博客园资深博主、极客学院特邀讲师分享多年的Hadoop使用经验

全面涵盖了Hadoop从基础部署到集群管理，再到底层设计等重点内容

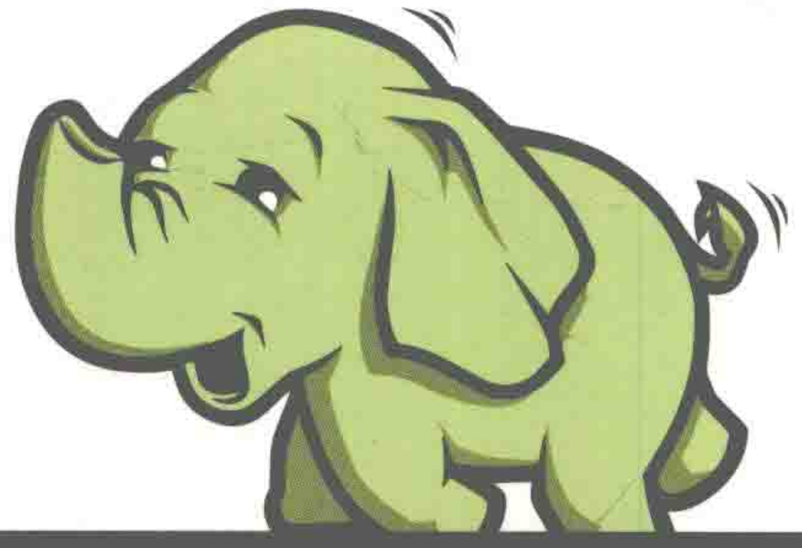
深度剖析Kafka开源监控工具Kafka Eagle的设计和架构思想

Hadoop大数据挖掘

从入门到进阶实战

（视频教学版）

邓杰◎编著



- 提供了近200分钟配套教学视频，手把手带领读者高效学习
- 详解51个实例和10个综合案例，带领读者通过实际动手提高编程水平
- 书中的所有实例和案例均来源于作者多年的工作经验积累和技术分享
- 给出了大量的“避坑”技巧，让读者在实际开发中少走弯路
- 用浅显易懂的语言进行讲解，读者阅读时不会有云山雾罩的感觉

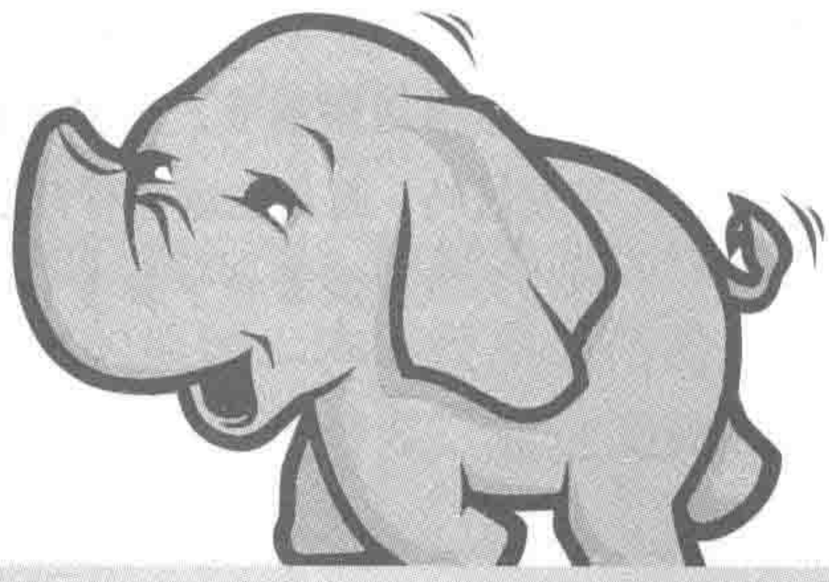
机械工业出版社
China Machine Press

Hadoop大数据挖掘

从入门到进阶实战

（视频教学版）

邓杰◎编著



Hadoop

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Hadoop大数据挖掘从入门到进阶实战：视频教学版/邓杰编著. —北京：机械工业出版社，2018.6

