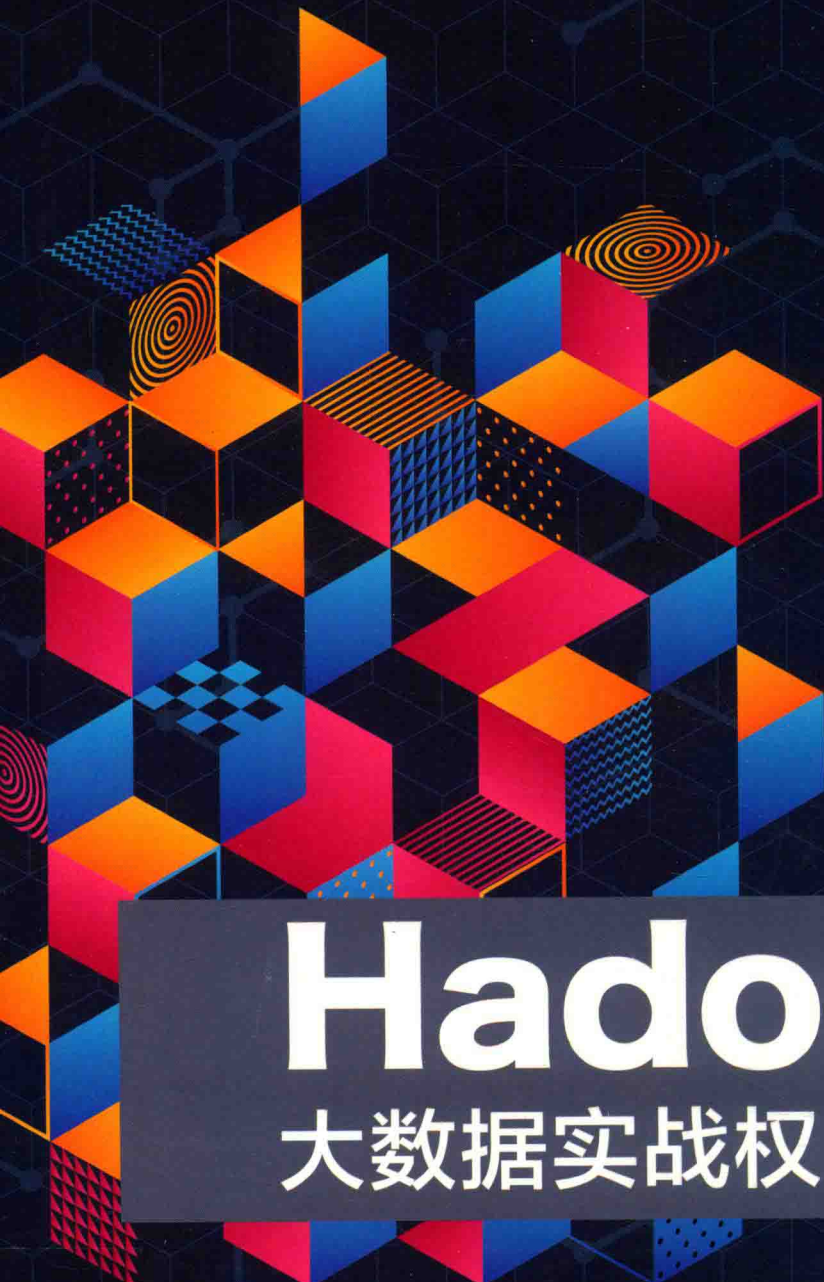


大数据科学与应用丛书



- 深入分析组件原理 ·
- 充分展示搭建过程 ·
- 详细指导应用开发 ·

Hadoop

大数据实战权威指南

黄东军 编著

 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
www.phei.com.cn

大数据科学与应用丛书



Hadoop

大数据实战权威指南

黄东军 编著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

大数据贵在落实!

