分布式系统实验报告

**实验二**

**MapReduce分布式计算及其应用**

姓名：

学号：

院系：

2025年 5 月 26 日

# 一、实验目的与原理

1. 理解 Hadoop 生态中 HDFS、MapReduce、YARN 的核心思想与协同工作机制。

2. 掌握在 Windows 伪分布式 / Linux 伪分布式环境下部署 Hadoop 集群的方法。

3. 学会编写并运行 MapReduce 程序，完成典型数据处理任务（词频统计、电影评分分析）。

4. 熟悉 YARN 资源调度及 JobHistory Web UI 的使用。

# 二、实验环境

硬件环境：个人Laptop + 虚拟机（或云服务器）（≥ 4 GB 内存）

操作系统：Windows 11 + PowerShell 7，Ubuntu 22.04（可选）

软件环境：

- JDK 1.8 （Hadoop 3.3.1 与 winutils 兼容）

- Hadoop 3.3.1 伪分布式集群（NameNode、DataNode、SecondaryNameNode、ResourceManager、NodeManager）

# 三、实验步骤

## 3.1 Hadoop 集群搭建与验证

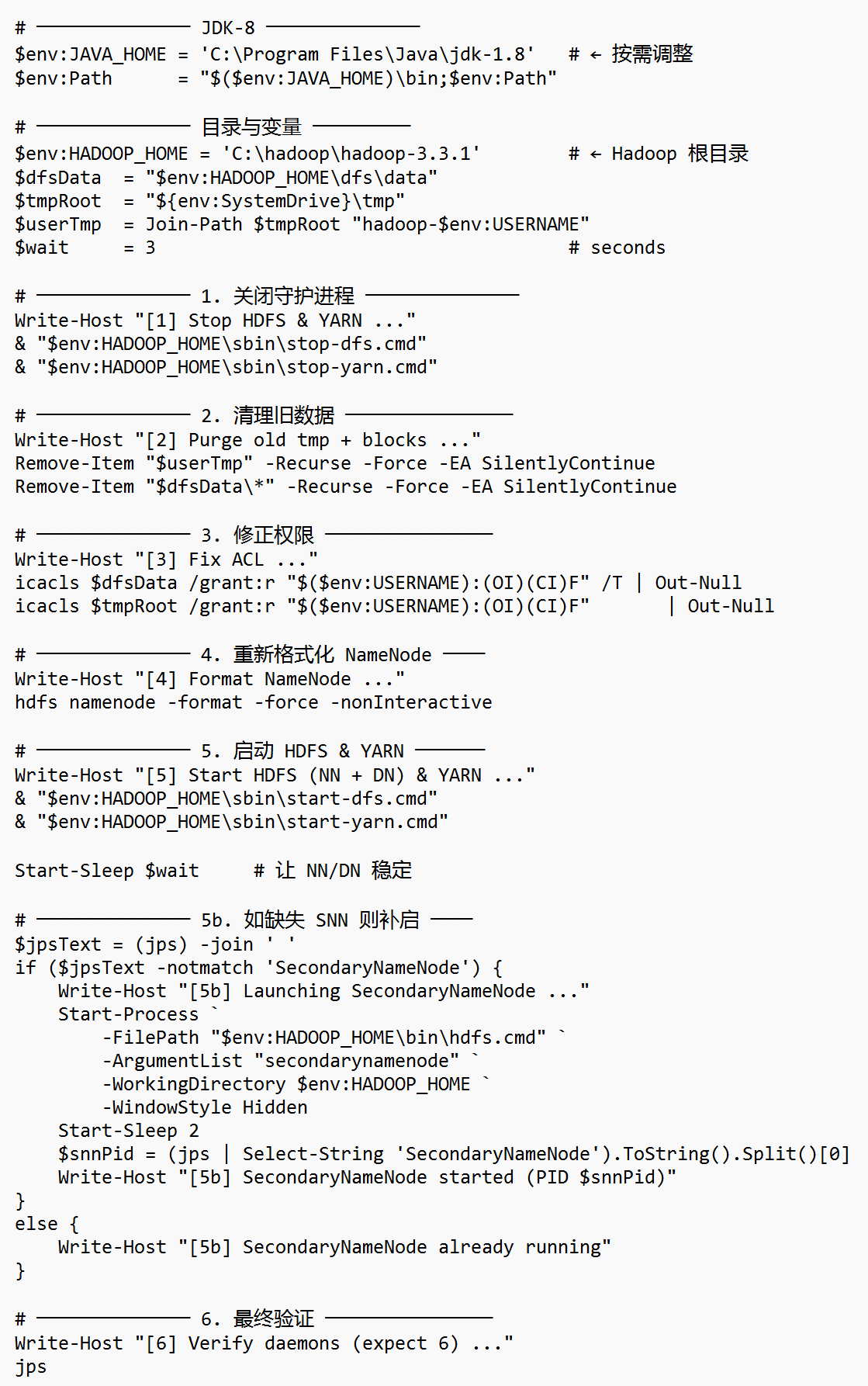
•执行 hadoop.ps1（或hadoop.sh），一键完成 JDK 切换、旧数据清理、NameNode 格式化、启动五大守护进程。

