教育部－浪潮集团产学合作协同育人项目成果

普通高等学校计算机教育“十三五”规划教材

inspur 浪潮

Hadoop

应用开发与案例实战

慕课版

浪潮优派◎策划

穆建平 王建 商程◎主编

崔瑞娟 郭建磊 尹丛丛 郭长友◎副主编

扫一扫

下载附赠资源

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

图书在版编目（CIP）数据

Hadoop应用开发与案例实战 : 慕课版 / 穆建平, 王建, 商程主编. — 北京 : 人民邮电出版社, 2021.4
普通高等学校计算机教育“十三五”规划教材
ISBN 978-7-115-53778-2

Ⅰ. ①H… Ⅱ. ①穆… ②王… ③商… Ⅲ. ①数据处理软件－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2020)第059688号

内 容 提 要

Hadoop是目前比较流行的大数据框架之一，它使用简单的高级编程模型即可实现大型数据集的分布式存储和处理。

本书以Hadoop的概念、集群搭建、核心组件、实战案例等为主线，较为全面地介绍了Hadoop大数据存储及处理技术的相关知识。全书共10章，前9章主要讲解了Hadoop的基础知识，内容包括初识Hadoop、Hadoop的安装与配置、高可用与联邦、分布式文件系统HDFS、集群资源管理系统YARN、分布式计算框架MapReduce、Hadoop的I/O操作、Hadoop 3.x的新特性、Hadoop商业发行版等；第10章是Hadoop实战案例，以实际Hadoop框架的运用为导向引入了三个实战案例：Avro文件合并及多目录输出、网页域名分区统计及电商平台商品评价数据分析。

本书既可作为高校大数据相关技术类专业的教材和辅导书，也可作为大数据技术爱好者的自学用书。

◆ 主　　编　穆建平　王　建　商　程
副 主 编　崔瑞娟　郭建磊　尹丛丛　郭长友
责任编辑　张　斌
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　https://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷
◆ 开本：787×1092　1/16
印张：13.5　　2021年4月第1版
字数：284千字　　2021年4月北京第1次印刷

定价：49.80元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东市监广登字20170147号

前言 PREFACE

Apache Hadoop 是一款由 Apache 基金会开发的，用于大数据分布式存储和处理的开源软件。它所提供的软件库允许用户在完全不了解底层实现细节的情况下，使用简单的编程模型实现在跨计算机集群中对大规模数据集进行分布式处理。Hadoop 的分布式集群架构可以搭建在廉价的 X86 服务器上，它具有高可靠性、高扩展性、高容错性以及低成本等优点。目前，Hadoop 已经成长为一个全栈式大数据技术生态圈，并且在互联网等领域得到了广泛的运用。

浪潮集团是我国重要的云计算、大数据服务商，旗下拥有浪潮信息、浪潮软件、浪潮国际三家上市公司，业务涵盖云数据中心、云服务大数据、智慧城市、智慧企业四大产业群组，形成了涵盖 IaaS、PaaS、SaaS 三个层面的整体解决方案服务能力。浪潮集团是先进的信息科技产品与解决方案服务商，也是“云+数+AI”新型互联网企业，引领信息科技浪潮，推动社会文明进步。

浪潮优派科技教育有限公司（以下简称浪潮优派）是浪潮集团的下属子公司，本书由浪潮优派具有多年开发经验和实训经验的 IT 培训讲师撰写，全书各章知识点讲解条理清晰、循序渐进。本书配有实战案例的源代码、视频资料和电子课件，读者可登录人邮教育社区（www.ryjiaoyu.com）下载。

本书共 10 章，各章内容如下。

第 1 章　初识 Hadoop：介绍了 Hadoop 的背景及发展历程、Hadoop 的核心组件、Hadoop 生态系统及相关技术，以及 Hadoop 的十大应用场景。

第 2 章　Hadoop 的安装与配置：主要讲解 Hadoop 集群的两种安装与配置方式——伪分布式安装和完全分布式安装，此外还介绍了如何在正常运行的集群环境中动态添加、删除节点。

第 3 章　高可用与联邦：主要介绍了高可用的概念、必要性以及 Hadoop 高可用的搭建过程，此外还讲解了联邦的概念以及联邦主要解决的问题。

第 4 章　分布式文件系统 HDFS：主要介绍了 HDFS 的概念、架构及读写数据的流程，此外还讲解了 HDFS 操作所涉及的 Shell 命令、HDFS 常用的 API 应用等。

第 5 章　集群资源管理系统 YARN：介绍了 YARN 的产生背景、基本架构和工作

流程。

第 6 章　分布式计算框架 MapReduce：重点讲解了 MapReduce 的处理过程，并通过一个入门案例详细演示了 MapReduce 的执行过程。

第 7 章　Hadoop 的 I/O 操作：主要讲解了序列化的概念，重点介绍了 Hadoop 常用序列化的接口以及从文件中读写数据所涉及的相关接口的使用。

第 8 章　Hadoop 3.x 的新特性：主要介绍了 Hadoop 3.x 的发展背景、Hadoop 3.x 相对于 Hadoop 2.x 的改进以及 Hadoop 3.x 其他的新特性。

第 9 章　Hadoop 商业发行版：重点讲解了当前比较流行的商业发行版 CDH 的部署与应用，此外还简单介绍了 HDP、MapR Hadoop 和华为 Hadoop 等其他商业发行版本。

第 10 章　Hadoop 实战案例：本章采用浪潮集团真实的大数据项目，重点讲解 Avro 文件合并及多目录输出、网页域名分区统计和电商平台商品评价数据分析三个实战案例。

本书由浪潮优派的穆建平、王建、商程担任主编，浪潮优派的崔瑞娟、山东电子职业技术学院的郭建磊、山东管理学院的尹丛丛、德州学院的郭长友担任副主编。他们对全书进行了审核和统稿。此外，参与本书编写的人员还有浪潮卓数大数据产业发展有限公司的姚民伟、杨胜华和杨担通。另外，为了使本书更适合高校的需要，与浪潮集团有合作关系的部分高校老师也协助了本书的编写工作，有山东女子学院胡蔚蔚，德州学院胡凯，山东管理学院常晓炜、刘乃文、刘涛、赵丽丽、李雅林和王高峰。感谢他们在本书撰写过程中所提供的帮助和支持。

由于时间仓促和编者水平有限，书中难免存在一些疏漏和不足之处，欢迎读者朋友批评指正。

编者

2020 年 10 月