本书是一本讲解大数据实战的图书,按照“深入分析组件原理、充分展示搭建过程、详细指导应用开发”编写。全书分为三篇,第一篇为大数据的基本概念和技术,主要介绍大数据的背景、发展及关键技术;第二篇为 Hadoop 大数据平台搭建与基本应用,内容涉及 Linux、HDFS、MapReduce、YARN、Hive、HBase、Sqoop、Kafk、Spark 等;第三篇为大数据处理与项目开发,包括交互式数据处理、协同过滤推荐系统、销售数据分析系统,并就京东的部分销售数据应用大数据进行处理分析。

本书适合初学者入门和进阶,也可供希望全面、系统地理解并掌握大数据实际应用的读者参考,对从事大数据项目开发的专业人员也有参考价值。

为了方便读者实践,本书配有开发资源包,读者可登录华信教育资源网(www.hxedu.com.cn)免费注册后下载。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

Hadoop 大数据实战权威指南 / 黄东军编著. — 北京: 电子工业出版社, 2017.7

(大数据科学与应用丛书)

ISBN 978-7-121-31821-4

I. ①H… II. ①黄… III. ①数据处理软件—指南 IV. ①TP274-62

中国版本图书馆 CIP 数据核字(2017)第 129534 号

责任编辑: 田宏峰

印 刷: 三河市鑫金马印装有限公司

装 订: 三河市鑫金马印装有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编: 100036

开 本: 787×980 1/16 印张: 23.5 字数: 526 千字

版 次: 2017 年 7 月第 1 版

印 次: 2017 年 7 月第 1 次印刷

定 价: 68.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888, 88258888。

质量投诉请发邮件至 zltz@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式: tianhf@phei.com.cn。

前 言

本书内容

本书分为三篇，共有 12 章。

第一篇 大数据的基本概念和技术

第 1 章 绪论，描述大数据的时代背景与国家大数据战略，探讨大数据的概念和特性，重点阐述大数据支撑体系，包括数据采集、存储、分布式计算和应用，并讨论大数据人才特点与能力要求。

第 2 章 Hadoop 大数据关键技术，详细介绍大数据系统涉及的主流技术，主要包括数据采集与生成、数据分布式存储、分布式计算框架、数据分析与挖掘等方面的技术和工具。

第二篇 Hadoop 大数据平台搭建与基本应用

第 3 章 Linux 操作系统与集群搭建，介绍 Linux 集群的安装、Java 开发包 JDK 的安装，以及集群的配置方法。

第 4 章 HDFS 安装与基本应用，介绍 Hadoop HDFS 的架构、工作原理，以及 Hadoop 安装、配置、启动和程序的运行。

第 5 章 MapReduce 与 YARN，介绍 MapReduce 的工作原理，描述 MapReduceV2（也就是 YARN）的架构和执行流程。本章重点介绍如何设计 MapReduce 程序，给出了在 Eclipse 中实现 Java 语言 MapReduce 程序的具体过程。

第 6 章 Hive 和 HBase 的安装与应用，主要介绍 Hive 和 HBase 的安装配置和应用方法，同时也介绍 MySQL 和 ZooKeeper 的安装与应用。

第 7 章 Sqoop 和 Kafka，介绍 Sqoop 和 Kafka 组件的安装及其基本应用方法。

第 8 章 Spark 集群安装与开发环境配置，介绍 Spark 架构及其工作原理，详细介绍 Spark 开发环境的安装与配置，包括热门的 IntelliJ IDEA 集成开发环境的安装与基本应用。

第 9 章 Spark 应用基础，介绍 Spark 程序的运行模式和应用设计方法，通过编写计算圆周率 Pi、基于随机森林模型的贷款风险预测 Scala 程序，展示了在集成开发环境 IDEA 中编写 Spark 程序的流程。

第三篇 大数据处理与项目开发

第 10 章 交互式数据处理，介绍如何利用 Hive 进行大数据处理和分析。Hive 是建立在 Hadoop MapReduce 基础上的数据仓库工具，用户借助 SQL 语句，可完成很多处理和分析，因此，对实际工作者有很大帮助。

