

Hadoop 大数据分布式计算 框架——原理与应用

杨成伟 祝翠玲 刘位龙 编著

中国财经出版传媒集团



经济科学出版社
Economic Science Press

图书在版编目 (CIP) 数据

责任编辑：于 源 姜思伊

责任校对：刘 昕

责任印制：范 艳

Hadoop 大数据分布式计算框架——原理与应用

杨成伟 祝翠玲 刘位龙 编著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：010-88191217 发行部电话：010-88191522

网址：www.esp.com.cn

电子邮箱：esp@esp.com.cn

天猫网店：经济科学出版社旗舰店

网址：http://jjkxcbs.tmall.com

印刷厂印刷

710×1000 16 开 印张 000000 字

2023 年 2 月第 1 版 2023 年 2 月第 1 次印刷

印数：000001—000000 册

ISBN 978-7-5218-0000-0 定价：00.00 元

(图书出现印装问题，本社负责调换。电话：010-88191545)

(版权所有 侵权必究 打击盗版 举报热线：010-88191661)

QQ: 2242791300 营销中心电话：010-88191537

电子邮箱：dbts@esp.com.cn)

前言

随着人类进入大数据时代，人们获取数据的方式、方法、途径不断扩展，各行各业能够获得的数据量越来越多，数据已经成为社会发展过程中至关重要的生产要素。各国政府高度重视大数据技术的研究和相关产业的发展进程。习近平总书记明确要求，要加快发展数字经济，促进数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。^① 党的二十大提出坚持人民至上、坚持自信自立、坚持守正创新、坚持问题导向、坚持系统观念、坚持胸怀天下等“六个坚持”，为我国大数据发展提供了世界观、方法论及科学指导。^② 在这样一个历史环境下，我们如何贯彻执行党的二十大精神 and 指示，快速适应大数据时代的新特征，已经成为构建数字中国和推动数字经济、数字社会、数字政府等各领域融合发展的首要任务。

本书以二十大精神为引领，探索和提供大数据管理与应用中的支撑技术及方法，加快促进政府和企业数字化转型步伐。大数据分布式框架技术具有高可靠性、高可扩展性、高可用性等特点，其中，Hadoop 作为大数据分布式计算框架的典型开源代表，在大数据存储、分布式计算、分布式资源管理等方面独具优势，已广泛应用于各行业领域。本书以 Hadoop 为例，详细介绍了分布式环境搭建、分布式数据存储、分布式数据计算、分布式资源管理及配置等技术方面的原理与应用。全书共分为六章，第 1 章主要介

① 唐朵朵．打造具有国际竞争力的数字产业集群 [N]．新华社，2022 - 12 - 05.

② 周佑勇．在法治中国建设中把握“六个必须坚持” [N]．光明日报，2023 - 06 - 19.

绍大数据的定义、特征及所带来的各种挑战；第2章主要介绍大数据环境中常用的Linux命令；第3、4、5章主要介绍Hadoop分布式框架中HDFS、YARN、Mapreduce三大关键技术的工作原理和配置方法；第6章通过一个项目案例介绍在分布式环境下进行大数据项目设计的方法和过程。

本书可作为国内本专科学校大数据专业课程的配套教材。根据教学经验，建议安排32学时理论课，16周教学，每周2学时。已经具备大数据实验实训环境的高校，可以安排32个学时的上机实践课，16周教学，每周2学时，并可搭配本书姊妹教材《Hadoop 大数据分布式计算框架编程与实践》一起使用，该书提供了大数据开发环境搭建的详细指导。

本书在编写过程中，很多老师给我们团队反馈了大量宝贵的意见，解答了编写过程中的疑惑。同学们做了大量辅助性工作，如付尹杰、卢冠霖、袁久存、李希茹、苏日新、徐展展、姜玲钰等学生在资料收集、制图、算法和模型实现等方面付出了很多时间和精力，在此向他们表示感谢。本书也得到山东省自然科学基金（ZR2019MG037）、山东省重点研发计划（重大科技创新工程）项目（2020CXGC010110）、山东省重点研发计划（重大科技示范工程）项目（2021SFGC0102）的资助以及学校学院的大力支持。另外，本书在撰写的过程中参考了部分国内外优秀的大数据教材及论文，在此不再一一列举，一并向这些作者致谢。

由于团队水平和能力有限，虽经多次校对，难免仍有疏漏和不足之处，请广大同行和读者批评指正。

杨成伟

山东财经大学大数据与人工智能研究院

2023年8月5日

目 录

第 1 章 Hadoop 大数据分布式计算框架概述	1
1.1 大数据基础	1
1.2 Hadoop 概述	12
1.3 Hadoop 关键技术	22
1.4 Hadoop 的优势与不足	31
1.5 本章小结	32
本章习题	33
本章主要参考文献	35
第 2 章 Hadoop 大数据分布式环境	37
2.1 大数据 Linux 基础	37
2.2 Hadoop 系统部署方式	64
2.3 Hadoop 3. x 环境搭建	67
2.4 本章小结	75
本章习题	75
本章主要参考文献	78
第 3 章 Hadoop 分布式文件系统原理	79
3.1 集群与分布式文件系统	79
3.2 HDFS 架构和组件	89
3.3 HDFS 的高可用机制	100
3.4 HDFS 的数据读写	115
3.5 HDFS 联邦机制	124
3.6 本章小结	132

本章习题·····	132
本章主要参考文献·····	134
第 4 章 分布式调度系统 YARN	135
4.1 YARN 分布式资源管理	135
4.2 YARN 体系结构	139
4.3 YARN 作业调度	150
4.4 YARN 高级特性	158
4.5 YARN 的应用实践	170
4.6 本章小结	178
本章习题·····	178
本章主要参考文献·····	180
第 5 章 MapReduce 分布式计算框架	182
5.1 MapReduce 概述	182
5.2 MapReduce 工作原理	188
5.3 Shuffle 过程	201
5.4 MapReduce 设计模式	207
5.5 MapReduce 的应用实践	279
5.6 MapReduce 优化	298
5.7 本章小结	300
本章习题·····	300
本章主要参考文献·····	302
第 6 章 分布式大数据分析项目案例	304
6.1 项目背景	304
6.2 功能需求	305
6.3 系统设计	306
6.4 项目实施	309
6.5 本章小结	351
本章习题·····	352