异步图书
www.epubit.com

“十三五”
国家重点图书出版规划项目

Official Study Kit for Wrox Certified
Big Data Developer Program

大数据开发者权威教程

大数据技术与编程基础

Wrox 国际 IT 认证项目组／编　顾晨／译　黄倩／审校

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 Developer Program

大数据开发者权威教程

大数据技术与编程基础

Wrox 国际 IT 认证项目组／编　顾晨／译　黄倩／审校

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据开发者权威教程 : 大数据技术与编程基础 / Wrox国际IT认证项目组编 ; 顾晨译. -- 北京 : 人民邮电出版社, 2018.12
书名原文: Official Study Kit for Wrox Certified Big Data Developer Program
ISBN 978-7-115-49350-7

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

中国版本图书馆CIP数据核字(2018)第213681号

内容提要

“大数据”近年成为IT领域的热点话题，人们每天都会通过互联网、移动设备等产生大量数据。如何管理大数据、掌握大数据的核心技术、理解大数据相关的生态系统等，是作为大数据开发者必须学习和熟练掌握的。本系列书以“大数据开发者”应掌握的技术为主线，共分两卷，以7个模块分别介绍如何管理大数据生态系统、如何存储和处理数据、如何利用Hadoop工具、如何利用NoSQL与Hadoop协同工作，以及如何利用Hadoop商业发行版和管理工具。本系列书涵盖了大数据开发工作的核心内容，全面且详尽地涵盖了大数据开发的各个领域。

本书为第1卷，共4个模块，分别介绍大数据基础知识、大数据生态系统的管理、HDFS和MapReduce以及Hadoop工具（如Hive、Pig和Oozie等）。

本书适用于想成为大数据开发者以及所有对大数据开发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　　顾　晨
审　　校　黄　倩
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
三河市君旺印务有限公司印刷

◆ 开本：800×1000　1/16
印张：32.5
字数：747千字　　2018年12月第1版
印数：1－2 600册　　2018年12月河北第1次印刷

著作权合同登记号　图字：01-2015-2407号

定价：109.00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

译者简介

顾晨，男，硕士、PMP、信息系统项目管理师。毕业于上海交通大学。曾获邀参加旧金山的Google I/O 大会。喜欢所有与编程相关的事物，拥有 14 年的编程经验。对于大数据、SAP HANA 数据库和思科技术有着极其浓厚的兴趣，是国内较早从事 HANA 数据库研究的人员之一。先后录制了 MCSE、CCNP 等多种教学视频，在多家知名网站发布。精通 C#、Java 编程，目前正致力于人脸识别、室内定位和门店人流统计方面的研究。

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”系列图书！

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意义的工作。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”系列图书（两卷）和“大数据开发者权威教程”系列图书（两卷）。

这两个系列共 4 本书，旨在培育新一代年轻 IT 专业人士，使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。

“大数据开发者权威教程”系列图书概览

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。

“大数据开发者权威教程”系列图书的目标是培养新一代的国际化全能大数据程序员、开发

者和技术专家，熟悉大数据的相关工具、平台和架构，帮助企业有效地存储、管理和处理海量和多样的数据。同时，该教程有助于读者了解如何有效地整合、实现、定制和管理大数据基础架构。

本系列图书旨在：

- 为参与者提供处理大数据的**技术、存储、处理、管理**和**安全基础架构**方面的技能；
- 为参与者提供与 **Hadoop** 及其**组件工具**协同工作的经验；
- 使参与者可以开发 **MapReduce** 和 **Pig** 程序，操纵分布式文件，以及了解支持 MapReduce 程序的 API；
- 参与者可以熟悉一些流行的 **Hadoop 商业发行版系统**，如 Cloudera、Hortonworks 和 Greenplum；
- 最后包含一个**完整的项目**，使参与者能够开发一个集成的大数据应用程序。

参与者的必备条件

要阅读这个系列图书，读者必须具备以下基础知识。

- 编程基础（含面向对象编程的基础）。
- 脚本语言的基础（如 Perl 或 Ruby）。
- 操作 Linux/Unix 操作系统的基础。
- 对 Java 编程语言有很好的理解：
  - Java 核心技术；
  - 了解 SQL 语句。

建议的学习时间

“大数据开发者权威教程”系列图书由 **7 个学习模块**（第 1 卷包括 4 个模块，第 2 卷包括 3 个模块）组成。

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**，建议学习 7 个模块，为成为合格的大数据开发者做好充足的就业前准备。**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

