大数据
BIG DATA

吴章勇 杨强 编著

大数据 Hadoop 3.X

分布式处理实战

01

项目大

深度剖析三大企业级项目实战案例

02

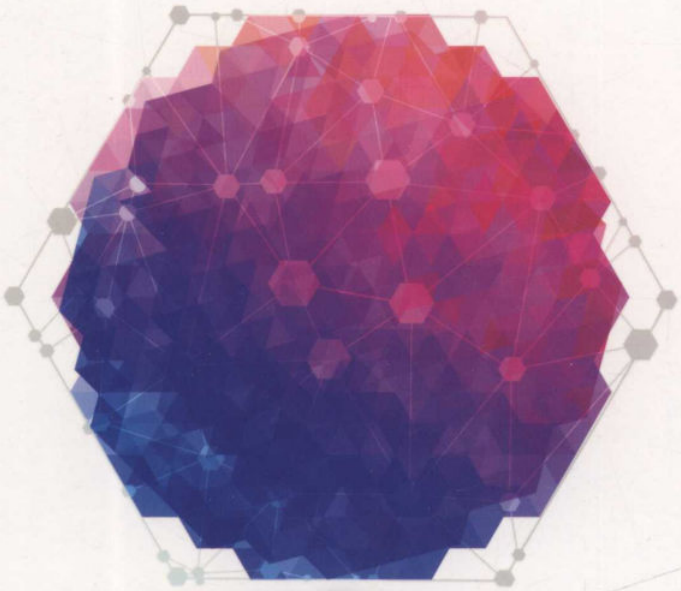
内容全

详细介绍 HDFS、MapReduce、HBase、Hive、Sqoop、Spark 等主流大数据工具

03

资源多

赠送 12 小时视频讲解和全书配套范例源码

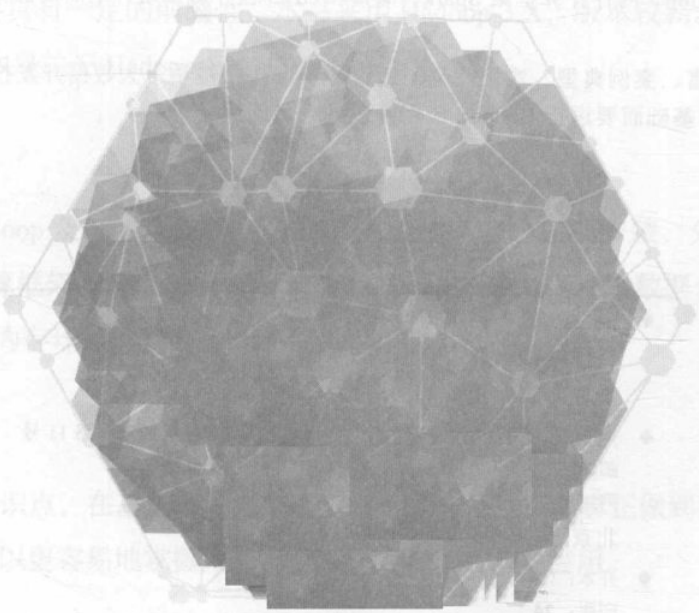


中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

大数据
BIG DATA

吴章勇 杨强 编著

大数据
Hadoop 3.X
分布式处理实战



人民邮电出版社
北京

图书在版编目（CIP）数据

大数据Hadoop 3.X分布式处理实战 / 吴章勇，杨强编著. -- 北京 : 人民邮电出版社，2020.4
ISBN 978-7-115-52466-9

Ⅰ. ①大… Ⅱ. ①吴… ②杨… Ⅲ. ①数据处理软件
Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2019)第252598号

内 容 提 要

本书以实战开发为原则，以 Hadoop 3.X 生态系统内的主要大数据工具整合应用及项目开发为主线，通过 Hadoop 大数据开发中常见的 11 个典型模块和 3 个完整项目案例，详细介绍 HDFS、MapReduce、HBase、Hive、Sqoop、Spark 等主流大数据工具的整合使用。本书附带资源包括本书核心内容的教学视频，本书所涉及的源代码、参考资料等。

全书共 14 章，分为 3 篇，涵盖的主要内容有 Hadoop 及其生态组件伪分布式安装和完全分布式安装、分布式文件系统 HDFS、分布式计算框架 MapReduce、NoSQL 数据库 HBase、分布式数据仓库 Hive、数据转换工具 Sqoop、内存计算框架 Spark、海量 Web 日志分析系统、电商商品推荐系统、分布式垃圾消息识别系统等。

本书内容丰富、案例典型、实用性强，适合各个层次希望学习大数据开发技术的人员阅读，尤其适合有一定 Java 基础而要进行 Hadoop 应用开发的人员阅读。

◆ 编　　著　吴章勇　杨　强
责任编辑　俞　彬
责任印制　王　郁　马振武

◆ 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
北京市艺辉印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：24
字数：478 千字　　2020 年 4 月第 1 版
印数：1 – 2 500 册　　2020 年 4 月北京第 1 次印刷

