外还支持编程范式

提示

本书提供配套的网上下载资源，包括预备知识内容、PowerPoint 幻灯片、模拟试题和其他附加资源（包括额外的面试题）。以上所有资源均为英文资料。[①]

“知识检测点”和“测试你的能力”环节中的问题可能需要使用特定数据集。读者可以使用本书配套的网上下载资源中提供的数据集，也可以使用从网上找到的合适的数据或者自己生成数据。

① 本书配套的网上下载资源请登录异步社区（www.epubit.com.cn），访问本书对应页面下载。——编者注

目 录

模块 1 大数据入门

模块 2 分析和 R 编程入门

模块 3　使用 R 进行数据分析

模块 4 使用 R 进行高级分析