• 通过 jps 命令确认进程：NameNode / DataNode / SecondaryNameNode / ResourceManager / NodeManager。

（1）Windows端启动结果：

• 在 PowerShell（Admin）中执行 hadoop.ps1，一键完成 JDK 切换、旧数据清理、NameNode 格式化、启动五大守护进程。

1. 脚本代码：



1. yarn-nodemanager运行状态截图：

文本

AI 生成的内容可能不正确。

1. yarn-resourcemanager运行状态截图：

文本

AI 生成的内容可能不正确。

1. hadoop-datanode运行状态截图：

文本

AI 生成的内容可能不正确。

1. hadoop-namenode运行状态截图：

文本

AI 生成的内容可能不正确。

1. jps打印结果，显示6个进程：

文本

AI 生成的内容可能不正确。

（2）Linux端启动结果：

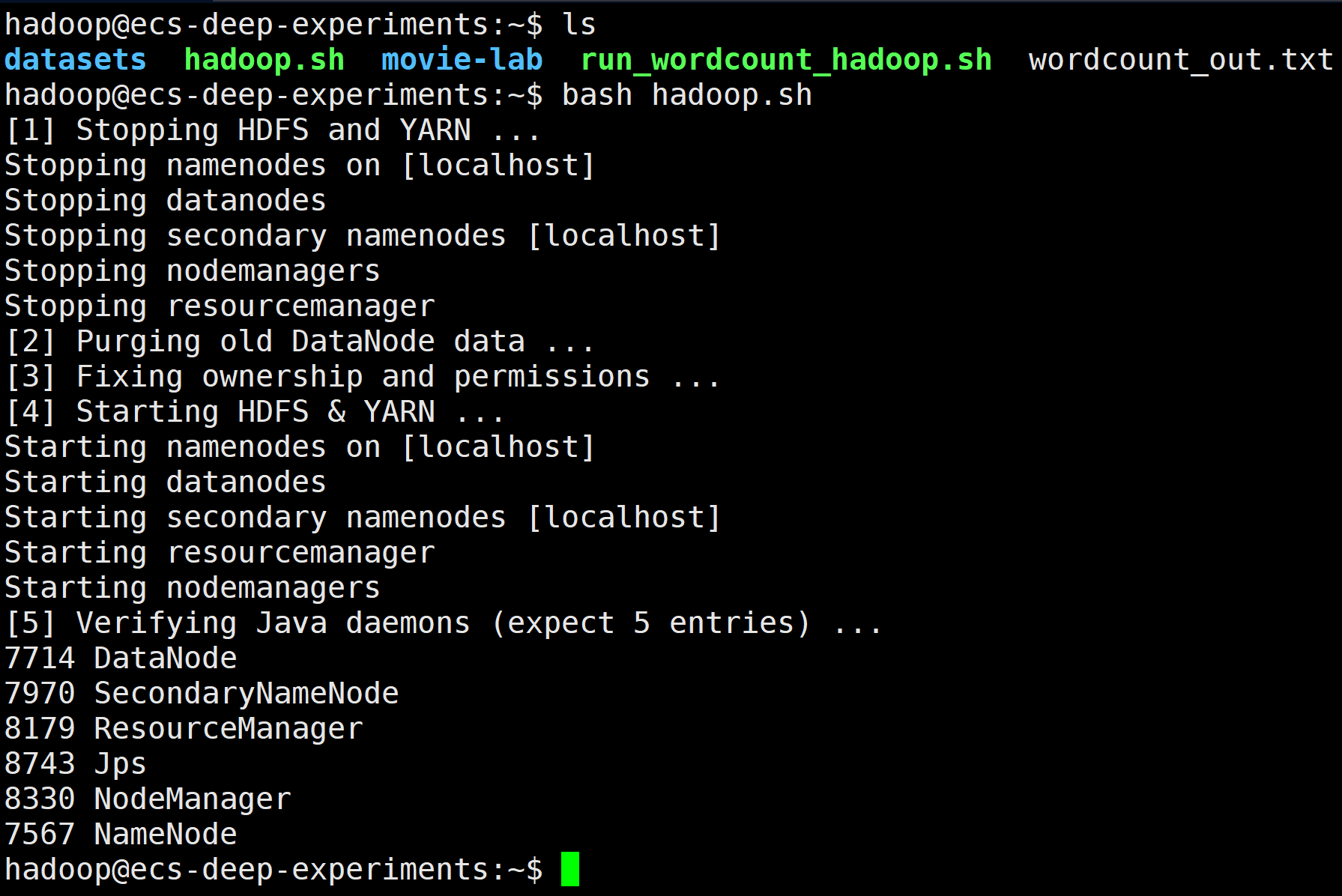
• 在 Linux终端中执行 hadoop.sh，一键完成 JDK 切换、旧数据清理、NameNode 格式化、启动五大守护进程。