ISBN 978-7-111-60010-7

Ⅰ. H… Ⅱ. 邓… Ⅲ. 数据处理 Ⅳ. TP274

中国版本图书馆CIP数据核字（2018）第107840号

本书采用“理论+实战”的形式编写，全面介绍了Hadoop大数据挖掘的相关知识。本书秉承循序渐进、易于理解、学以致用和便于查询的讲授理念，讲解时结合了大量实例和作者多年积累的一线开发经验。本书作者拥有丰富的视频制作与在线教学经验，曾经与极客学院合作开设过在线视频教学课程。为了帮助读者高效、直观地学习本书内容，作者特意为本书录制了配套教学视频，这些教学视频和本书配套源代码文件读者都可以免费获取。

本书共分为13章，涵盖的主要内容有：集群及开发环境搭建；快速构建一个Hadoop项目并线上运行；Hadoop套件实战；Hive编程——使用SQL提交MapReduce任务到Hadoop集群；游戏玩家的用户行为分析——特征提取；Hadoop平台管理与维护；Hadoop异常处理解决方案；初识Hadoop核心源码；Hadoop通信机制和内部协议；Hadoop分布式文件系统剖析；ELK实战案例——游戏应用实时日志分析平台；Kafka实战案例——实时处理游戏用户数据；Hadoop拓展——Kafka剖析。

本书通俗易懂，案例丰富，实用性强，不但适合初学者系统学习Hadoop的各种基础语法和开发技巧，而且也适合有开发经验的程序员进阶提高。另外，本书还适合社会培训机构和相关院校作为教材或者教学参考书。

Hadoop 大数据挖掘从入门到进阶实战（视频教学版）

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：欧振旭　李华君　　责任校对：姚志娟
印　　刷：中国电影出版社印刷厂　　版　　次：2018 年 6 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：26
书　　号：ISBN 978-7-111-60010-7　　定　　价：99.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

前言

大数据时代，数据的存储与挖掘至关重要。企业在追求高可靠性、高扩展性及高容错性的大数据处理平台的同时还希望能够降低成本，而 Hadoop 为实现这些需求提供了解决方案。

Hadoop 在分布式计算与存储上具有先天优势。它作为 Apache 软件基金会的顶级开源项目，其版本迭代持续至今，而且已经拥有一个非常活跃的社区和全球众多开发者，并且成为了当前非常流行的大数据处理平台。很多公司，特别是互联网公司，都纷纷开始使用或者已经使用 Hadoop 来做海量数据存储与数据挖掘。

Hadoop 简单易学，其学习曲线平缓且学习周期短。它的操作命令和 Linux 命令非常相似。一个熟悉 Linux 的开发者只需要短短的一周时间，就可以学会 Hadoop 开发，完成一个高可用集群的部署和高可用应用程序的编写。

面对 Hadoop 的普及和学习热潮，笔者愿意分享自己多年的开发经验，带领读者比较轻松地掌握 Hadoop 数据挖掘的相关知识。这便是笔者编写本书的原因。本书使用通俗易懂的语言进行讲解，从基础部署到集群的管理，再到底层设计等内容均有涉及。通过阅读本书，读者可以较为轻松地掌握 Hadoop 大数据挖掘与分析的相关技术。

本书特色

1. 提供专业的配套教学视频，高效、直观

笔者曾接受过极客学院的专业视频制作指导，并在极客学院录制过多期 Hadoop 和 Kafka 实战教学视频课程，得到了众多学习者的青睐及好评。为了便于读者更加高效、直观地学习本书内容，笔者特意为本书实战部分的内容录制了配套教学视频，读者可以在教学视频的辅助下学习，从而更加轻松地掌握 Hadoop。

