

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

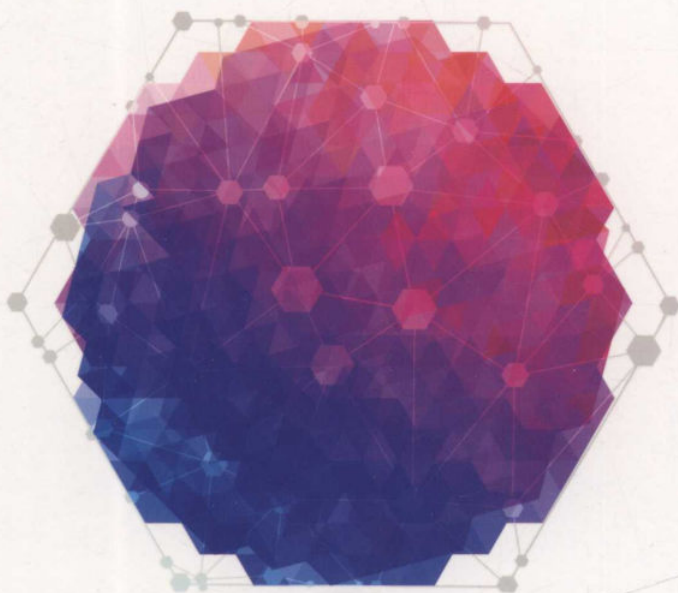
内容全

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

03

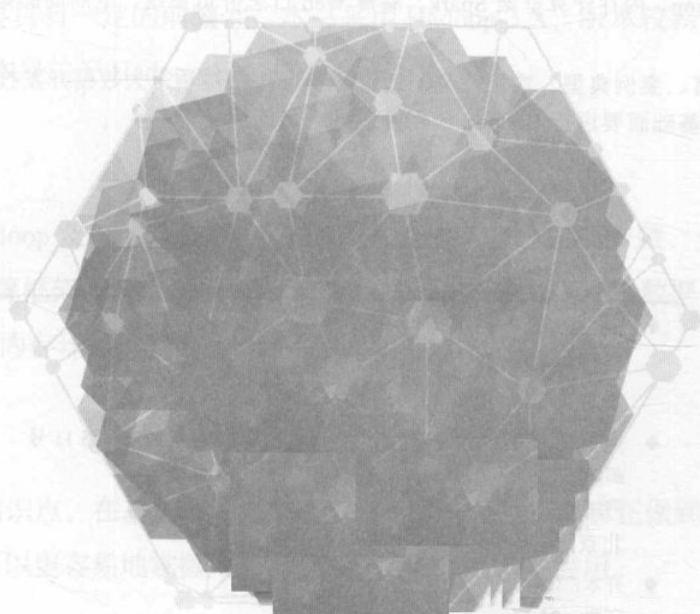
资源多

赠送 12 小时视频讲解和
全书配套范例源码



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 (C I P) 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

F oreword

前言

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1章 大数据与 Hadoop 概述.....03

1.1 大数据概述.....03

1.1.1 大数据的定义.....03

1.1.2 大数据行业的发展.....04

1.1.3 大数据的典型应用.....04

1.2 Hadoop 概述.....06

1.2.1 Hadoop 简介.....06

1.2.2 Hadoop 生态子项目.....07

1.2.3 Hadoop 3.X 的新特性.....09

1.3 小结.....09

1.4 配套视频.....10

第2章 Hadoop 伪分布式安装.....11

2.1 Hadoop 伪分布式安装前的准备.....11

2.1.1 安装 VMware.....11

2.1.2 安装 CentOS 7.....12

2.1.3 配置 CentOS 7: 接受协议.....15

2.1.4 配置 CentOS 7: 登录系统.....16

2.1.5 配置 CentOS 7: 设置 IP.....16

2.1.6 配置 CentOS 7: 修改主机名.....17

2.1.7 配置 CentOS 7: 配置 hosts 文件.....18

2.1.8 配置 CentOS 7: 关闭防火墙.....18

2.1.9 配置 CentOS 7: 禁用 selinux.....19

2.1.10 配置 CentOS 7: 设置 SSH 免密码登录.....19

2.1.11 配置 CentOS 7: 重启.....20

2.2 Hadoop 伪分布式安装	21
2.2.1 安装 WinSCP	21
2.2.2 安装 PieTTY	22
2.2.3 安装 JDK	23
2.2.4 安装 Hadoop	24
2.3 Hadoop 验证	28
2.3.1 格式化	28
2.3.2 启动 Hadoop	29
2.3.3 查看 Hadoop 相关进程	29
2.3.4 浏览文件	30
2.3.5 浏览器访问	30
2.4 小结	31
2.5 配套视频	31