1. 脚本代码：

**文本

AI 生成的内容可能不正确。**

1. jps打印结果，显示6个进程：



③ 在Linux终端测试端口连接：

文本

AI 生成的内容可能不正确。

（3）访问 Web（Windows和Linux）：

• NameNode Web UI：http://localhost:9870

图形用户界面

AI 生成的内容可能不正确。

图形用户界面, 文本, 应用程序, 电子邮件

AI 生成的内容可能不正确。

• YARN ResourceManager：http://localhost:8088

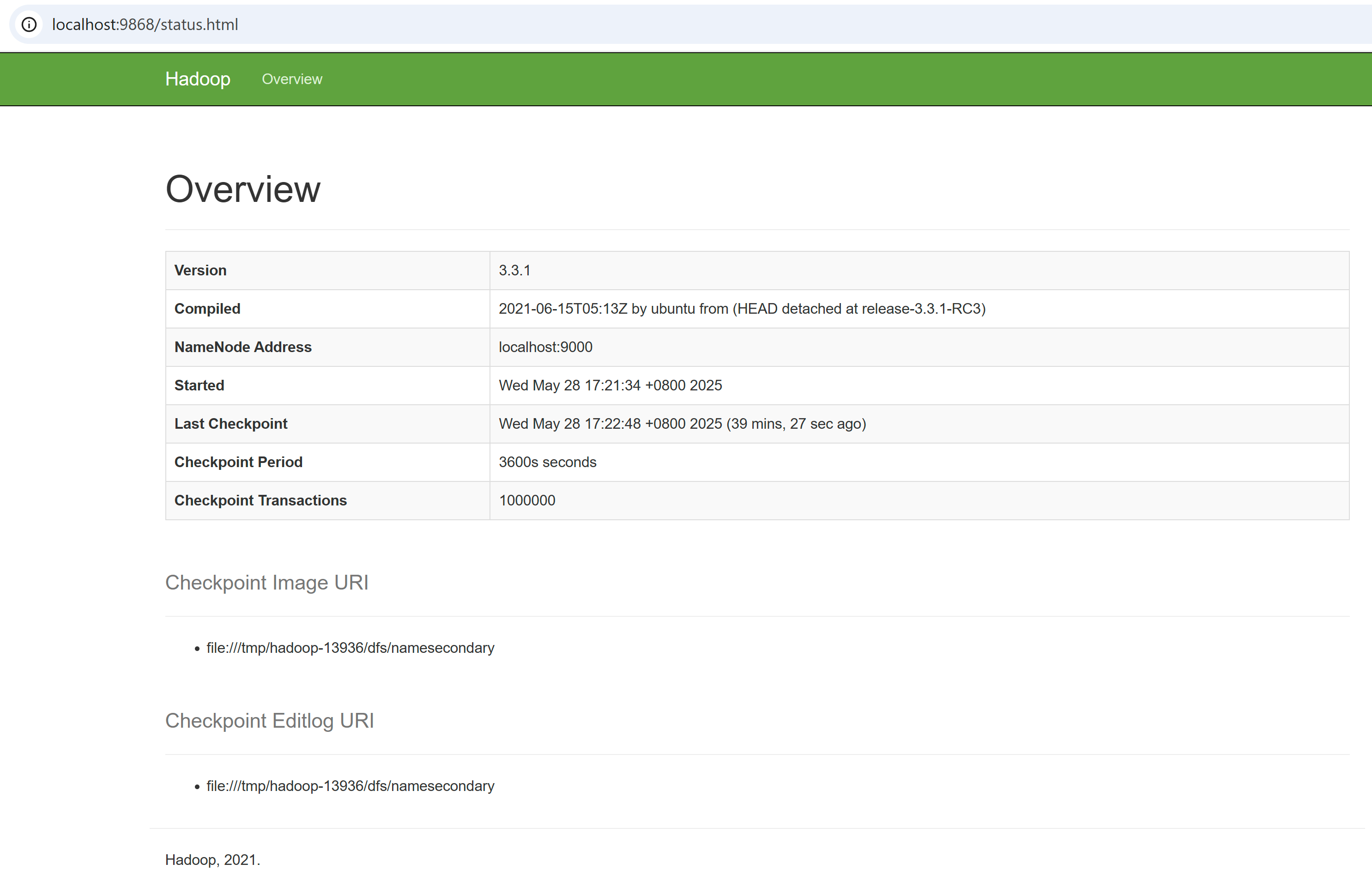
图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