2. 分享大量来自一线的开发经验，贴近实际开发

本书给出的代码讲解和实例大多数来自于笔者多年的教学积累和技术分享，几乎都是得到了学习者一致好评的干货。另外，笔者还是一名开源爱好者，编写了业内著名的 Kafka

Eagle 监控系统。本书第 13 章介绍了该系统的使用，以帮助读者掌握如何监控大数据集群的相关知识。

3．分享多个来自一线的实例，有很强的实用性

本书精心挑选了多个实用性很强的例子，如 Hadoop 套件实战、Hive 编程、Hadoop 平台管理与维护、ELK 实战和 Kafka 实战等。读者不但可以从这些例子中学习和理解 Hadoop 及其套件的相关知识点，而且还可以将这些例子应用于实际开发中。

4．讲解通俗易懂，力争触类旁通，举一反三

本书用通俗易懂的语言讲解，避免“云山雾罩”，让读者不知所云。书中在讲解一些常用知识点时将 Hadoop 命令与 Linux 命令进行了对比，便于熟悉 Linux 命令的读者能够迅速掌握 Hadoop 的操作命令。

本书内容

第1章　集群及开发环境搭建

本章介绍的主要内容包括：环境准备；安装 Hadoop；演示 Hadoop 版 Hello World 示例程序，以及搭建 Hadoop 开发环境。

第2章　实战：快速构建一个Hadoop项目并线上运行

本章首先介绍了快速构建项目工程的方法，如 Maven 和 Java Project；然后介绍了分布式文件系统的操作命令，以及利用 IDE 提交 MapReduce 作业的相关知识；最后介绍了编译应用程序并打包，以及部署与调度等内容。

第3章　Hadoop套件实战

本章介绍了 Hadoop 生态圈中常见的大数据套件的背景知识和使用方法，涵盖 Sqoop、Flume、HBase、Zeppelin、Drill 及 Spark 等套件。

第4章　Hive编程——使用SQL提交MapReduce任务到Hadoop集群

本章主要介绍了 Hive 数据仓库的相关内容：Hive 底层设计组成；安装和配置 Hive；基于 Hive 应用接口进行编程；开源监控工具 Hive Cube。

第5章　游戏玩家的用户行为分析——特征提取

本章首先对 Hadoop 的基础知识进行了梳理；然后介绍了项目的背景和平台架构；接着对项目进行了整体分析与指标设计，并进行了技术选型；最后对分析的指标进行了

编码实践。

第6章　Hadoop平台管理与维护

本章介绍了 Hadoop 平台管理与维护的重要方法。本章首先介绍了 Hadoop 分布式文件系统的特性，然后介绍了 HDFS 的基础命令，并对 NameNode 进行了解读。另外，本章对 Hadoop 平台维护时的常规操作，如节点管理、HDFS 快照和安全模式等内容也进行了讲解。

第7章　Hadoop异常处理解决方案

本章介绍了 Hadoop 异常处理解决方案的几个知识点。主要内容包括：跟踪日志；分析异常信息；利用搜索引擎检索关键字；查看 Hadoop JIRA；阅读 Hadoop 源代码。

本章最后以实战案例的形式分析了几种异常情况：启动 HBase 集群失败；HBase 表查询失败；Spark 的临时数据不自动清理等。

第8章　初识Hadoop核心源码

本章首先介绍了 Hadoop 源码基础环境准备及源代码编译；接着介绍了 Hadoop 的起源和两代 MapReduce 框架间的差异；最后介绍了 Hadoop 的序列化机制。

第9章　Hadoop通信机制和内部协议

本章首先介绍了 Hadoop 通信模型和 Hadoop RPC 的特点；然后通过编码实践介绍了 Hadoop RPC 的使用，同时还介绍了与之类似的开源 RPC 框架；最后介绍了 MapReduce 的通信协议和 RPC 协议的实现过程。