第3章 Hadoop 分布式文件系统——HDFS 32

3.1 HDFS 原理	32
3.1.1 HDFS 的假设前提和设计目标	32
3.1.2 HDFS 的组件	33
3.1.3 HDFS 数据复制	36
3.1.4 HDFS 健壮性	36
3.1.5 HDFS 数据组织	38
3.2 HDFS Shell	39
3.2.1 Hadoop 文件操作命令	39
3.2.2 Hadoop 系统管理命令	44
3.3 HDFS Java API	46
3.3.1 搭建 Linux 下 Eclipse 开发环境	46
3.3.2 为 Eclipse 安装 Hadoop 插件	47
3.3.3 HDFS Java API 示例	49
3.4 小结	56
3.5 配套视频	56

第4章	分布式计算框架 MapReduce	57
4.1	MapReduce 原理	57
4.1.1	MapReduce 概述	57
4.1.2	MapReduce 的主要功能	59
4.1.3	MapReduce 的处理流程	59
4.2	MapReduce 编程基础	61
4.2.1	内置数据类型介绍	61
4.2.2	WordCount 入门示例	63
4.2.3	MapReduce 分区与自定义数据类型	67
4.3	MapReduce 综合实例——数据去重	71
4.3.1	实例描述	71
4.3.2	设计思路	72
4.3.3	程序代码	73
4.3.4	运行结果	74
4.4	MapReduce 综合实例——数据排序	75
4.4.1	实例描述	75
4.4.2	设计思路	76
4.4.3	程序代码	77
4.4.4	运行结果	79
4.5	MapReduce 综合实例——求学生平均成绩	79
4.5.1	实例描述	79
4.5.2	设计思路	80
4.5.3	程序代码	81
4.5.4	运行结果	83
4.6	MapReduce 综合实例——WordCount 高级示例	84
4.7	小结	87
4.8	配套视频	87

第二篇 Hadoop 生态系统的主要大数据工具整合应用

第5章	NoSQL 数据库 HBase	91
------------	------------------------	-----------

5.1 HBase 原理	91
5.1.1 HBase 概述	91
5.1.2 HBase 核心概念	92
5.1.3 HBase 的关键流程	95
5.2 HBase 伪分布式安装	97
5.2.1 安装 HBase 的前提条件	98
5.2.2 解压并配置环境变量	98
5.2.3 配置 HBase 参数	99
5.2.4 验证 HBase	100
5.3 HBase Shell	103
5.3.1 HBase Shell 常用命令	103
5.3.2 HBase Shell 综合示例	109
5.3.3 HBase Shell 的全部命令	112
5.4 小结	114
5.5 配套视频	114

第6章 HBase 高级特性 115

6.1 HBase Java API	115
6.1.1 HBase Java API 介绍	115
6.1.2 HBase Java API 示例	120
6.2 HBase 与 MapReduce 的整合	130
6.2.1 HBase 与 MapReduce 的整合概述	130
6.2.2 HBase 与 MapReduce 的整合示例	130
6.3 小结	134
6.4 配套视频	134

第7章 分布式数据仓库 Hive 135

7.1 Hive 概述	135
7.1.1 Hive 的定义	135
7.1.2 Hive 的设计特征	136

7.1.3	Hive 的体系结构	136
7.2	Hive 伪分布式安装	137
7.2.1	安装 Hive 的前提条件	137
7.2.2	解压并配置环境变量	138
7.2.3	安装 MySQL	139
7.2.4	配置 Hive	143
7.2.5	验证 Hive	145
7.3	Hive QL 的基础功能	146
7.3.1	操作数据库	146
7.3.2	创建表	147
7.3.3	数据准备	150
7.4	Hive QL 的高级功能	153
7.4.1	select 查询	154
7.4.2	函数	154
7.4.3	统计函数	154
7.4.4	distinct 去除重复值	155
7.4.5	limit 限制返回记录的条数	156
7.4.6	为列名取别名	156
7.4.7	case when then 多路分支	156
7.4.8	like 模糊查询	157
7.4.9	group by 分组统计	157
7.4.10	having 过滤分组统计结果	157
7.4.11	inner join 内联接	158
7.4.12	left outer join 和 right outer join 外联接	159
7.4.13	full outer join 外部联接	159
7.4.14	order by 排序	160
7.4.15	where 查找	160
7.5	小结	161
7.6	配套视频	162
第8章	Hive 高级特性	163
8.1	Beeline	163