图形用户界面, 文本, 应用程序

AI 生成的内容可能不正确。

• SecondaryNamenode Web UI：http://localhost:9868

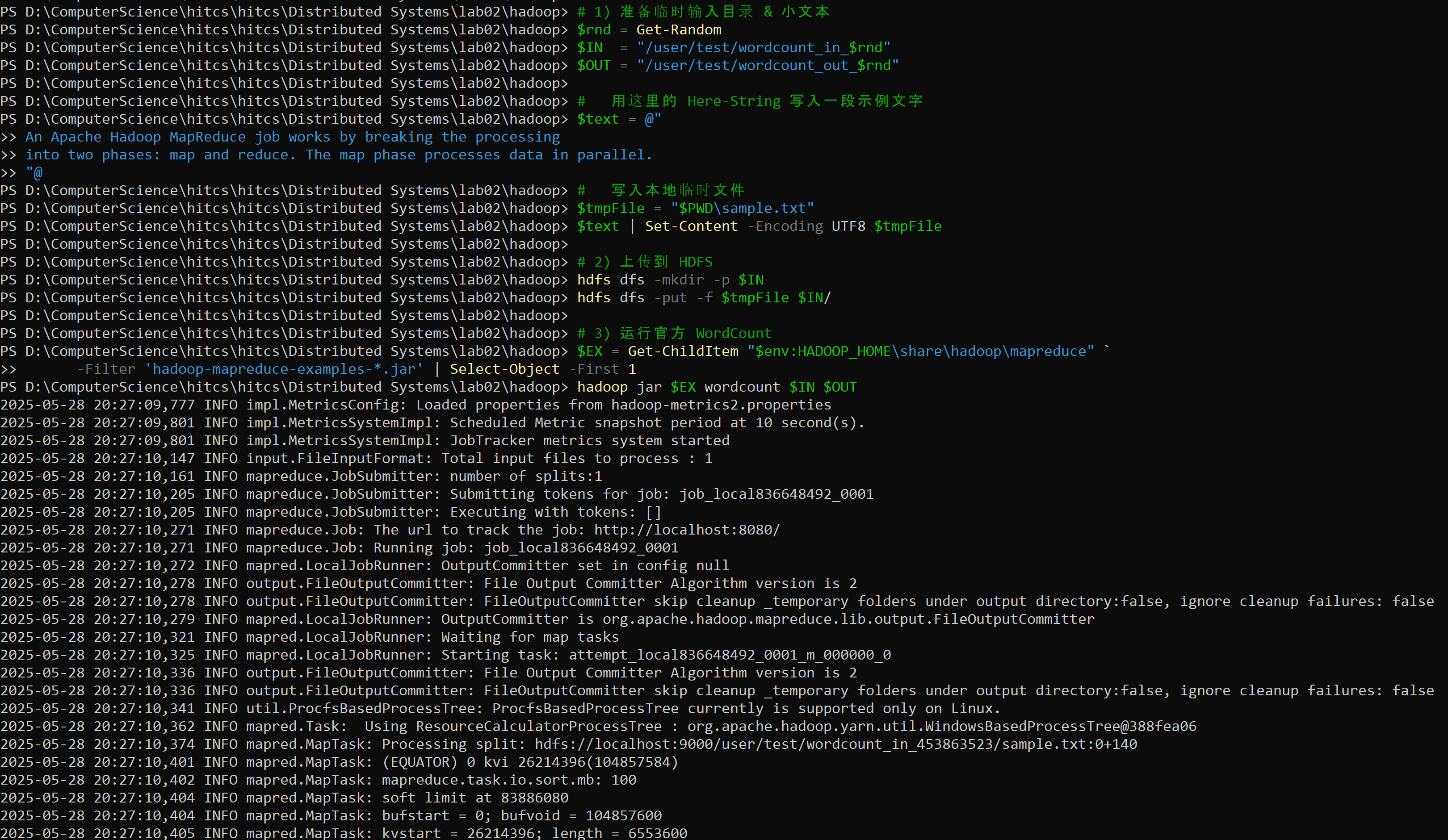


# 3.2 应用 1：WordCount 词频统计

（1）运行官方示例源代码测试并验证：

• Windows端：

脚本运行：



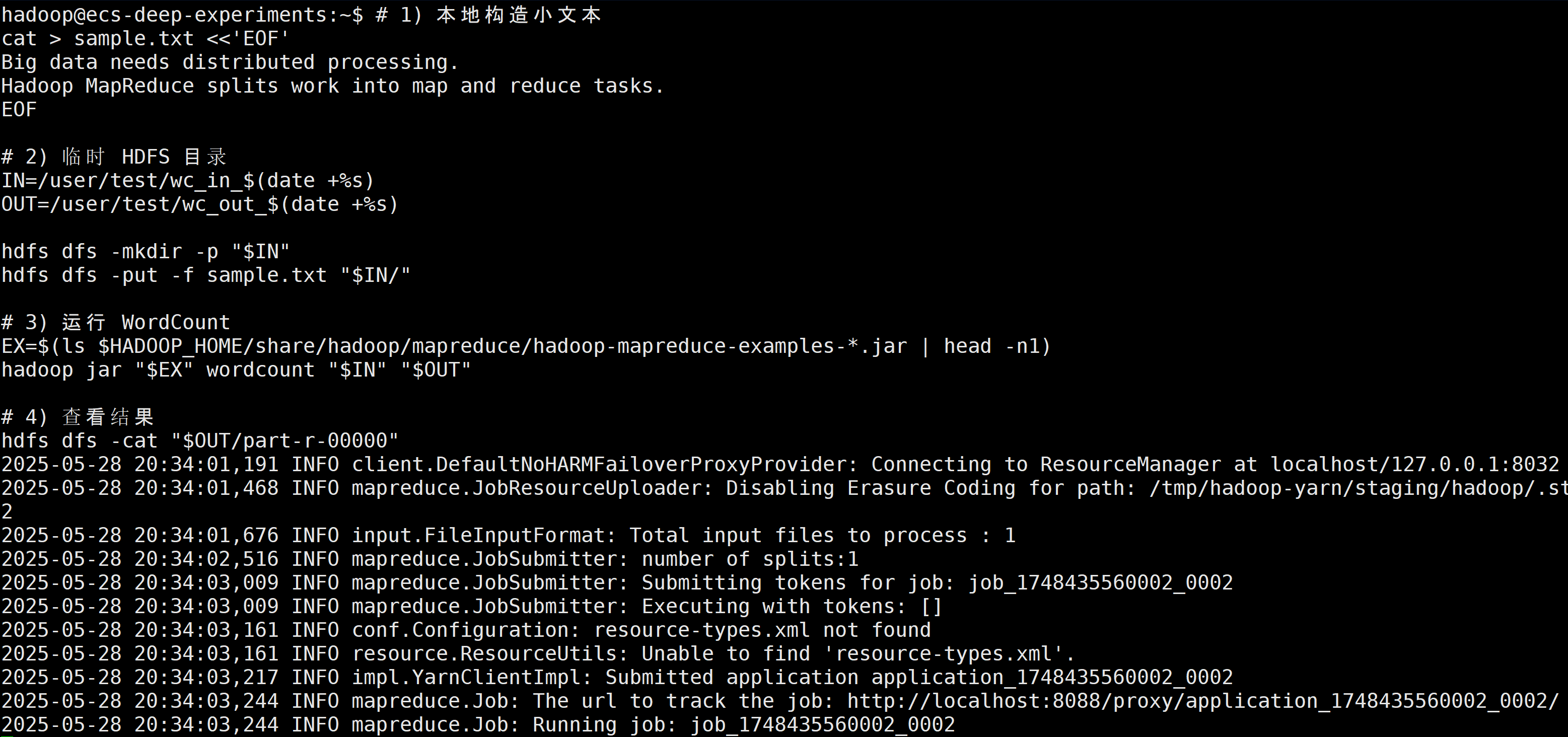