第10章　Hadoop分布式文件系统剖析

本章主要介绍了 Hadoop 分布式文件系统的设计特点、命令空间和节点、数据备份策略等内容，最后以实战的形式演示了跨平台数据迁移的过程。

第11章　ELK实战案例——游戏应用实时日志分析平台

本章介绍了常用的 ELK 套件：Logstash——实时日志采集、分析和传输；Elasticsearch——分布式存储及搜索引擎；Kibana——可视化管理系统。

第12章　Kafka实战案例——实时处理游戏用户数据

本章首先介绍了 Kafka 项目的背景，以及 Kafka 集群和 Storm 集群的安装过程；然后对项目案例进行了分析与指标设计，并利用笔者多年的大数据开发经验设计项目体系架构；最后演示了各个模块的编码实现，如生产模块、消费模块、数据持久化实现及应用调度实现等。

第13章　Hadoop拓展——Kafka剖析

本章主要介绍了 Kafka 的基本特性与结构，以及笔者设计并开发的开源 Kafka 监控工具 Kafka Eagle。本章关键知识点包括：Kafka 开发与维护；开源监控工具 Kafka Eagle 的使用；Kafka 源代码分析，如分布式选举算法剖析、Kafka Offset 解读、Kafka 存储机制和副本剖析等。

本书配套学习资源

本书提供了配套教学视频和实例源代码文件等超值资源。请在机械工业出版社华章公司的网站 www.hzbook.com 上搜索到本书页面，然后在“资料下载”模块下载这些学习资源。

本书读者对象

- Hadoop 初学者；
- Hadoop 进阶人员；
- 后端程序初学者；
- 前端转后端的开发人员；
- 熟悉 Linux 和 Java 而需要学习 Hadoop 的编程爱好者；
- 想用 Hadoop 快速编写海量数据处理程序的开发者；
- 相关培训机构的学员和高等院校的学生。

致谢

感谢我的女朋友邹苗苗对我生活上的细心照顾与琐事上的宽容，使得我能安心写作！感谢我的父母对我的养育之恩！

感谢刘旨阳、王珏辉、贺祥、张翠菊等人（排名不分先后）在写作本书时给我提供的各种帮助！

另外，本书的编写得到了吴宏伟先生的大力帮助。他对本书的写作提出了很多有益建议，并对内容做了细致入微的审核，这使得本书条理更为清晰，语言更加通俗易懂。在此深表感谢！

最后感谢各位读者选择了本书！希望本书能对您的学习有所助益。

虽然笔者对书中所述内容都尽量核实，并多次进行文字校对，但因时间有限，加之水平所限，书中可能还存在疏漏和错误，敬请广大读者批评指正。联系邮件：hzbook2017@163.com 和 whw010@163.com。

邓杰

目录

第 1 章　集群及开发环境搭建

工欲善其事，必先利其器。在学习和研究一门技术之前，需要做一些必要的准备，比如搭建和使用 Hadoop 集群。由于 Hadoop 是一个分布式系统，具有相当程度的复杂性，所以对于 Hadoop 相关的项目开发，仅仅掌握以上知识是远远不够的。在笔者看来还需要掌握 Hadoop 生态圈中其他套件的集成与使用，这样才能在 Hadoop 项目开发中游刃有余。

本章的知识都是 Hadoop 基础，学习起来会非常轻松。本章将介绍如何搭建一个高可用的 Hadoop 集群，内容包含 Hadoop 2.7 版本的安装、Zookeeper 套件的集成和 Hadoop 应用程序的运行等。

本书的所有演示环境都是基于分布式环境进行的，所以本章内容也是基于分布式环境基础上的。

1.1　环境准备

现如今，大部分企业在测试和生产环境中所使用的服务器操作系统均是基于 Linux 操作系统。考虑到 Linux 操作系统的市场占有率，Hadoop 设计之初便是以 Linux 操作系统为前提，因而其在 Linux 操作系统中具有完美的支持。

