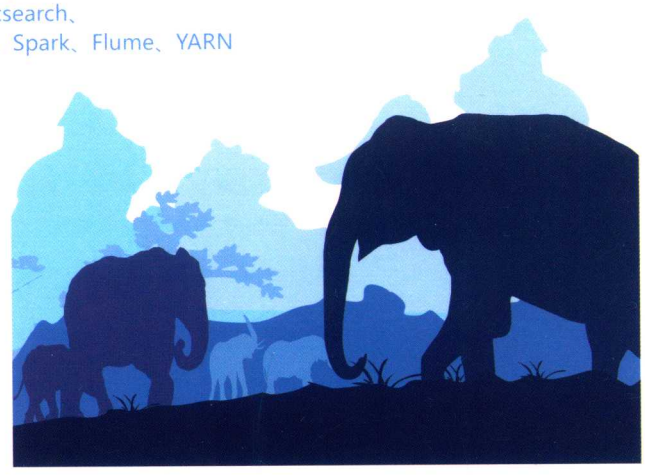
HDFS、MapReduce、ZooKeeper、HBase、Hive、
Sqoop、Kafka、Storm、
Elasticsearch、
Scala、Spark、Flume、YARN



Hadoop生态系统及大数据开发技术全接触

基础知识 / 架构原理 / 集群搭建 / 常用Shell命令
/ API操作 / 源码剖析 / 案例开发

Hadoop
大数据技术开发实战

张伟洋 著

清華大學出版社

Hadoop
大数据技术开发实战

张伟洋 著

清華大學出版社
北京

内 容 简 介

本书以 Hadoop 及其周边框架为主线，介绍了整个 Hadoop 生态系统主流的大数据开发技术。全书共 16 章，第 1 章讲解了 VMware 中 CentOS 7 操作系统的安装；第 2 章讲解了大数据开发之前对操作系统集群环境的配置；第 3~16 章讲解了 Hadoop 生态系统各框架 HDFS、MapReduce、YARN、ZooKeeper、HBase、Hive、Sqoop 和数据实时处理系统 Flume、Kafka、Storm、Spark 以及分布式搜索系统 Elasticsearch 等的基础知识、架构原理、集群环境搭建，同时包括常用的 Shell 命令、API 操作、源码剖析，并通过实际案例加深对各个框架的理解与应用。通过阅读本书，读者即使没有任何大数据基础，也可以对照书中的步骤成功搭建属于自己的大数据集群并独立完成项目开发。

本书可作为 Hadoop 新手入门的指导书，也可作为大数据开发人员的随身手册以及大数据从业者的参考用书。

图书在版编目（CIP）数据

Hadoop 大数据技术开发实战/张伟洋著. —北京：清华大学出版社，2019
ISBN 978-7-302-53402-0

Ⅰ. ①H… Ⅱ. ①张… Ⅲ. ①数据处理软件－程序设计 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 178532 号

责任编辑： 王金柱
封面设计： 王　翔
责任校对： 闫秀华
责任印制： 杨　艳

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： 三河市铭诚印务有限公司
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 29.75　　**字　　数：** 762 千字
版　　次： 2019 年 10 月第 1 版　　**印　　次：** 2019 年 10 月第 1 次印刷
定　　价： 99.00 元

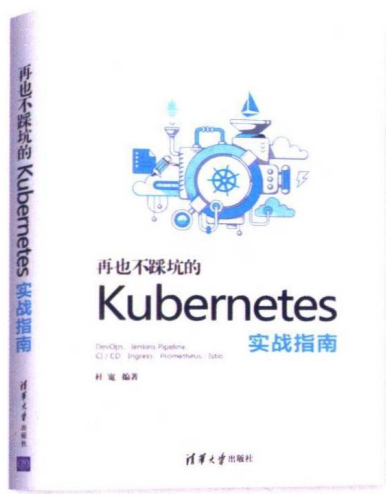
产品编号：081781-01

作 / 者 / 简 / 介

张伟洋

毕业于中国地质大学计算机科学与技术专业，先后就职于知名互联网公司百度、慧聪网，任Java高级软件工程师，互联网旅游公司任软件研发事业部技术经理。目前供职于青岛英谷教育科技股份有限公司，任大数据项目讲师，为数十所高校先后举行多次大数据专题讲座，对Hadoop及周边框架ZooKeeper、Hive、HBase、Storm、Spark等有深入的研究。高等院校云计算与大数据专业课改教材《云计算与大数据概论》《大数据开发与应用》的主要编写者，百度文库、百度阅读签约作者。

