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：

**年 月 日**

**实验名称：MapReduce分布式计算及其应用**

**实验学时：3学时**

**实验内容**

Hadoop是一个能够对大量数据进行分布式处理的软件框架，已被广泛应用到各个领域，Hadoop框架最核心的设计是分布式文件系统（HDFS）、分布式计算框架（MapReduce）和集群资源管理系统（YARN）。本实验要求实现如下内容。

1.搭建分布式Hadoop集群环境，采用伪分布式或完全分布式模式。

2.应用MapReduce模型，实现词频统计。

3.应用MapReduce模型，对电影数据分析，实现用户喜好评定。**（选作）**

**实验要求**

1.Hadoop环境搭建完毕后，通过命令及Web访问方式查看集群的启动情况，附程序运行和结果截图。

2.应用1：将Hadoop官方提供的WordCount程序示例源码分别在本机（windows端）和hadoop平台下运行验证，然后将“词频统计数据”文档作为输入词频统计，给出实验步骤（附程序运行和结果截图）。

3.应用2：电影网站提供了相关的2份数据，分别为用户对电影的评分数据（ratings.dat）、电影信息数据（movies.dat），使用MapReduce，对2份影评数据进行统计分析，依据每部电影的**评价次数**来分析用户的观影喜好。

两份数据的介绍说明如下：

（1）用户对电影的评分数据ratings.dat包含4个字段，即UserID（用户ID）、MovieID（电影ID）、Rating（评分）和Timestamp（时间戳），如右表。其中UserID的范围是1~6040，MovieID的范围是1~3952，Rating采用5分好评制度，即最高分为5分，最低分为1分。



（2）电影信息数据movies.dat包含3个数据字段，分别为MovieID（电影ID）、MovieName（电影名称）和Genres（电影类型），如下表。数据中共记录18种电影类型，包括喜剧片、动作片、爱情片等类型。