Hadoop 的源代码是基于 Java 语言编写的。虽然 Java 语言具有跨平台的特性，但由于 Hadoop 的部分功能对 Linux 操作系统有一定依赖，因而 Hadoop 对其他平台（如 Windows 操作系统）的兼容性不是很好。

本节以 64 bit CentOS（Community Enterprise Operating System，社区企业操作系统，是 Linux 发行版之一）的 6.6 版本为例，介绍如何在 Linux 操作系统下完成基础软件的配置与部署。

1.1.1　基础软件下载

由于 Hadoop 采用的开发语言是 Java 编程语言，所以搭建 Hadoop 集群的首要任务是先安装 Java 语言的基础开发包 Java Development Kit（简称 JDK）。

本书选择的 JDK 版本是 Oracle 官方的 JDK 8，版本号为 8u144，如图 1-1 所示。

这里选择 rpm 安装包和 tar.gz 安装包均可。本书选择的是 x64.tar.gz 安装包，版本信息与下载地地址如表 1-1 所示。

Java SE Development Kit 8u144

You must accept the Oracle Binary Code License Agreement for Java SE to download this software.

Thank you for accepting the Oracle Binary Code License Agreement for Java SE; you may now download this software.

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	77.89 MB	jdk-8u144-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	74.83 MB	jdk-8u144-linux-arm64-vfp-hflt.tar.gz
Linux x86	164.65 MB	jdk-8u144-linux-i586.rpm
Linux x86	179.44 MB	jdk-8u144-linux-i586.tar.gz
Linux x64	162.1 MB	jdk-8u144-linux-x64.rpm
Linux x64	176.92 MB	jdk-8u144-linux-x64.tar.gz
Mac OS X	226.6 MB	jdk-8u144-macosx-x64.dmg
Solaris SPARC 64-bit	139.87 MB	jdk-8u144-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	99.18 MB	jdk-8u144-solaris-sparcv9.tar.gz
Solaris x64	140.51 MB	jdk-8u144-solaris-x64.tar.Z
Solaris x64	96.99 MB	jdk-8u144-solaris-x64.tar.gz
Windows x86	190.94 MB	jdk-8u144-windows-i586.exe
Windows x64	197.78 MB	jdk-8u144-windows-x64.exe

图 1-1　JDK 下载预览

表 1-1　版本信息与下载地址

软　　件	下载地址	推荐版本
JDK	http://www.oracle.com/technetwork/java/javase/downloads/index.html	1.8
CentOS	https://www.centos.org/download/	6.6

1.1.2　准备 Linux 操作系统

本节主要讲述 Linux 操作系统的选择，以及 JDK 环境的安装配置。

如今，市场上 Linux 操作系统的版本有很多，如 RedHat、Ubuntu、CentOS 等。本书选择的操作系统是基于 64bit CentOS 6.6 类型的，读者可以根据自己的喜好选取合适的 Linux 操作系统，这个对学习本书的影响不大。CentOS 6.6 安装包下载预览图，如图 1-2 所示。这里选择 64 bit CentOS 6.6 的镜像文件进行下载。

CentOS Linux 6

Release	Based on RHEL Source (Version)	Archived Tree
6.9	6.9	Tree
6.8	6.8	Tree
6.7	6.7	Tree
6.6	6.6	Tree
6.5	6.5	Tree
6.4	6.4	Tree
6.3	6.3	Tree
6.2	6.2	Tree
6.1	6.1	Tree
6.0	6.0	Tree

图 1-2　CentOS 下载预览

1. 安装配置JDK

由于 CentOS 操作系统会自带 OpenJDK 环境，在安装 Oracle 官网下载的 JDK 版本之前，需要先检查 CentOS 操作系统中是否存在 OpenJDK 环境，如果存在，则需先行卸载自带的 JDK 环境，具体步骤如下：

