

大数据分布式基础与编程实践

祝翠玲 杨成伟 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

· 北 京 ·

图书在版编目 (CIP) 数据

大数据分布式基础与编程实践 / 祝翠玲, 杨成伟,
刘位龙编著. -- 北京 : 经济科学出版社, 2024. 11.
ISBN 978 - 7 - 5218 - 6435 - 9
I. TP274

中国国家版本馆 CIP 数据核字第 2024NM2501 号

责任编辑: 冯 蓉
责任校对: 靳玉环
责任印制: 范 艳

大数据分布式基础与编程实践

祝翠玲 杨成伟 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](http://www.esp.com.cn)

电子邮箱: [esp@ esp. com. cn](mailto:esp@esp.com.cn)

天猫网店: 经济科学出版社旗舰店

网址: [http://jjkxcs. tmall. com](http://jjkxcs.tmall.com)

北京季蜂印刷有限公司印装

710 × 1000 16 开 32.25 印张 545000 字

2024 年 11 月第 1 版 2024 年 11 月第 1 次印刷

ISBN 978 - 7 - 5218 - 6435 - 9 定价: 78.00 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](mailto:dbts@esp.com.cn))

前言

在数字化浪潮席卷全球的今天，大数据已成为推动社会进步与产业升级的重要力量。其庞大的数据量、高速的流动性和多样化的类型，不仅挑战着传统数据处理技术的极限，也孕育了前所未有的创新机遇。从党的十八大以来，我国深入实施国家大数据战略，协同推进数字产业化和产业数字化。党的二十大报告也指出，要加快建设网络强国、数字中国。习近平总书记深刻指出：“加快数字中国建设，就是要适应我国发展新的历史方位，全面贯彻新发展理念，以信息化培育新动能，用新动能推动新发展，以新发展创造新辉煌。”^① 在这样一个历史机遇下，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。

本书以党的二十大精神为引领，探索和提供大数据管理与应用中的相关支撑技术和方法，以期能够为加快数字中国建设尽绵薄之力。分布式框架能够从数据存储、数据处理、资源管理等方面提供一种新的解决方案。Hadoop 作为 Apache 软件基金会旗下的一种开源分布式框架，以其高扩展性、高容错性和低成本的优势，成为处理大规模数据集的首选工具，它不仅能够处理 PB 级的数据量，还能在成百上千的廉价硬件上实现高效运行，为数据密集型应用提供了强有力的支持。本书聚焦于 Hadoop 的分布式环境部署和编程实践，通过理论讲解、环境部署与编程实践相结合的方式，

^① 习近平关于网络强国论述摘编 [M]. 北京：中央文献出版社，2021：46.

每一章都辅以丰富的示例代码、图片展示和详细的操作步骤，使读者能够轻松上手，快速掌握相关知识、相关环境部署及编程技巧。

本书共分为三部分：工具实践篇、数据库实践篇和编程实践篇。工具实践篇主要介绍在大数据分布式环境中的各种基础环境的安装和部署及简单应用；数据库实践篇主要介绍在大数据分布式环境中关系型数据库、非关系型数据库以及数据仓库的安装部署及使用；编程实践篇主要介绍在已经部署的分布式环境中进行HDFS程序设计、MapReduce 分布式编程、Python 分布式编程实践以及利用 Spark 环境进行编程实践。

全书共分为 13 章，具体内容简介如下：

工具实践篇：

第 1 章主要介绍大数据与 Linux 基础实践。首先介绍了大数据的基础环境 Linux 操作系统概况、特点及常用的 Ubuntu 操作系统概况；其次介绍了 Linux 虚拟机的概念及常用的虚拟机软件 VirtualBox 和 VMware 工作站；再次分别介绍了 VirtualBox 和 VMWare 工作站两种虚拟机软件的下载及安装过程；最后分别介绍了在 VirtualBox 和 VMWare 工作站中安装 Ubuntu 操作系统的具体过程及 Linux 虚拟机的导入/导出操作。

第 2 章主要介绍 Hadoop 分布式框架实践。首先介绍 Hadoop 分布式框架概况、运行模式以及优势和不足；其次介绍了安装 Hadoop 所需要的基础 JDK 工具包；最后介绍了在 Linux 虚拟机中利用伪分布式模式安装部署 Hadoop 的具体过程。