图/书/推/荐



本书以Kubernetes实战为主线，从企业应用场景出发，介绍了Kubernetes的架构模式、集群安装、组件、应用的容器化、Jenkins的持续集成和持续部署以及服务网格Service Mesh等内容，可帮助读者从零开始快速搭建容器集群，并应用于公司业务实践。

前　言

当今互联网已进入大数据时代，大数据技术已广泛应用于金融、医疗、教育、电信、政府等领域。各行各业每天都在产生大量的数据，数据计量单位已从 B、KB、MB、GB、TB 发展到 PB、EB、ZB、YB 甚至 BB、NB、DB。预计未来几年，全球数据将呈爆炸式增长。谷歌、阿里巴巴、百度、京东等互联网公司都急需掌握大数据技术的人才，而大数据相关人才却出现了供不应求的状况。

Hadoop 作为大数据生态系统中的核心框架，专为离线和大规模数据处理而设计。Hadoop 的核心组成 HDFS 为海量数据提供了分布式存储；MapReduce 则为海量数据提供了分布式计算。很多互联网公司都使用 Hadoop 来实现公司的核心业务，例如华为的云计算平台、淘宝的推荐系统等，只要和海量数据相关的领域都有 Hadoop 的身影。

本书作为 Hadoop 及其周边框架的入门书，知识面比较广，涵盖了当前整个 Hadoop 生态系统主流的大数据开发技术。内容全面，代码可读性强，以实操为主，理论为辅，一步一步手把手对常用的离线计算以及实时计算等系统进行了深入讲解。

全书共 16 章，第 1 章讲解了 VMware 中 CentOS 7 操作系统的安装；第 2 章讲解了大数据开发之前对操作系统集群环境的配置；第 3~16 章讲解了 Hadoop 生态系统各框架 HDFS、MapReduce、YARN、ZooKeeper、HBase、Hive、Sqoop 和数据实时处理系统 Flume、Kafka、Storm、Spark 以及分布式搜索系统 Elasticsearch 等的基础知识、架构原理、集群环境搭建，同时包括常用的 Shell 命令、API 操作、源码剖析，并通过实际案例加深对各个框架的理解与应用。

那么如何学习本书呢？

本书推荐的阅读方式是按照章节顺序从头到尾完成阅读，因为后面的很多章节是以前面的章节为基础，而且这种一步一个脚印、由浅入深的方式将使你更加顺利地掌握大数据的开发技能。

学习本书时，首先根据第 1、2 章搭建好开发环境，然后依次学习第 3~16 章，学习每一章时先了解该章的基础知识和框架的架构原理，然后再进行集群环境搭建、Shell 命令操作等实操练习，这样学习效果会更好。当书中的理论和实操知识都掌握后，可以进行举一反三，自己开发一个大数据程序，或者将所学知识运用到自己的编程项目上，也可以到各种在线论坛与其他大数据爱好者进行讨论，互帮互助。

本书可作为 Hadoop 新手入门的指导书籍或者大数据开发人员的参考用书，要求读者具备一定的 Java 语言基础和 Linux 系统基础，即使没有任何大数据基础的读者，也可以对照书中的步骤成

功搭建属于自己的大数据集群，是一本真正的提高读者动手能力、以实操为主的入门书籍。通过对本书的学习，读者能够对大数据相关框架迅速理解并掌握，可以熟练使用 Hadoop 集成环境进行大数据项目的开发。

读者若对书中讲解的知识有任何疑问，可关注下面的公众号联系笔者，还可以在该公众号中获取大数据相关的学习教程和资源。

扫描下述二维码可以下载本书源代码：

由于时间原因，书中难免出现一些错误或不准确的地方，恳请读者批评指正。

张伟洋

2019 年 5 月于青岛

目　录

第 1 章　VMware 中安装 CentOS 7 1

1.1　下载 CentOS 7 镜像文件 1
1.2　新建虚拟机 5
1.3　安装操作系统 9

第 2 章　CentOS 7 集群环境配置 16

2.1　系统环境配置 16
2.1.1　新建用户 17
2.1.2　修改用户权限 17
2.1.3　关闭防火墙 17
2.1.4　设置固定 IP 18
2.1.5　修改主机名 22
2.1.6　新建资源目录 23
2.2　安装 JDK 23
2.3　克隆虚拟机 25
2.4　配置主机 IP 映射 29