（1）卸载 CentOS 操作系统自带的 JDK 环境（如果系统自带的 JDK 环境不存在，可跳过此步骤）。

```
# 查找 Java 安装依赖库
[hadoop@nna ~]$ rpm -qa | grep Java
# 卸载 Java 依赖库
[hadoop@nna ~]$ yum -y remove Java*
```

（2）将下载的 JDK 安装包解压缩到指定目录下（可自行指定），详细操作命令如下：

```
# 解压 JDK 安装包到当前目录
[hadoop@nna ~]$ tar -zxvf jdk-8u144-linux-x64.tar.gz
# 移动 JDK 到/data/soft/new 目录下，并改名为 jdk
[hadoop@nna ~]$ mv jdk-8u144-linux-x64 /data/soft/new/jdk
```

（3）编辑环境变量，具体操作命令如下：

```
# 打开全局环境变量配置文件
[hadoop@nna ~]$ vi /etc/profile
# 添加具体内容如下
export JAVA_HOME=/data/soft/new/jdk
export $PATH:$JAVA_HOME/bin
# 编辑完成后保存并退出
```

（4）保存刚刚编辑完成后的文件，若要配置的内容立即生效，则执行如下命令：

```
# 使用 source 或者英文点(.)命令，立即生效配置文件
[hadoop@nna ~]$ source /etc/profile
```

（5）验证安装的 JDK 环境是否成功，具体操作命令如下：

```
# Java 语言版本验证命令
[hadoop@nna ~]$ java -version
```

如果操作终端上显示对应的 JDK 版本号，即可认为 JDK 环境配置成功。

2. 同步安装包

将该节点（这里将每台服务器称为“Hadoop 集群中的某一个节点”，后续章节中都以“节点”来称呼服务器）上的 JDK 安装包，使用 Linux 同步命令传输到其他节点上，具体操作命令如下：

```
# Linux 传输命令（将 nna 节点的 JDK 文件复制到 nns 节点）
[hadoop@nna ~]$ scp -r /data/soft/new/jdk hadoop@nns:/data/soft/new
```

1.2 安装 Hadoop

由于本书通篇都是基于分布式环境来演示和讲解的，所以安装 Hadoop 集群需要准备至少 5 台虚拟机或者物理机。

本节将介绍如何在 Linux 系统环境下搭建基础的 Hadoop 集群，其内容包含基础环境的配置、Hadoop 集群核心文件的配置及验证集群搭建成功后的效果等。

1.2.1 基础环境配置

在搭建 Hadoop 集群之前，需要确保 5 台基于 Linux 系统的节点已准备就绪。本书的 Hadoop 集群的硬件配置如表 1-2 所示。

表 1-2 Hadoop集群硬件配置表

主 机 名	角 色	内 存	CPU
nna	NameNode Active	2GB	2核
nns	NameNode Standby	2GB	2核
dn1	DataNode	1GB	1核
dn2	DataNode	1GB	1核
dn3	DataNode	1GB	1核

提示： 由于是在学习阶段，在硬盘选择方面可以不用纠结 TB 级别容量或是 PB 级别容量，一般来说，1TB 的硬盘容量足够我们学习完本书的所有内容了。

1. 创建Hadoop账号

在搭建 Hadoop 集群环境时，不推荐使用 root 账号来操作，可以创建一个 Hadoop 账号用来专门管理集群环境。

创建 Hadoop 账号的过程均在 root 账号下完成，具体创建命令如下：

```
# 创建名为 hadoop 的账号
[root@nna ~]# useradd hadoop
# 给名为 hadoop 的账号设置密码
[root@nna ~]# passwd hadoop
```

然后根据系统提示，设置账号登录密码；接着给 Hadoop 账号设置免密码登录权限，当然也可以自行添加其他权限，操作命令内容如下：

```
# 给 sudoers 文件赋予写权限
[root@nna ~]# chmod +w /etc/sudoers
```