第 11 章 协同过滤推荐系统，介绍推荐算法的基本概念和应用，展示基于 Spark 的机器学习库 MLlib 实现的协同推荐应用。

第 12 章 销售数据分析系统，通过一个完整的销售数据分析系统设计，展示如何利用 Hadoop 的各种组件开发实际的大数据应用系统。本章运用到的组件包括 HDFS、MySQL、Eclipse、Phoenix、HBase、WebCollector、Sevlet、Tomcat 等，所展示的数据和应用均来自真实场景，对读者有较高参考价值。

本书特点

把原理、架构、运行流程分析与实际应用融合起来介绍，融合性阐述框架优于单纯的原理分析，因为原理最终要付诸应用。

本书高度重视实践能力的培养，对系统安装、配置和应用过程给出了十分详细的描述，所有实验都是基于实际完成的操作介绍的，并配有现场截图，为读者展示了真实、详尽、可重现的场景，十分方便读者自学和钻研。

与很多大数据技术书籍不同，本书突出了数据处理本身，深入介绍了如何运用技术进行实际的数据分析，所采用的数据样本来自生产一线，所展示的项目具有实用的参考价值，读者掌握这些技术之后，就可以开始进行项目开发了。

本书的读者群

本书十分适合初学者入门和进阶。

本书也可供那些已经学习过 Hadoop 组件技术，但希望全面、系统地理解并掌握实际应用的读者参考。

本书对从事大数据项目开发的专业人员也有参考价值，书中所描述的 Hadoop 组件应用中遇到的各种问题及其解决办法，十分实用。

本书特别适合自学，读者完全可以利用本书给出的资源和示例，一步一步地完成各项操作和应用，体验一种登堂入室的成就感。

致谢

感谢大数据时代，感谢开源社区，感谢 Apache 基金会，感谢 Google，感谢所有关心和热爱大数据的人们！

作者在创作本书中借鉴了中科普开（北京）科技公司的部分培训资源，在此谨表示衷心的感谢。特别感谢中南大学郑瑾副教授，本书的部分内容使用了她编撰的书稿。由衷地感谢王建新教授、李建彬教授、张祖平教授，他们耐心地审阅了本书，提出了中肯的意见和建议。非常感谢电子工业出版社田宏峰编辑，他细心专业的工作方式，给作者留下深刻印象，并为本书的高质量印装提供了保障。

由于作者水平有限，本书的错误和疏漏在所难免，恳请广大读者提出宝贵意见和建议。联系邮箱：djhuang@csu.edu.cn。

作者

2017年6月于长沙

目 录

第一篇 大数据的基本概念和技术

第 1 章 绪论	3
1.1 时代背景	3
1.1.1 全球大数据浪潮	3
1.1.2 我国的大数据国家战略	5
1.2 大数据的概念	7
1.2.1 概念	7
1.2.2 特征	8
1.3 技术支撑体系	9
1.3.1 概览	9
1.3.2 大数据采集层	9
1.3.3 大数据存储层	10
1.3.4 大数据分析（处理与服务）层	11
1.3.5 大数据应用层	11
1.3.6 垂直视图	13
1.4 大数据人才及其能力要求	14
1.4.1 首席数据官	14
1.4.2 数据科学家（数据分析师）	15
1.4.3 大数据开发工程师	16
1.4.4 大数据运维工程师	17
1.5 本章小结	17
第 2 章 Hadoop 大数据关键技术	19
2.1 Hadoop 生态系统	19
2.1.1 架构的基本理论	19
2.1.2 主要组件及其关系	21

2.2	数据采集	24
2.2.1	结构化数据采集工具	24
2.2.2	日志文件采集工具与技术	25
2.3	大数据存储技术	29
2.3.1	相关概念	29
2.3.2	分布式文件存储系统	34
2.3.3	数据库与数据仓库	38
2.4	分布式计算框架	43
2.4.1	离线计算框架	43
2.4.2	实时流计算平台	50
2.5	数据分析平台与工具	57
2.5.1	面向大数据的数据挖掘与分析工具	57
2.5.2	机器学习	61
2.6	本章小结	66