第 3 章主要介绍 Spark 分布式框架实践。首先介绍了 Spark 的发展史、Spark 与 MapReduce 的区别、Spark 的优势；其次介绍了 Spark 生态系统以及 Spark 的三种部署方式；最后介绍了在 Ubuntu 环境中安装和部署 Spark 的过程。

第 4 章主要介绍 HDFS 分布式文件系统实践。首先介绍了分布式文件系统 DFS 和常见的分布式文件系统及其特点；其次介绍

了 Hadoop 分布式文件系统 HDFS 概况；最后介绍了分布式文件系统 HDFS 中的文件和目录操作的相关命令及使用方式。

第 5 章主要介绍在 Linux 环境中的 Anaconda 应用实践。首先介绍了 Python 的开发环境 Anaconda 概况；其次介绍了在 Linux 中对 Anaconda 的安装部署过程；最后介绍了 conda 的环境管理和包管理情况。

第 6 章主要介绍在分布式环境中的 Python 应用实践。首先介绍 Python 的发展史、特点、常用的内置库和标准库；其次介绍在 Ubuntu 中 Python 的安装部署；最后，介绍在 IDEA 集成开发环境中利用安装 Python 插件的方式进行 Python 项目开发的过程和具体步骤。

数据库实践篇：

第 7 章主要介绍 MySQL 关系型数据库实践。首先介绍关系型数据模型、结构化查询语言 SQL 和 MySQL 数据库及其特点；其次介绍了在 Linux 虚拟机中安装 MySQL 的具体过程和安装过程中经常遇到的问题及解决方法；最后介绍在 MySQL 中对数据库、数据表和数据记录的常用操作。

第 8 章主要介绍 HBase 非关系型数据库实践。首先介绍非关系型数据库（NoSQL 数据库）的相关概念、NoSQL 数据库和 SQL 数据库的区别以及常见的四种 NoSQL 数据库及其特点；其次详细介绍了数据库 HBase 概况、所使用的数据模型和在 HBase 中的常用的 Shell 命令及操作实例；最后介绍在 Hadoop 分布式环境中 HBase 的安装、部署及使用。

第 9 章主要介绍 Redis 非关系型数据库实践。首先介绍内存数据库的相关概念、Redis 内存数据库概况以及 Redis 数据库的数据类型和主要功能；其次介绍内存数据库 Redis 的安装和部署；最后介绍 Redis 数据库的具体使用。

第 10 章主要介绍 Hive 数据仓库实践。首先介绍数据仓库的概念和特点；其次详细介绍了数据仓库 Hive 概况、体系结构、数

据模型、三种部署方式、HiveQL 的执行过程以及常用的 HiveQL 命令；最后介绍数据仓库 Hive 的安装、部署与使用。

编程实践篇：

第 11 章主要介绍 HDFS 分布式文件系统编程实践。首先介绍在分布式文件系统 HDFS 中进行程序设计的相关概念、HDFS 中的 Java API 操作以及在 HDFS 中进行程序设计的集成开发环境 IDEA 概况；其次介绍 IDEA 集成开发环境的安装和部署；最后详细介绍在 IDEA 中进行 HDFS 程序设计的相关操作。

第 12 章主要介绍了 MapReduce 分布式编程实践。首先介绍分布式计算的相关概念并引出分布式编程模型 MapReduce；其次介绍 MapReduce 的架构演变及在相应架构下 MapReduce 的组成和作业执行过程；再次介绍 MapReduce 编程模型中的五大核心组件、工作流程、主要功能及在 MapReduce 中的 Shuffle 过程；最后通过求聚合、求均值、去重和单表 join 连接四个例子详细介绍利用集成开发环境 IDEA 开发 MapReduce 程序的详细过程。

第 13 章主要介绍 Python 分布式编程实践。首先介绍了四种常用的 Python 开发环境；其次介绍了在 IDEA 中编写 Python 爬虫程序的具体实践过程。