**实验说明**

（1）上传内容：实验报告；运行结果文件；程序源文件。

（2）上传格式：打包为压缩文件，命名学号-姓名-实验编号。

（3）所需软件下载地址：

linux端软件 https://pan.hit.edu.cn/l/AFKshs

Windows端软件 https://pan.hit.edu.cn/l/8FsYgc

实验数据集 https://pan.hit.edu.cn/l/EFOOVU

（3）报告上传地址：ftp://10.70.25.227/(按照上课时间选择文件夹)。。

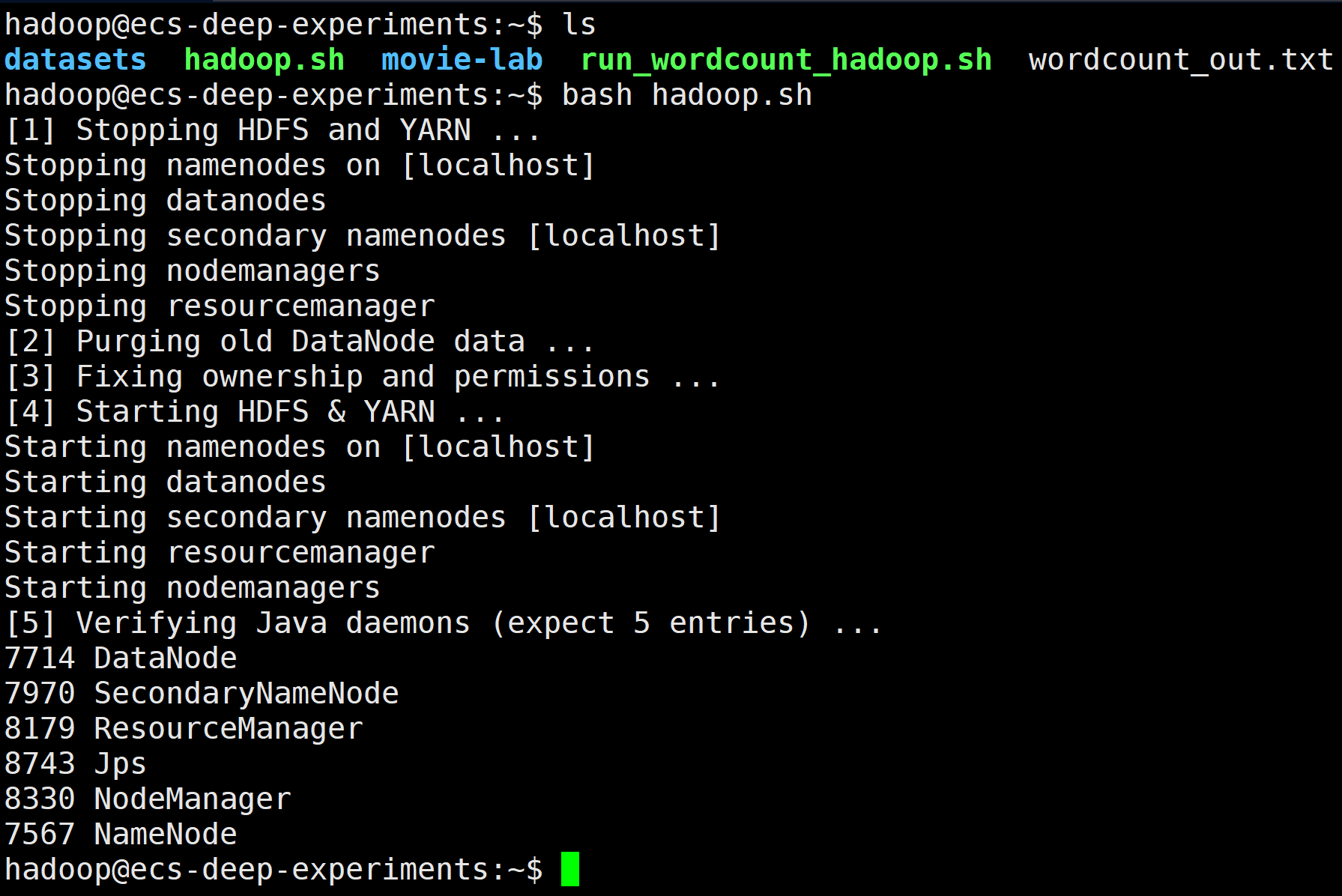
**实验提示**

实验报告及内容提示，见QQ群文件。