8.1.1	使用 Beeline 的前提条件	163
8.1.2	Beeline 的基本操作	164
8.1.3	Beeline 的参数选项与管理命令	166
8.2	Hive JDBC	167
8.2.1	运行 Hive JDBC 的前提条件	167
8.2.2	Hive JDBC 基础示例	167
8.2.3	Hive JDBC 综合示例	169
8.3	Hive 函数	174
8.3.1	内置函数	174
8.3.2	自定义函数	175
8.4	Hive 表的高级特性	181
8.4.1	外部表	181
8.4.2	分区表	182
8.5	小结	185
8.6	配套视频	185

第9章 数据转换工具 Sqoop

9.1	Sqoop 概述与安装	186
9.1.1	Sqoop 概述	186
9.1.2	Sqoop 安装	187
9.2	Sqoop 导入数据	189
9.2.1	更改 MySQL 的 root 用户密码	189
9.2.2	准备数据	190
9.2.3	导入数据到 HDFS	191
9.2.4	查看 HDFS 数据	192
9.2.5	导入数据到 Hive	193
9.2.6	查看 Hive 数据	193
9.3	Sqoop 导出数据	194
9.3.1	准备 MySQL 表	194
9.3.2	导出数据到 MySQL	194
9.3.3	查看 MySQL 中的导出数据	195
9.4	深入理解 Sqoop 的导入与导出	196

9.5 小结.....	203
9.6 配套视频.....	203

第10章 内存计算框架 Spark 204

10.1 Spark 入门.....	204
10.1.1 Spark 概述.....	204
10.1.2 Spark 伪分布式安装	205
10.1.3 由 Java 到 Scala.....	209
10.1.4 Spark 的应用	212
10.1.5 Spark 入门示例	217
10.2 Spark Streaming	220
10.2.1 Spark Streaming 概述	220
10.2.2 Spark Streaming 示例	221
10.3 Spark SQL.....	224
10.3.1 Spark SQL 概述	224
10.3.2 spark-sql 命令	225
10.3.3 使用 Scala 操作 Spark SQL	227
10.4 小结.....	228
10.5 配套视频.....	229

第11章 Hadoop 及其常用组件集群安装 230

11.1 Hadoop 集群安装	230
11.1.1 安装并配置 CentOS.....	230
11.1.2 安装 JDK.....	236
11.1.3 安装 Hadoop.....	237
11.1.4 远程复制文件.....	241
11.1.5 验证 Hadoop.....	242
11.2 HBase 集群安装.....	244
11.2.1 解压并配置环境变量	244
11.2.2 配置 HBase 参数	245

11.2.3	远程复制文件	246
11.2.4	验证 HBase	247
11.3	Hive 集群安装	249
11.3.1	解压并配置环境变量	249
11.3.2	安装 MySQL	250
11.3.3	配置 Hive	252
11.3.4	验证 Hive	254
11.4	Spark 集群安装	254
11.4.1	安装 Scala	254
11.4.2	安装 Spark	254
11.4.3	配置 Spark	255
11.4.4	远程复制文件	256
11.4.5	验证 Spark	257
11.5	小结	259
11.6	配套视频	259

第三篇 实战篇

第12章 海量 Web 日志分析系统 263

12.1	案例介绍	263
12.1.1	分析 Web 日志数据的目的	263
12.1.2	Web 日志分析的典型应用场景	265
12.1.3	日志的不确定性	265
12.2	案例分析	266
12.2.1	日志分析的 KPI	267
12.2.2	案例系统结构	267
12.2.3	日志分析方法	268
12.3	案例实现	273
12.3.1	定义日志相关属性字段	273
12.3.2	数据合法标识（在分析时是否被过滤）	274
12.3.3	解析日志	274
12.3.4	日志合法性过滤	275

12.3.5	页面访问量统计的实现	276
12.3.6	页面独立 IP 访问量统计的实现	278
12.3.7	用户单位时间 PV 的统计实现	280
12.3.8	用户访问设备信息统计的实现	282
12.4	小结	283
12.5	配套视频	283