第二篇 Hadoop 大数据平台搭建与基本应用

第 3 章	Linux 操作系统与集群搭建	69
3.1	Linux 操作系统	69
3.1.1	概述	69
3.1.2	特点	70
3.1.3	Linux 的组成	72
3.2	Linux 安装与集群搭建	75
3.2.1	安装 VMware Workstation	75
3.2.2	在 VMware 上安装 Linux (CentOS7)	79
3.3	集群的配置	91
3.3.1	设置主机名	91
3.3.2	网络设置	93
3.3.3	关闭防火墙	98
3.3.4	安装 JDK	99
3.3.5	免密钥登录配置	102
3.4	Linux 基本命令	105
3.5	本章小结	112
第 4 章	HDFS 安装与基本应用	113
4.1	HDFS 概述	113

4.1.1	特点	113
4.1.2	主要组件与架构	114
4.2	HDFS 架构分析	114
4.2.1	数据块	114
4.2.2	NameNode	115
4.2.3	DataNode	116
4.2.4	SecondaryNameNode	117
4.2.5	数据备份	117
4.2.6	通信协议	118
4.2.7	可靠性保证	118
4.3	文件操作过程分析	119
4.3.1	读文件	119
4.3.2	写文件	120
4.3.3	删除文件	122
4.4	Hadoop HDFS 安装与配置	122
4.4.1	解压 Hadoop 安装包	122
4.4.2	配置 Hadoop 环境变量	123
4.4.3	配置 Yarn 环境变量	124
4.4.4	配置核心组件文件	125
4.4.5	配置文件系统	125
4.4.6	配置 yarn-site.xml 文件	126
4.4.7	配置 MapReduce 计算框架文件	128
4.4.8	配置 Master 的 slaves 文件	129
4.4.9	复制 Master 上的 Hadoop 到 Slave 节点	129
4.5	Hadoop 集群的启动	130
4.5.1	配置操作系统环境变量	130
4.5.2	创建 Hadoop 数据目录	131
4.5.3	格式化文件系统	132
4.5.4	启动和关闭 Hadoop	133
4.5.5	验证 Hadoop 是否启动成功	133
4.6	Hadoop 集群的基本应用	136
4.6.1	HDFS 基本命令	136
4.6.2	在 Hadoop 集群中运行程序	139
4.7	本章小结	141

第 5 章	MapReduce 与 Yarn	143
5.1	MapReduce 程序的概念	143
5.1.1	基本编程模型	143
5.1.2	计算过程分析	144
5.2	深入理解 Yarn	147
5.2.1	Yarn 的基本架构	147
5.2.2	Yarn 的工作流程	151
5.3	在 Linux 平台安装 Eclipse	152
5.3.1	Eclipse 简介	153
5.3.2	安装并启动 Eclipse	154
5.4	开发 MapReduce 程序的基本方法	155
5.4.1	为 Eclipse 安装 Hadoop 插件	156
5.4.2	WordCount: 第一个 MapReduce 程序	160
5.5	本章小结	175
第 6 章	Hive 和 HBase 的安装与应用	177
6.1	在 CentOS7 下安装 MySQL	177
6.1.1	下载或复制 MySQL 安装包	177
6.1.2	执行安装命令	178
6.1.3	启动 MySQL	179
6.1.4	登录 MySQL	179
6.1.5	使用 MySQL	181
6.1.6	问题与解决办法	182
6.2	Hive 安装与应用	183
6.2.1	下载并解压 Hive 安装包	183
6.2.2	配置 Hive	184
6.2.3	启动并验证 Hive	187
6.2.4	Hive 的基本应用	189
6.3	ZooKeeper 集群安装	190
6.3.1	ZooKeeper 简介	190
6.3.2	安装 ZooKeeper	191
6.3.3	配置 ZooKeeper	191
6.3.4	启动和测试	193
6.4	HBase 的安装与应用	195