第 3 章　Hadoop 31

3.1　Hadoop 简介 31
3.1.1　Hadoop 生态系统架构 32
3.1.2　Hadoop 1.x 与 2.x 的架构对比 33
3.2　YARN 基本架构及组件 34
3.3　YARN 工作流程 37
3.4　配置集群各节点 SSH 无密钥登录 38
3.4.1　无密钥登录原理 38
3.4.2　无密钥登录操作步骤 39
3.5　搭建 Hadoop 2.x 分布式集群 41

第 4 章　HDFS 48

4.1　HDFS 简介 48
4.1.1　设计目标 49

4.1.2 总体架构 49
4.1.3 主要组件 50
4.1.4 文件读写 53
4.2 HDFS 命令行操作 54
4.3 HDFS Web 界面操作 57
4.4 HDFS Java API 操作 59
4.4.1 读取数据 59
4.4.2 创建目录 61
4.4.3 创建文件 62
4.4.4 删除文件 63
4.4.5 遍历文件和目录 64
4.4.6 获取文件或目录的元数据 65
4.4.7 上传本地文件 66
4.4.8 下载文件到本地 66

第 5 章 MapReduce 68

5.1 MapReduce 简介 68
5.1.1 设计思想 69
5.1.2 任务流程 70
5.1.3 工作原理 71
5.2 MapReduce 程序编写步骤 74
5.3 案例分析：单词计数 76
5.4 案例分析：数据去重 82
5.5 案例分析：求平均分 86
5.6 案例分析：二次排序 89
5.7 使用 MRUnit 测试 MapReduce 程序 97

第 6 章 ZooKeeper 100

6.1 ZooKeeper 简介 100
6.1.1 应用场景 101
6.1.2 架构原理 101
6.1.3 数据模型 102
6.1.4 节点类型 103
6.1.5 Watcher 机制 103
6.1.6 分布式锁 105
6.2 ZooKeeper 安装配置 106
6.2.1 单机模式 106
6.2.2 伪分布模式 108
6.2.3 集群模式 109

6.3 ZooKeeper 命令行操作 112
6.4 ZooKeeper Java API 操作 114
6.4.1 创建 Java 工程 114
6.4.2 创建节点 115
6.4.3 修改数据 118
6.4.4 获取数据 118
6.4.5 删除节点 123
6.5 案例分析：监听服务器动态上下线 124

第 7 章 HDFS 与 YARN HA 129

7.1 HDFS HA 搭建 129
7.1.1 架构原理 130
7.1.2 搭建步骤 131
7.1.3 结合 ZooKeeper 进行 HDFS 自动故障转移 137
7.2 YARN HA 搭建 142
7.2.1 架构原理 142
7.2.2 搭建步骤 142

第 8 章 HBase 147

8.1 什么是 HBase 147
8.2 HBase 基本结构 148
8.3 HBase 数据模型 149
8.4 HBase 集群架构 151
8.5 HBase 安装配置 153
8.5.1 单机模式 153
8.5.2 伪分布模式 155
8.5.3 集群模式 156
8.6 HBase Shell 命令操作 160
8.7 HBase Java API 操作 164
8.7.1 创建 Java 工程 164
8.7.2 创建表 164
8.7.3 添加数据 166
8.7.4 查询数据 168
8.7.5 删除数据 169
8.8 HBase 过滤器 170
8.9 案例分析：HBase MapReduce 数据转移 174
8.9.1 HBase 不同表间数据转移 174
8.9.2 HDFS 数据转移至 HBase 180
8.10 案例分析：HBase 数据备份与恢复 183