查看结果：

文本

AI 生成的内容可能不正确。

• Linux端：

脚本运行：



查看结果：

文本

AI 生成的内容可能不正确。

（2）使用脚本run\_wordcount\_hadoop处理 datasets\词频统计数据.txt：

• 上传数据 → 运行 MapReduce → 导出结果到本地 wordcount\_out.txt.

1. Windows端使用脚本 run\_wordcount\_hadoop.ps1运行词频统计测试用例：

• 脚本代码：



• 运行测试脚本：

文本

AI 生成的内容可能不正确。

文本

AI 生成的内容可能不正确。

• 测试结果前 20 行：

文本

AI 生成的内容可能不正确。

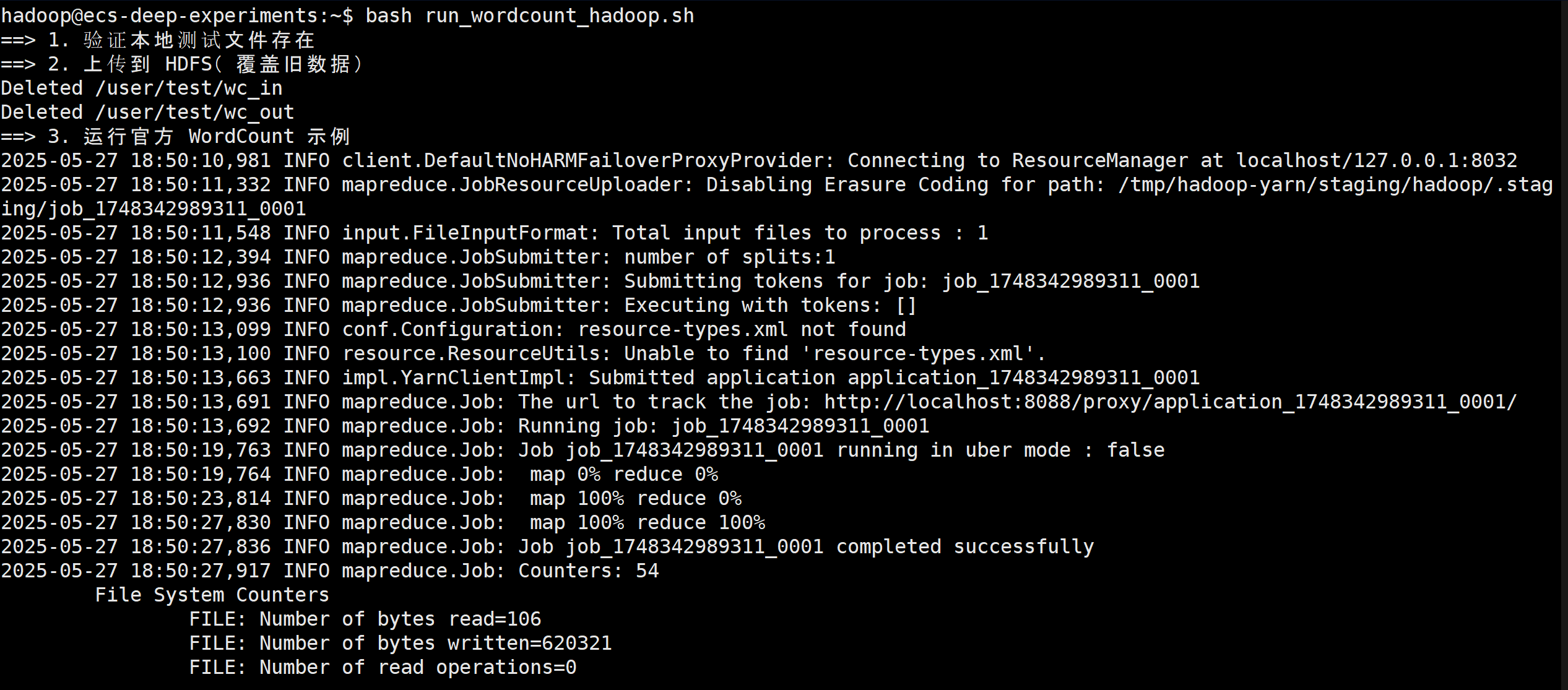
1. Linux端使用脚本 run\_wordcount\_hadoop.sh运行词频统计测试用例：