**个人建议**

**文本

AI 生成的内容可能不正确。**

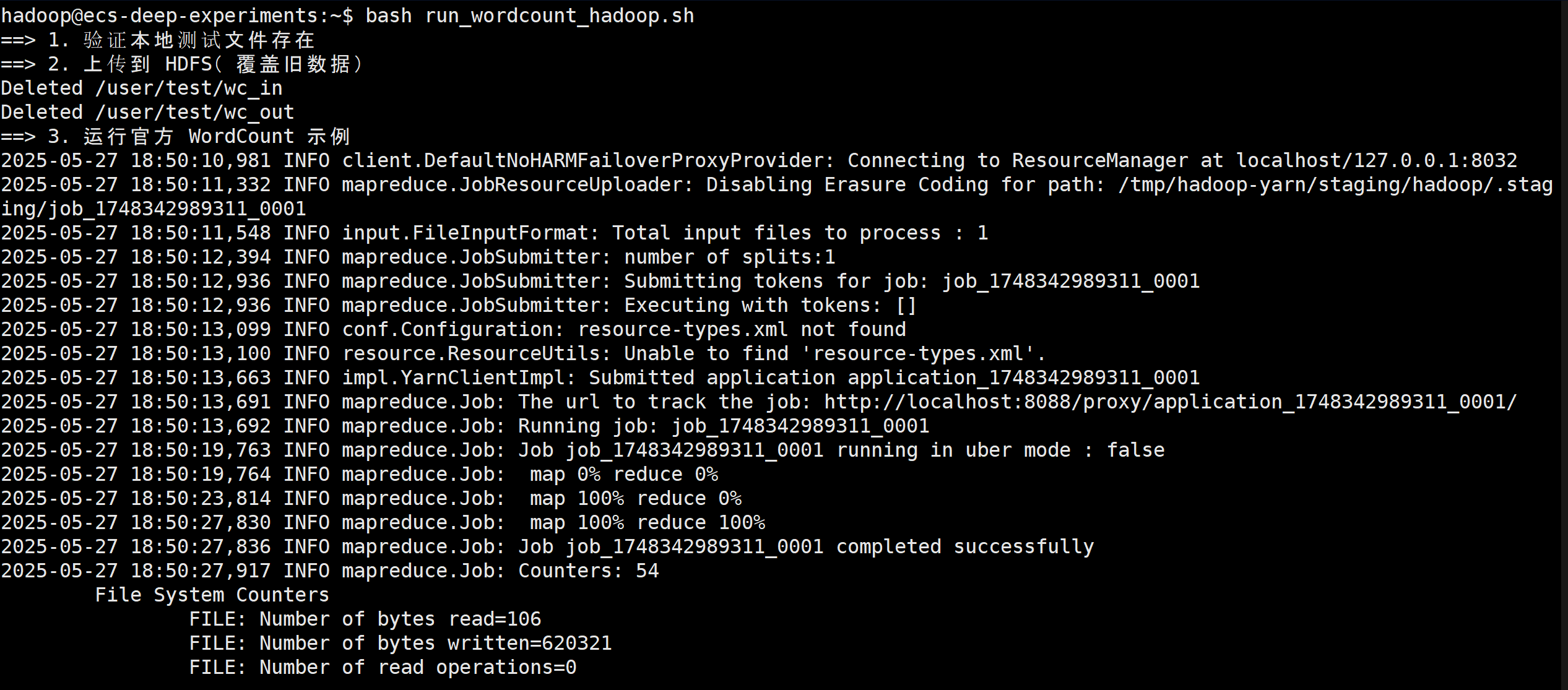


文本

AI 生成的内容可能不正确。

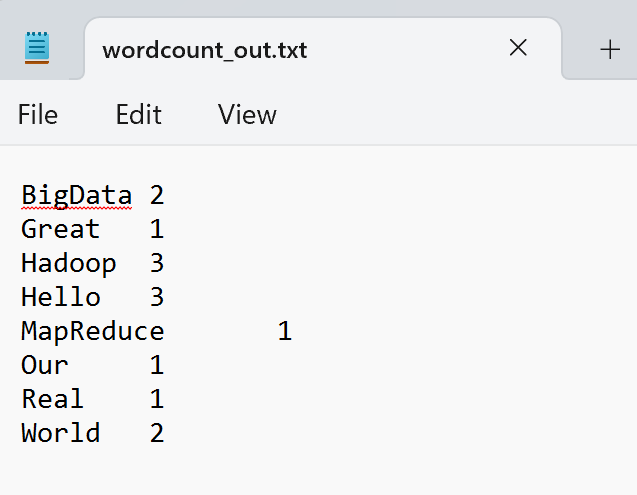
**表格

AI 生成的内容可能不正确。**

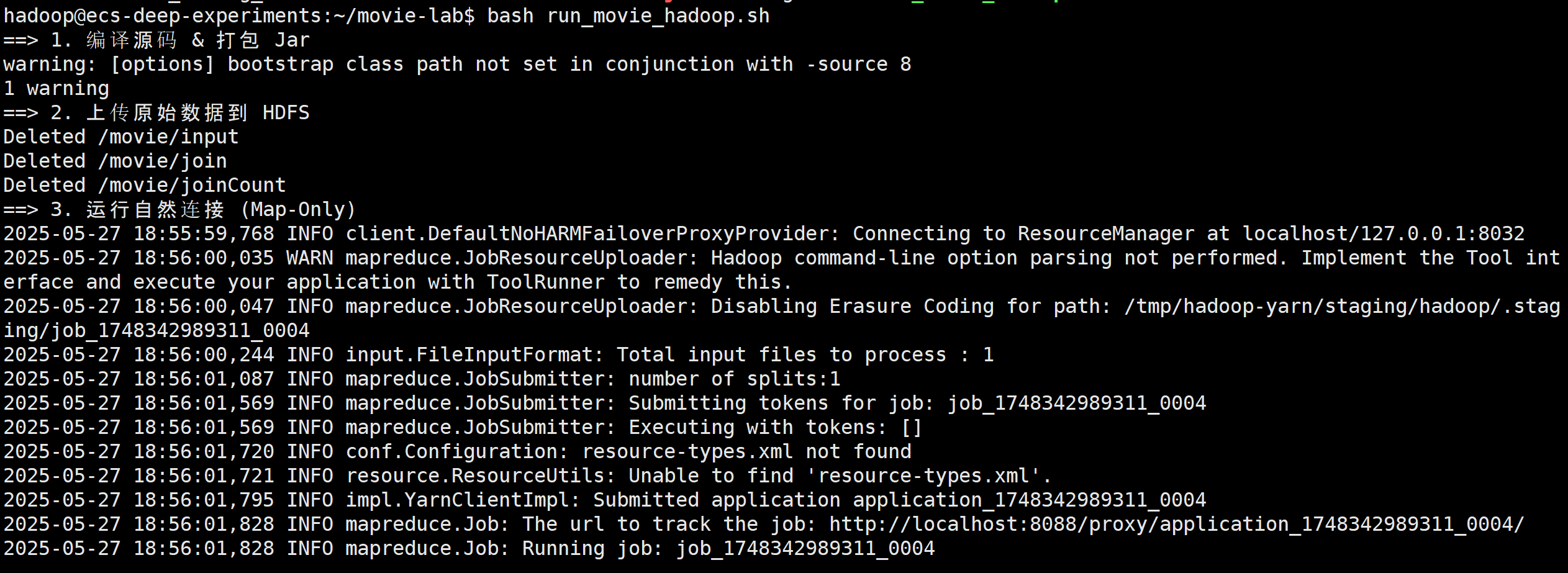
****

**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。文本

AI 生成的内容可能不正确。**

**表格

AI 生成的内容可能不正确。**