附录部分介绍了 Hadoop3.2.0 中 HDFS 常用命令的语法格式及各命令中包含的参数及其含义，以方便读者进行使用参考。

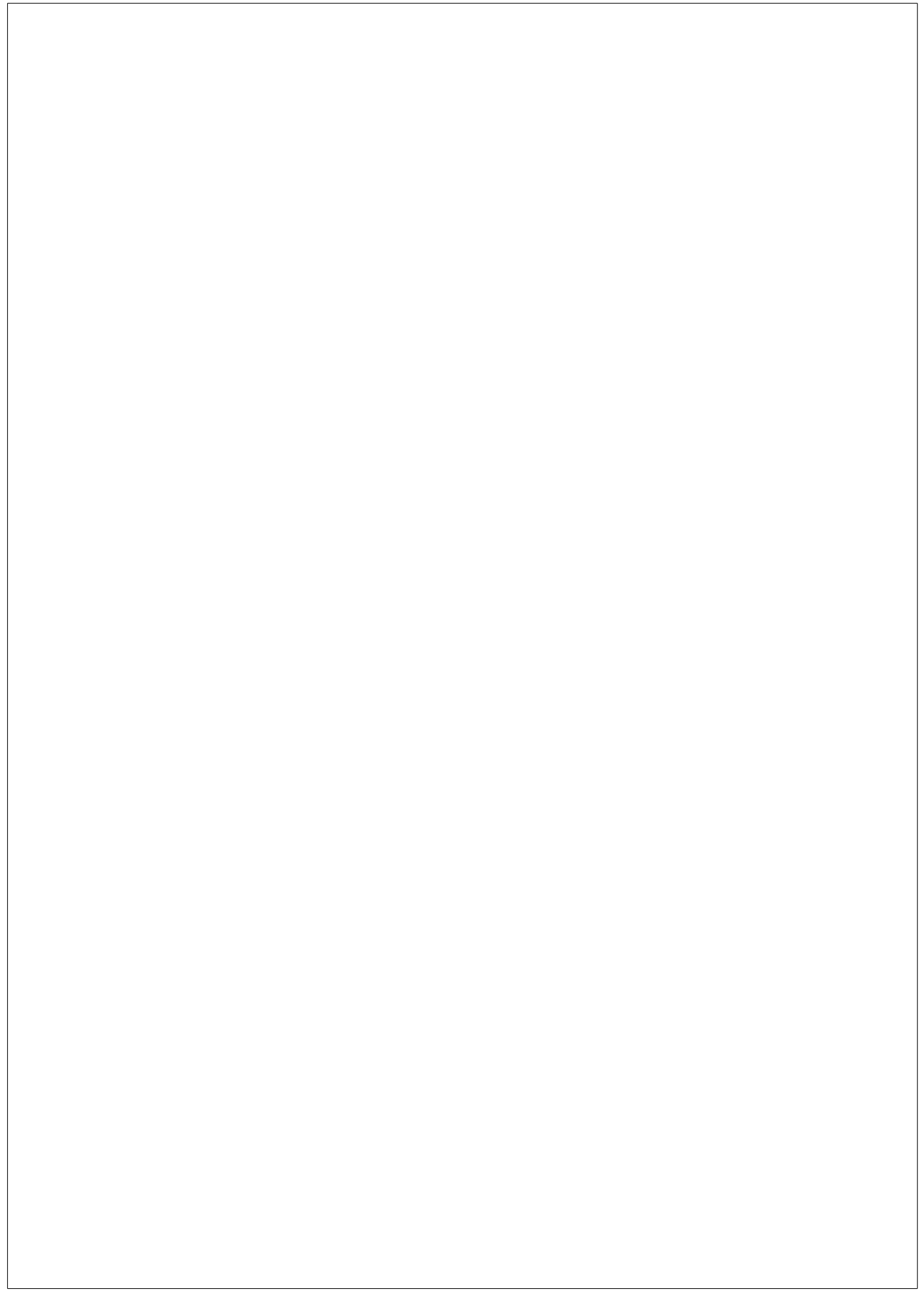
本书在编写过程中，很多老师给团队给予了大量宝贵的意见，山东财经大学管理科学与工程学院的学生胡明远、梁佼佼、王允芃、袁久存、胡博毓、蒋开通、王鹏、于新荷等提供了大量辅助性工作，在此对他们表示感谢。本书的写作还得到了山东省自然科学基金（编号：ZR2019MG037）；山东省重点研发计划（重大科技创新工程）项目（编号：2020CXGC010110）；山东省重点研发计划（重大科技示范工程）项目（编号：2021SFGC0102）、2024 年度山东省社科规划研究青年项目（编号：24DGLJ08）和 2024 年度教育部人文社会科学研究青年基金（AI 道德失范识别赋能基层