脚本代码：

**表格

AI 生成的内容可能不正确。**

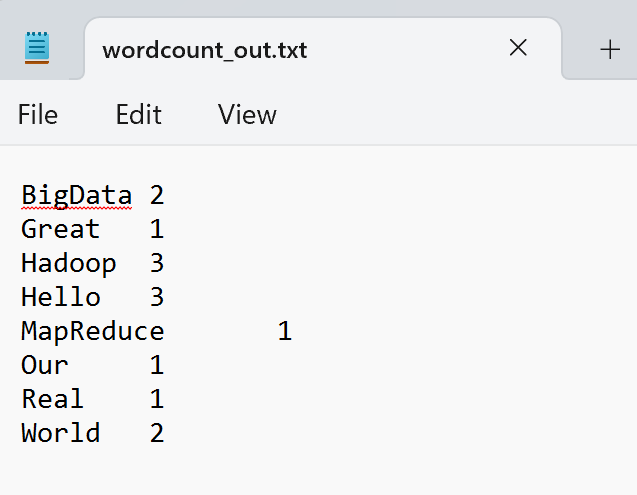
• 运行测试脚本：

****

**文本

AI 生成的内容可能不正确。**

• 测试结果文件：

****

# 3.3 应用 2：电影评分次数统计

• 数据集：ratings.dat、movies.dat（7 MB / 24 MB）。

• 编译 MoviesLab.jar，编写 MoviesJoinDriver（Map-Only）+ MoviesCountDriver（Map + Reduce）。

• 运行脚本 run\_movie\_hadoop.sh 完成自然连接 + 评分计数。

• JobHistory 运行记录：

1. 代码脚本run\_movie\_hadoop.sh：

**文本

AI 生成的内容可能不正确。**② 运行测试用例：

**文本

AI 生成的内容可能不正确。**

**文本

AI 生成的内容可能不正确。文本

AI 生成的内容可能不正确。**

③测试结果文件（部分）：

**表格

AI 生成的内容可能不正确。**

# 四、实验心得

通过本实验，成功在 Windows系统和Linux系统伪分布式环境完成 Hadoop 集群部署，并利用 MapReduce 处理结构化与文本数据。掌握了：

• NameNode 数据格式化与 DataNode 块清理技巧；

• YARN 调度与作业历史追踪；

• Map-Only 与 MapReduce 组合的两种编程范式。

不足：本机环境 IO/内存有限，大规模数据仍需多节点集群；后续可尝试 Spark 对比。

# 五、参考文献

[1] Apache Hadoop Documentation, https://hadoop.apache.org/

[2] “Hadoop: The Definitive Guide”, Tom White, O’Reilly Media.