6.4.1	解压并安装 HBase	195
6.4.2	配置 HBase	196
6.4.3	启动并验证 HBase	199
6.4.4	HBase 的基本应用	200
6.4.5	应用 HBase 中常见问题及其解决办法	203
6.5	本章小结	204
第 7 章	Sqoop 和 Kafka 的安装与应用	205
7.1	安装部署 Sqoop	205
7.1.1	下载或复制 Sqoop 安装包	205
7.1.2	解压并安装 Sqoop	206
7.1.3	配置 Sqoop	206
7.1.4	启动并验证 Sqoop	208
7.1.5	测试 Sqoop 与 MySQL 的连接	209
7.2	安装部署 Kafka 集群	211
7.2.1	下载或复制 Kafka 安装包	211
7.2.2	解压缩 Kafka 安装包	211
7.2.3	配置 Kafka 集群	211
7.2.4	Kafka 的初步应用	213
7.3	本章小结	218
第 8 章	Spark 集群安装与开发环境配置	219
8.1	深入理解 Spark	219
8.1.1	Spark 系统架构	219
8.1.2	关键概念	221
8.2	安装与配置 Scala	224
8.2.1	下载 Scala 安装包	225
8.2.2	安装 Scala	225
8.2.3	启动并应用 Scala	226
8.3	Spark 集群的安装与配置	226
8.3.1	安装模式	226
8.3.2	Spark 的安装	227
8.3.3	启动并验证 Spark	230
8.3.4	几点说明	234
8.4	开发环境安装与配置	236

8.4.1	IDEA 简介	236
8.4.2	IDEA 的安装	236
8.4.3	IDEA 的配置	238
8.5	本章小结	243
第 9 章	Spark 应用基础	245
9.1	Spark 程序的运行模式	245
9.1.1	Spark on Yarn-cluster	245
9.1.2	Spark on Yarn-client	246
9.2	Spark 应用设计	247
9.2.1	分布式估算圆周率	248
9.2.2	基于 Spark MLlib 的贷款风险预测	265
9.3	本章小结	285

第三篇 数据处理与项目开发术

第 10 章	交互式数据处理	289
10.1	数据预处理	289
10.1.1	查看数据	289
10.1.2	数据扩展	291
10.1.3	数据过滤	292
10.1.4	数据上传	293
10.2	创建数据仓库	294
10.2.1	创建 Hive 数据仓库的基本命令	294
10.2.2	创建 Hive 分区表	296
10.3	数据分析	299
10.3.1	基本统计	299
10.3.2	用户行为分析	301
10.3.3	实时数据	303
10.4	本章小结	304
第 11 章	协同过滤推荐系统	305
11.1	推荐算法概述	305
11.1.1	基于人口统计学的推荐	305
11.1.2	基于内容的推荐	306
11.1.3	协同过滤推荐	307

11.2	协同过滤推荐算法分析	308
11.2.1	基于用户的协同过滤推荐	308
11.2.2	基于物品的协同过滤推荐	310
11.3	Spark MLlib 推荐算法应用	312
11.3.1	ALS 算法原理	312
11.3.2	ALS 的应用设计	315
11.4	本章小结	329
第 12 章	销售数据分析系统	331
12.1	数据采集	331
12.1.1	在 Windows 下安装 JDK	331
12.1.2	在 Windows 下安装 Eclipse	334
12.1.3	将 WebCollector 项目导入 Eclipse	335
12.1.4	在 Windows 下安装 MySQL	336
12.1.5	连接 JDBC	339
12.1.6	运行爬虫程序	340
12.2	在 HBase 集群上准备数据	342
12.2.1	将数据导入到 MySQL	342
12.2.2	将 MySQL 表中的数据导入到 HBase 表中	344
12.3	安装 Phoenix 中间件	347
12.3.1	Phoenix 架构	347
12.3.2	解压安装 Phoenix	348
12.3.3	Phoenix 环境配置	349
12.3.4	使用 Phoenix	350
12.4	基于 Web 的前端开发	353
12.4.1	将 Web 前端项目导入 Eclipse	353
12.4.2	安装 Tomcat	355
12.4.3	在 Eclipse 中配置 Tomcat	355
12.4.4	在 Web 浏览器中查看执行结果	359
12.5	本章小结	361