“德治”中的双刃剑效应研究)的支持以及学校、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教材、论文,在此不再一一列举,一并向他们表示感谢。

由于团队水平和能力有限,虽经多次校对,书中难免有疏漏和不足之处,衷心希望广大同行和读者批评指正。

祝翠玲

2024年8月29日



目 录

第一篇 工具实践篇

第 1 章 大数据与 Linux 系统实践	3
本章学习目的.....	3
1.1 Linux 操作系统.....	3
1.2 Linux 虚拟机	13
1.3 虚拟机软件安装部署.....	17
1.4 Linux Ubuntu 安装部署	30
1.5 Linux 虚拟机导入/导出.....	65
1.6 本章小结.....	71
本章习题	71
第 2 章 Hadoop 分布式框架实践	74
本章学习目的	74
2.1 Hadoop 分布式框架	75
2.2 JDK 工具包	79
2.3 Hadoop 分布式部署	81
2.4 本章小结.....	99
本章习题.....	100
第 3 章 Spark 分布式框架实践.....	102
本章学习目的.....	102
3.1 Spark 介绍.....	102
3.2 在 Ubuntu 环境中安装 Spark	112

3.3 本章小结	120
本章习题	120
第4章 HDFS 分布式文件系统实践	122
本章学习目的	122
4.1 分布式文件系统	122
4.2 分布式文件系统 HDFS 操作	128
4.3 本章小结	145
本章习题	146
第5章 Anaconda 应用实践	148
本章学习目的	148
5.1 Anaconda 介绍	148
5.2 Linux 中 Anaconda 的部署与使用	153
5.3 本章小结	166
本章习题	167
第6章 Python 应用实践	168
本章学习目的	168
6.1 Python 简介	168
6.2 在 Ubuntu 中安装 Python	176
6.3 在 IDEA 中进行 Python 实践	179
6.4 本章小结	187
本章习题	188

第二篇 数据库实践篇

第7章 MySQL 关系型数据库实践	191
本章学习目的	191
7.1 关系数据库	191
7.2 安装 MySQL	194
7.3 MySQL 常用操作	205

7.4 本章小结	213
本章习题	213
第 8 章 HBASE 非关系型数据库实践	215
本章学习目的	215
8.1 NoSQL 数据库	215
8.2 HBase 概述	220
8.3 HBase 的部署与使用	231
8.4 本章小结	242
本章习题	243
第 9 章 Redis 非关系型数据库实践	246
本章学习目的	246
9.1 Redis 数据库简介	246
9.2 Redis 安装与部署	253
9.3 Redis 的使用	260
9.4 本章小结	272
本章习题	273
第 10 章 Hive 数据仓库实践	275
本章学习目的	275
10.1 数据仓库	275
10.2 数据仓库 Hive	278
10.3 Hive 的部署与使用	295
10.4 本章小结	309
本章习题	309

第三篇 编程实践篇

第 11 章 HDFS 分布式文件系统编程实践	313
本章学习目的	313
11.1 HDFS 程序设计	313

11.2 HDFS 程序设计实践	323
11.3 本章小结	353
本章习题	353
第 12 章 MapReduce 分布式编程实践	355
本章学习目的	355
12.1 分布式计算基础	355
12.2 MapReduce 架构演变	360
12.3 MapReduce 编程模型	370
12.4 MapReduce 编程实践——求聚合	385
12.5 MapReduce 编程实践——求均值	396
12.6 MapReduce 编程实践——去重	404
12.7 MapReduce 编程实践——单表 join 连接	414
12.8 本章小结	428
本章习题	428
第 13 章 Python 分布式编程实践	433
本章学习目的	433
13.1 Python 开发环境	433
13.2 在 IDEA 中编写 Python 爬虫程序	442
13.3 本章小结	461
本章习题	462
参考文献	464
附录 1 Hadoop 3.2.0 HDFS 命令指南	466
1. 用户命令	466
2. 管理命令	474
3. 调试命令	488
4. Hadoop 常用命令快速一览	489
5. HDFS 命令参考	491
6. 分布式复制	493

7. 管理员命令	493
附录 2 课程实践报告	495
部分课后题答案	496