第 9 章 Hive 185
9.1 什么是 Hive 185
9.1.1 数据单元 186
9.1.2 数据类型 187
9.2 Hive 架构体系 189
9.3 Hive 三种运行模式 190
9.4 Hive 安装配置 191
9.4.1 内嵌模式 192
9.4.2 本地模式 195
9.4.3 远程模式 198
9.5 Hive 常见属性配置 200
9.6 Beeline CLI 的使用 201
9.7 Hive 数据库操作 205
9.8 Hive 表操作 208
9.8.1 内部表 209
9.8.2 外部表 213
9.8.3 分区表 215
9.8.4 分桶表 219
9.9 Hive 查询 223
9.9.1 SELECT 子句查询 224
9.9.2 JOIN 连接查询 230
9.10 其他 Hive 命令 233
9.11 Hive 元数据表结构分析 235
9.12 Hive 自定义函数 237
9.13 Hive JDBC 操作 239
9.14 案例分析：Hive 与 HBase 整合 242
9.15 案例分析：Hive 分析搜狗用户搜索日志 246
第 10 章 Sqoop 251
10.1 什么是 Sqoop 251
10.1.1 Sqoop 基本架构 252
10.1.2 Sqoop 开发流程 252
10.2 使用 Sqoop 253
10.3 数据导入工具 254
10.4 数据导出工具 259
10.5 Sqoop 安装与配置 261
10.6 案例分析：将 MySQL 表数据导入到 HDFS 中 262
10.7 案例分析：将 HDFS 中的数据导出到 MySQL 中 263
10.8 案例分析：将 MySQL 表数据导入到 HBase 中 264

第 11 章 Kafka 267

11.1 什么是 Kafka 267
11.2 Kafka 架构 268
11.3 主题与分区 269
11.4 分区副本 271
11.5 消费者组 273
11.6 数据存储机制 274
11.7 集群环境搭建 276
11.8 命令行操作 278
11.8.1 创建主题 278
11.8.2 查询主题 279
11.8.3 创建生产者 280
11.8.4 创建消费者 280
11.9 Java API 操作 281
11.9.1 创建 Java 工程 281
11.9.2 创建生产者 281
11.9.3 创建消费者 283
11.9.4 运行程序 285
11.10 案例分析：Kafka 生产者拦截器 287

第 12 章 Flume 294

12.1 什么是 Flume 294
12.2 架构原理 295
12.2.1 单节点架构 295
12.2.2 组件介绍 296
12.2.3 多节点架构 297
12.3 安装与简单使用 299
12.4 案例分析：日志监控（一） 302
12.5 案例分析：日志监控（二） 304
12.6 拦截器 306
12.6.1 内置拦截器 307
12.6.2 自定义拦截器 310
12.7 选择器 313
12.8 案例分析：拦截器和选择器的应用 315
12.9 案例分析：Flume 与 Kafka 整合 319

第 13 章 Storm 322

13.1 什么是 Storm 322

13.2 Storm Topology 323
13.3 Storm 集群架构 324
13.4 Storm 流分组 326
13.5 Storm 集群环境搭建 329
13.6 案例分析：单词计数 332
13.6.1 设计思路 332
13.6.2 代码编写 333
13.6.3 程序运行 339
13.7 案例分析：Storm 与 Kafka 整合 341

第 14 章 Elasticsearch 347

14.1 什么是 Elasticsearch 347
14.2 基本概念 348
14.2.1 索引、类型和文档 348
14.2.2 分片和副本 348
14.2.3 路由 349
14.3 集群架构 350
14.4 集群环境搭建 352
14.5 Kibana 安装 355
14.6 REST API 357
14.6.1 集群状态 API 357
14.6.2 索引 API 358
14.6.3 文档 API 360
14.6.4 搜索 API 363
14.6.5 Query DSL 365
14.7 Head 插件安装 371
14.8 Java API 操作：员工信息 375

第 15 章 Scala 379

15.1 什么是 Scala 379
15.2 安装 Scala 380
15.2.1 Windows 中安装 Scala 380
15.2.2 CentOS 7 中安装 Scala 381
15.3 Scala 基础 382
15.3.1 变量声明 382
15.3.2 数据类型 383
15.3.3 表达式 385
15.3.4 循环 386
15.3.5 方法与函数 388