第一篇

大数据的基本概念和技术

最早提出“大数据”时代到来的全球知名咨询公司麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来”。

本章主要分析大数据的时代背景与国家大数据战略，给出大数据的概念，并分析其特性，重点介绍大数据技术支撑体系，包括数据采集、存储、分布式计算和应用，最后简要讨论大数据人才特点与能力要求。

1.1 时代背景

1.1.1 全球大数据浪潮

为什么最近几年里大数据变得如此引人注目？大数据到底有多大？

一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满 1.68 亿张 DVD；发出的邮件有 2940 亿封之多；发出的社区帖子达 200 万个（相当于《时代》杂志 770 年的文字量）；卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万。

目前，全球数据量已经从 TB（ $1024\text{ GB}=1\text{ TB}$ ）级别跃升到 PB（ $1024\text{ TB}=1\text{ PB}$ ）、EB（ $1024\text{ PB}=1\text{ EB}$ ）乃至 ZB（ $1024\text{ EB}=1\text{ ZB}$ ）级别。国际数据公司（IDC）的研究结果表明，2008 年全球产生的数据量为 0.49 ZB，2009 年的数据量为 0.8 ZB，2010 年增长到 1.2 ZB，2011 年的数量更是高达 1.82 ZB，相当于全球每人产生 200 GB 以上的数据。而

到 2016 年，人类生产的所有印刷材料的数据量是 300 PB，全人类历史上说过的所有话的数据量大约是 5 EB。IBM 的研究称，整个人类文明所获得的全部数据中，有 90% 是过去两年内产生的。而到了 2020 年，全世界所产生的数据规模将达到今天的 44 倍。

这样的趋势将会持续下去。我们现在还处于大数据的初级阶段，随着技术的进步，我们的设备、交通工具和迅速发展的“可穿戴”科技将能互连互通。科技的进步已经使创造、采集和管理信息的成本降至十年前的六分之一，而从 2005 年起，用在硬件、软件、人才及服务之上的商业投资也增长了整整 50%，达到了 4000 亿美元。

正如《纽约时报》2012 年 2 月的一篇专栏文章所称，“大数据”时代已经降临，在商业、经济及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉。哈佛大学社会学教授加里金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

越来越多的政府、企业等机构开始意识到数据正在成为组织最重要的资产，数据分析能力正在成为组织的核心竞争力。

2012 年 3 月 22 日，美国政府宣布投资 2 亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。美国政府将数据定义为“未来的新石油”，并表示一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来，对数据的占有和控制甚至将成为陆权、海权、空权之外的另一种国家核心资产。

联合国也在 2012 年发布了大数据政务白皮书，指出大数据对于联合国和各国政府来说是一个历史性的机遇，人们如今可以使用极为丰富的数据资源，来对社会经济进行前所未有的实时分析，帮助政府更好地响应社会和经济运行。

最为积极的还是众多的 IT 企业。麦肯锡在一份名为“大数据，是下一轮创新、竞争和生产力的前沿”的专题研究报告中提出，“对于企业来说，海量数据的运用将成为未来竞争和增长的基础”，该报告在业界引起广泛反响。麦肯锡的报告发布后，大数据迅速成为了计算机行业争相传诵的热门概念，也引起了包括金融界在内的各行各业的高度关注。随着互联网技术的不断发展，数据本身是资产，这一点在业界已经形成共识。如果说云计算为数据资产提供了保管、访问的场所和渠道，那么如何盘活数据资产，使其为国家治理、企业决策乃至个人生活服务，则是大数据的核心议题，也是云计算内在的灵魂和必然的升级方向。事实上，全球互联网巨头都已意识到了“大数据”时代，数据的重要意义，包括谷歌、苹果、惠普、IBM、微软在内的全球 IT 巨头纷纷通过收购“大数据”相关厂商来实现技术整合，可见其对“大数据”的重视。