第 1 卷《大数据开发者权威教程：大数据技术与编程基础》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
模块 2	管理大数据生态系统	• 讨论大数据所需的关键技术基础 • 把传统数据管理系统与大数据管理系统进行对比 • 评估大数据分析的关键需求 • 讨论整合数据的流程 • 解释实时数据的相关性 • 在企业中评估实施大数据的需求 • 解释如何使用大数据和实时数据作为业务规划工具
模块 3	存储和处理数据：HDFS 和 MapReduce	• 分析 Hadoop 的大数据的 HDFS 和 HBase 存储模型 • 开发基本的 MapReduce 程序 • 利用 MapReduce 的可扩展性，进行定制执行 • 在设计时进行 MapReduce 程序的测试和调试 • 在给定的场景下实现 MapReduce 程序
模块 4	利用 Hadoop 工具 Hive、Pig 和 Oozie 提升效率	• 讨论了 Hive 的数据存储原理 • 在 Hive 中执行数据操作 • 实现 Hive 的提前查询特性 • 解释 Hive 环境支持的文件格式和记录格式 • 利用 Pig 使 MapReduce 的设计和实现自动化 • 使用 Oozie 分析工作流的设计和管理 • 设计和实现一个 Oozie 工作流

第 2 卷《大数据开发者权威教程：NoSQL、Hadoop 组件及大数据实施》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	额外的 Hadoop 工具：ZooKeeper、Sqoop、Flume、YARN 和 Storm	• 利用 Apache Zookeeper 实现分布式协同服务 • 将数据从非 Hadoop 的存储系统加载到 Hive 和 HBase 中 • 描述 Flume 的角色 • 使用 Flume 进行数据汇总 • 解释 YARN 的角色，并将它与 Hadoop 1.0 中的 MapReduce 进行对比 • 解释如何利用运行在 YARN 上的 Storm 管理 Hadoop 上的实时数据 • 开发运行在 YARN 上的 Storm 应用程序
模块 2	利用 NoSQL 和 Hadoop：实时、安全和云	• 与 NoSQL 的界面和交互 • 执行 CRUD 操作和各种 NoSQL 数据库查询 • 分析在 Hadoop 中安全是如何实现的 • 配置运行在 Amazon Web Services（AWS）中的 Hadoop 应用 • 设计 Hadoop 实时应用

续表

模块编号	模块名称	模块目标
模块 3	Hadoop 商业发行版和管理工具	• 探讨 Cloudera 管理器平台 • 利用 Cloudera 管理器进行服务的添加和管理 • 为各种平台配置 Hive 的元数据 • 为 Hive 安装 Cloudera 管理器 4.5 版 • 为大数据分析部署 Hortonworks 数据平台（HDP）集群 • 使用 Talend Open Studio 进行数据分析 • 解释 Greenplum Pivotal HD 架构 • 讨论并安装 InfoSphere BigInsights • 讨论并安装 MapR 和 MapR 沙盒 • 为求职面试做有效的准备

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。

- 涵盖了大数据开发者必备的**所有大数据**和 **Hadoop 基础组件及相关组件的基本知识**，使参与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**大数据架构、大数据应用程序开发**以及与**大数据实施**相关的**产业相关技术**有着极密切关联的编程和技术领域中，锻炼自己全面的和结构化的本领。
- **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，鼓励参与者积极、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**，帮助参与者深入理解巧妙且符合道德伦理的专业实践及其对业务的影响。

学习工具

下列学习工具将确保参与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**：说明对某一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**：将整个模块中的相关概念联系起来，启发参与者理解分析工具的不同功能、职责和挑战，确保概念不被孤立地学习。

- **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参考。
- **快速提示**：提供高效地运用概念的技巧。
- **与现实生活的联系**：提供简短的案例分析和简报，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方法和信息。
- **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**：提供这一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的硬件和软件
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提取、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
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

续表

术　语	定　义
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 和关系型表之间移动数据
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 作为数据库的主要检索，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
R	一种开源的语言和环境，用以统计计算和图形化
实时	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

续表

术　语	定　义
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 文件中的能力
Storm	分布式、容错、实时分析处理的开源框架
结构化数据	有预先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（https://www.epubit.com），访问本书对应页面下载。——编者注

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

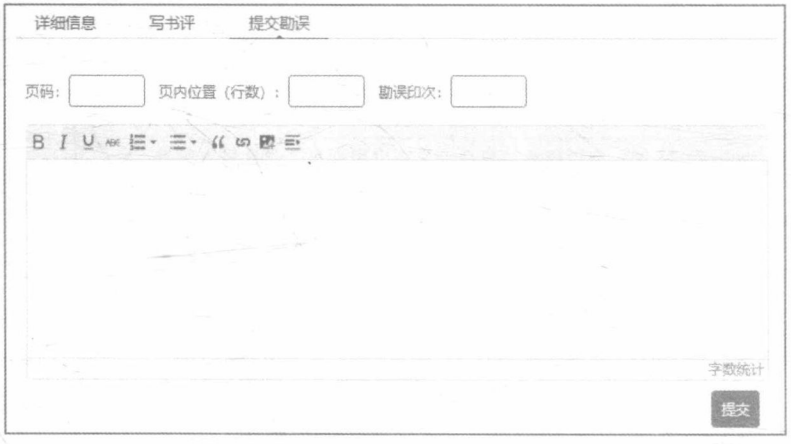
本书提供一些配套资源，要获得这些配套资源，请在异步社区本书页面中点击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目 录

模块 1 大数据入门

模块 2 管理大数据生态系统