定价：79.00 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

前言
Foreword

随着云时代的来临，移动互联网、电子商务、物联网以及社交媒体快速发展，全球的数据正在以几何速度呈爆炸性增长，大数据也吸引了越来越多的人关注。大数据的核心技术就是 Hadoop。目前市面上关于 Hadoop 的书有很多，但基本是关于 Hadoop 1.X 或 Hadoop 2.X 的，而且偏重理论讲述，缺少实践案例。本书从 Hadoop 3.X 实例出发，通过“理论 + 实践 + 视频”的方式，帮助读者轻松掌握大数据技术。特别值得一提的是，本书讲解了日志分析、推荐系统、垃圾消息识别 3 个企业级的综合大数据项目案例，读者稍加改造，即可在生产环境中使用，具有重大的实用价值，也可供在校大学生或研究生毕业设计时参考。

本书有何特色

1．版本较新

技术研究需要具有一定的前瞻性，本书采用 Hadoop 3.X，版本较新。目前国内关于 Hadoop 的图书基本是关于 Hadoop 1.X 或 Hadoop 2.X 的。

2．知识全面

本书包括 Hadoop 及其生态组件伪分布式安装和完全分布式安装、分布式文件系统 HDFS、分布式计算框架 MapReduce、NoSQL 数据库 HBase、分布式数据仓库 Hive、数据转换工具 Sqoop、内存计算框架 Spark 等主要大数据技术。

3．重视实战

针对每一个知识点，在基本理论讲述后都提供了实战项目，真正做到学以致用。读者通过实战项目，可以更容易地掌握大数据技术在具体工作中的应用。

4．视频讲解

本书作者具有丰富的 IT 培训和视频录制经验，针对每章内容精心录制了多个讲解视频。全书有 32 个相关视频，视频总时长超过 12 小时，特别是环境搭建、项目运行、源码分析等场景，通过视频学习将更加轻松。

5．图文并茂

一图胜过千言万语，全书共有超过 200 幅插图，用于展示语言难以描述的内容，同时插图也有助于增加阅读的趣味性。

6．在线答疑

本书提供答疑 QQ 群，在线答疑，群号是 243363382。也可以通过作者的 QQ 号进行在线交流，作者的 QQ 号是 107964558。

7．电子资源

本书在附带的电子资源中，提供了每章的相关视频、源代码及测试数据，用 Eclipse 工具打开源代码即可运行。通过运行效果来分析源代码，理解会更容易、更深刻。

读者可扫描“职场研究社”二维码，关注后回复“52466”即可获取电子资源下载链接，也可以扫描云课二维码，手机端在线观看视频。

职场研究社

云课

适合阅读本书的读者

（1）渴望转型进入大数据领域的程序员。

（2）希望学习大数据技术的在校大学生或研究生。

（3）希望提升技能的初级大数据领域从业人员。

（4）希望研究“推荐系统”等大数据典型应用的大数据开发工程师。

目录

Contents

第一篇 Hadoop 技术

第三篇 实战篇

2.2 Hadoop 伪分布式安装

Hadoop 伪分布式安装主要包括安装 JDK 和安装 Hadoop 两步，但为了使 Windows 系统能与虚拟机 CentOS 进行通信，还需安装 WinSCP 和 PieTTY 两个工具软件。

2.2.1 安装 WinSCP

WinSCP 工具可以实现 Windows 系统和 Linux 系统之间共享文件。WinSCP 工具安装非常简单，从网上下载 winscp516setup.exe，一直单击“下一步”按钮即可完成安装。安装完成后打开 WinSCP，输入 CentOS 7 的主机名、用户名、密码，先单击“保存”按钮，再单击“登录”按钮，即可登录到 CentOS 7，如图 2.28 所示。

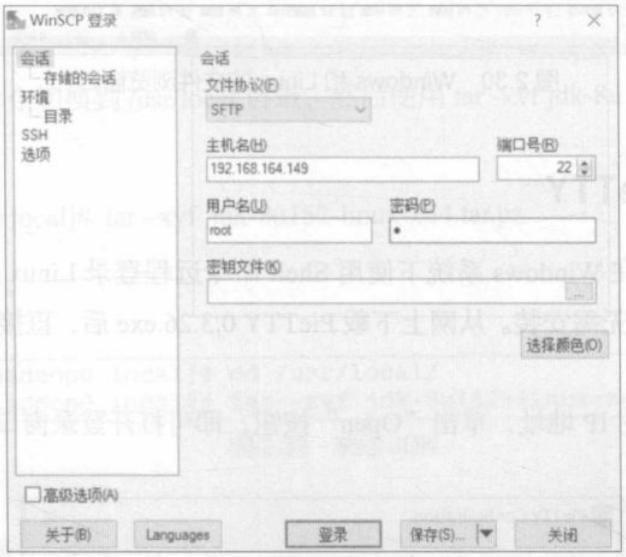


图 2.28 WinSCP 登录

在弹出的警告中单击“是”按钮，如图 2.29 所示。

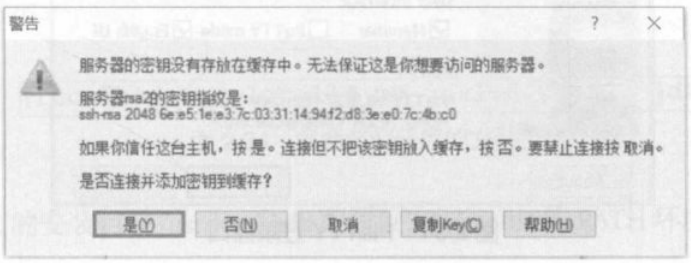


图 2.29 密钥指纹