第13章 电商商品推荐系统.....284

13.1	案例介绍	284
13.1.1	推荐算法	284
13.1.2	案例的意义	285
13.1.3	案例需求	285
13.2	案例设计	286
13.2.1	协同过滤	286
13.2.2	基于用户的协同过滤算法	289
13.2.3	基于物品的协同过滤算法	292
13.2.4	算法实现设计	295
13.2.5	推荐步骤与架构设计	298
13.3	案例实现	298
13.3.1	实现 HDFS 文件操作工具	299
13.3.2	实现任务步骤 1: 汇总用户对所有物品的评分信息	302
13.3.3	实现任务步骤 2: 获取物品同现矩阵	305
13.3.4	实现任务步骤 3: 合并同现矩阵和评分矩阵	307
13.3.5	实现任务步骤 4: 计算推荐结果	310
13.3.6	实现统一的任务调度	316
13.4	小结	317
13.5	配套视频	317

第14章 分布式垃圾消息识别系统.....318

14.1	案例介绍	318
14.1.1	案例内容	318

14.1.2	案例应用的主体结构	319
14.1.3	案例运行结果	321
14.2	RPC 远程方法调用的设计	322
14.2.1	Java EE 的核心优势: RMI	322
14.2.2	RMI 的基本原理	324
14.2.3	自定义 RPC 组件分析	325
14.3	数据分析设计	328
14.3.1	垃圾消息识别算法——朴素贝叶斯算法	328
14.3.2	进行分布式贝叶斯分类学习时的全局计数器	330
14.3.3	数据清洗分析结果存储	332
14.4	案例实现	333
14.4.1	自定义的 RPC 组件服务端相关实现	333
14.4.2	自定义的 RPC 组件客户端相关实现	342
14.4.3	业务服务器实现	347
14.4.4	业务客户端实现	367
14.5	小结	370
14.6	配套视频	370

第12章	海量 Web 日志分析系统	261
12.1	案例介绍	263
12.1.1	海量 Web 日志分析系统背景	263
12.1.2	海量 Web 日志分析系统需求	265
12.1.3	海量 Web 日志分析系统架构	265
12.2	案例设计	267
12.2.1	日志分析需求	267
12.2.2	日志分析系统架构	267
12.2.3	日志分析方法	269
12.3	案例实现	273
12.3.1	定义日志相关函数	273
12.3.2	数据清洗与预处理	273
12.3.3	数据清洗与预处理	273
12.3.4	日志清洗与预处理	273

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。

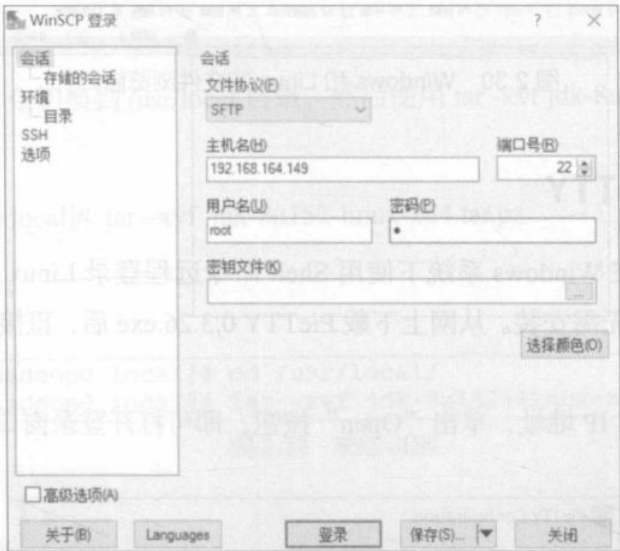


图 2.28 WinSCP 登录

在弹出的警告中单击“是”按钮，如图 2.29 所示。

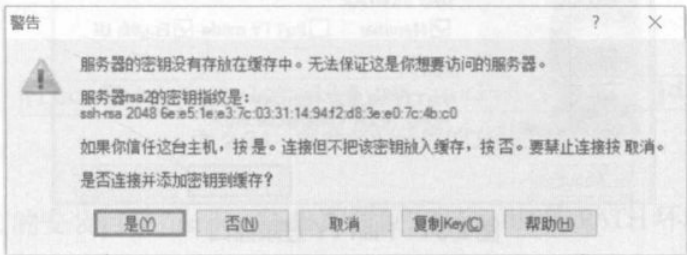


图 2.29 密钥指纹