15.4 集合 391
15.4.1 数组 391
15.4.2 List 393
15.4.3 Map 映射 394
15.4.4 元组 396
15.4.5 Set 396
15.5 类和对象 398
15.5.1 类的定义 398
15.5.2 单例对象 399
15.5.3 伴生对象 399
15.5.4 get 和 set 方法 400
15.5.5 构造器 402
15.6 抽象类和特质 404
15.6.1 抽象类 404
15.6.2 特质 406
15.7 使用 Eclipse 创建 Scala 项目 408
15.7.1 安装 Scala for Eclipse IDE 408
15.7.2 创建 Scala 项目 409
15.8 使用 IntelliJ IDEA 创建 Scala 项目 410
15.8.1 IDEA 中安装 Scala 插件 410
15.8.2 创建 Scala 项目 414

第 16 章 Spark 416

16.1 Spark 概述 416
16.2 Spark 主要组件 417
16.3 Spark 运行时架构 419
16.3.1 Spark Standalone 模式 419
16.3.2 Spark On YARN 模式 421
16.4 Spark 集群环境搭建 423
16.4.1 Spark Standalone 模式 423
16.4.2 Spark On YARN 模式 425
16.5 Spark HA 搭建 426
16.6 Spark 应用程序的提交 430
16.7 Spark Shell 的使用 433
16.8 Spark RDD 435
16.8.1 创建 RDD 435
16.8.2 RDD 算子 436
16.9 案例分析：使用 Spark RDD 实现单词计数 441
16.10 Spark SQL 448

16.10.1 DataFrame 和 Dataset 448
16.10.2 Spark SQL 基本使用 449
16.11 案例分析：使用 Spark SQL 实现单词计数 452
16.12 案例分析：Spark SQL 与 Hive 整合 454
16.13 案例分析：Spark SQL 读写 MySQL 457

第 1 章

VMware 中安装 CentOS 7

本章内容

本章讲解在 VMware Workstation（以下简称 VMware）中安装 CentOS 操作系统的步骤。使用的 VMware 版本为 12.5.2，CentOS 操作系统的版本为 7.3（1611）。

本章目标

- 了解 CentOS 7 操作系统的下载
- 掌握 VMware 中虚拟机的创建步骤
- 掌握 CentOS 7 操作系统的安装步骤

1.1 下载 CentOS 7 镜像文件

请参考下述安装步骤。

步骤01 在浏览器中输入网址 https://www.centos.org/，进入 CentOS 官网，单击官网主页面中的【Get CentOS Now】按钮，如图 1-1 所示。

步骤02 在出现的下载页面中单击【DVD ISO】按钮，可进入目前 CentOS 操作系统最新版的下载链接页面，如图 1-2 所示。

步骤03 若想下载 CentOS 操作系统的历史版本，可以在浏览器中访问网址 http://vault.centos.org/，然后单击对应的版本所在的文件夹，此处选择【7.3.1611/】，如图 1-3 所示。

图 1-1　CentOS 官网主页面

图 1-2　CentOS 操作系统下载页面（下载最新版本）

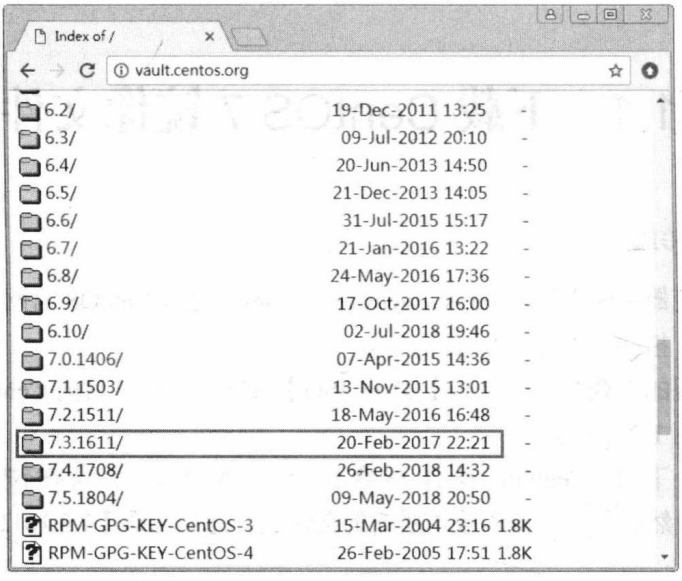


图 1-3　CentOS 操作系统下载页面（下载历史版本）