

（深圳）

实验报告

开课学期： 2024秋季

课程名称： 大数据导论

实验名称：

实验性质： 设计型

实验学时： 2 地点： T2608

学生班级： 22级5班

学生学号： 220110519

学生姓名： 邢瑞龙

评阅教师：

报告成绩：

实验与创新实践教育中心制

2024年9月

# 1. 实验目的

1. 熟悉Hadoop分布式集群的配置方法和基本操作；
2. 理解MapReduce的基本原理和框架；
3. 掌握MapReduce的基础编程方法，操作和运行Mapreduce作业。

# 2. 实验内容

1．在大数据教学管理平台完成Hadoop完全分布式集群搭建；

2. 在IntelliJ IDEA中创建MapReduce工程，编码解决以下3个问题：

（1）获取词频统计Top 20 关键词

（2）获取成绩表的最高分记录

（3）统计网站每日的访问次数

# 3. 实验环境

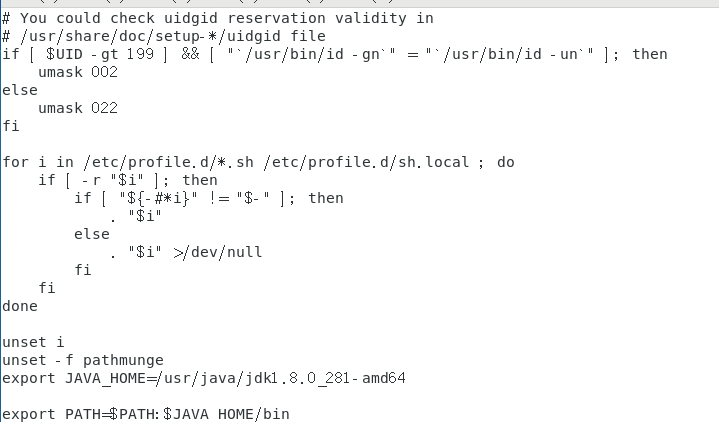
* CentOS 7.9
* JDK 1.8
* Hadoop3.1.4
* IntelliJ IDEA 2022.2

# 实验过程及结果

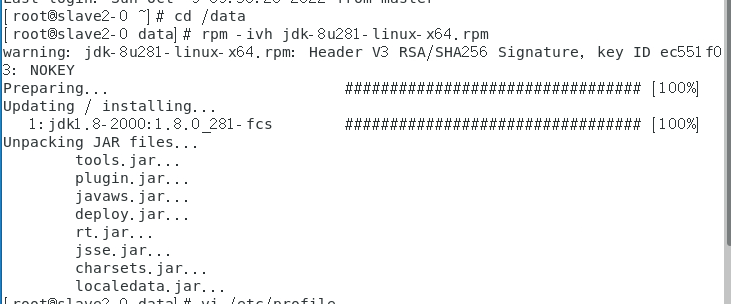
## 4.1 Hadoop 集群环境搭建

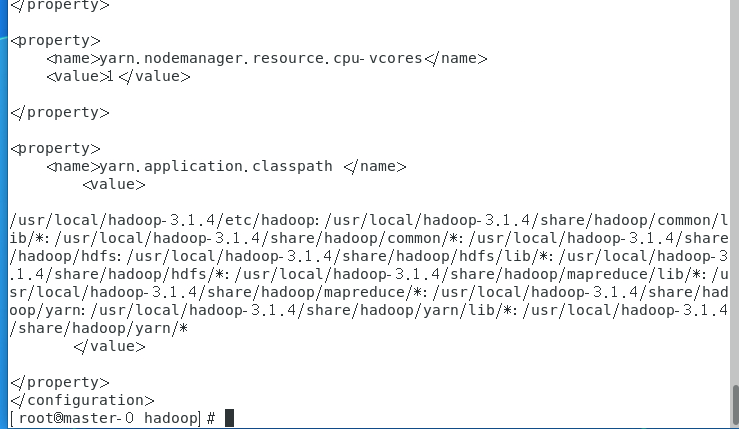
*Hadoop分布式环境搭建过程中的关键步骤截图，如命令运行结果，修改后的配置文件（使用cat命令查看）等。词频统计的结果文件需提交。*

为etc\profile添加java的环境变量：

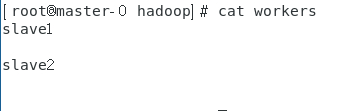


安装JDK:

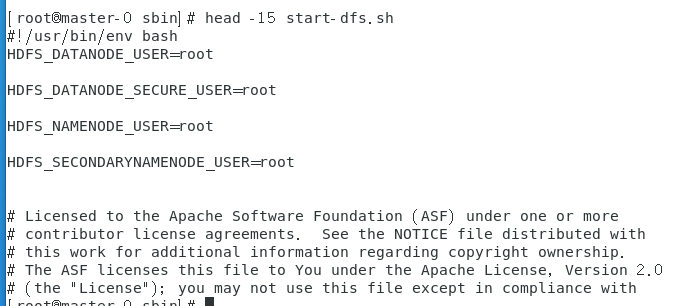
修改Hadoop的配置文件：（图为yarn-site.xml实例）：

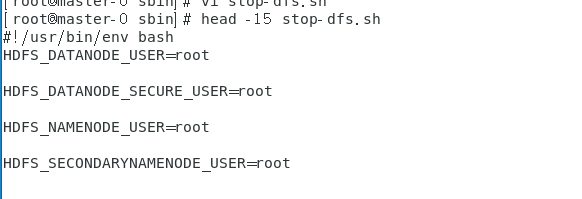


修改workers文件：

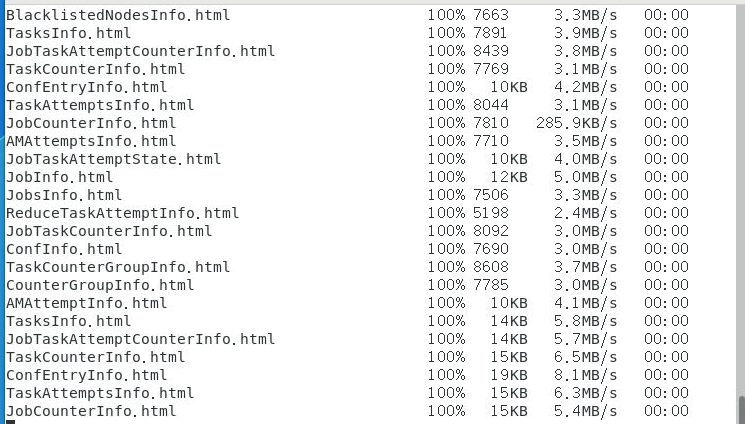


修改HDFS的启动脚本start-dfs.sh与停止脚本stop-dfs.sh：

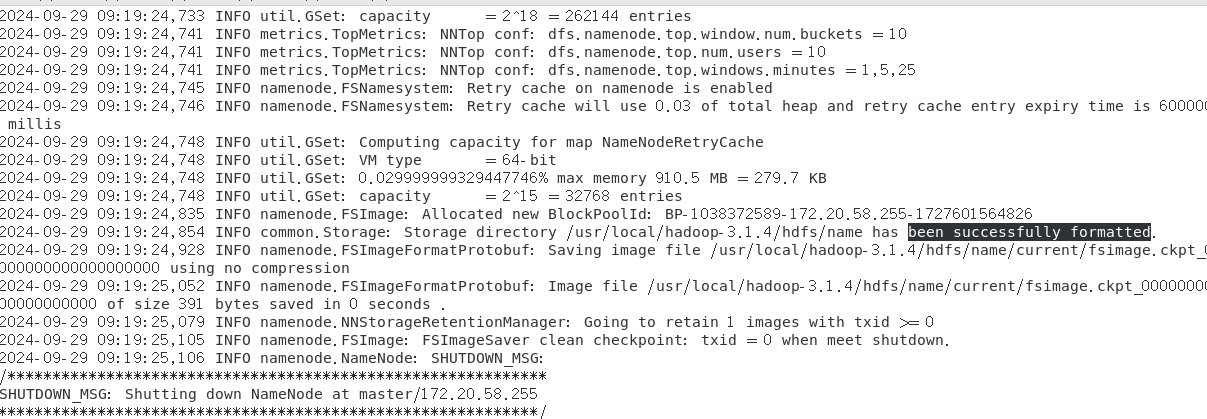




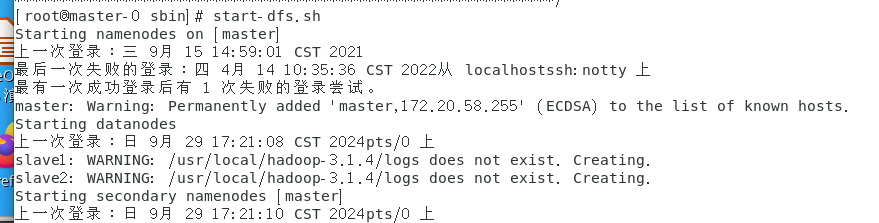
master节点已经部署好的Hadoop与/etc/profile文件复制传输到slave1、slave2节点：

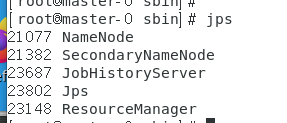


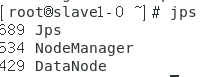
格式化成功：

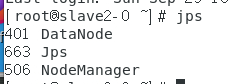


启动集群：

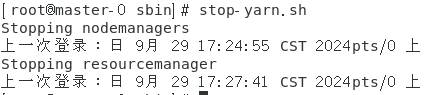


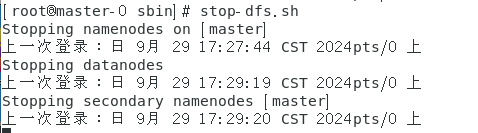




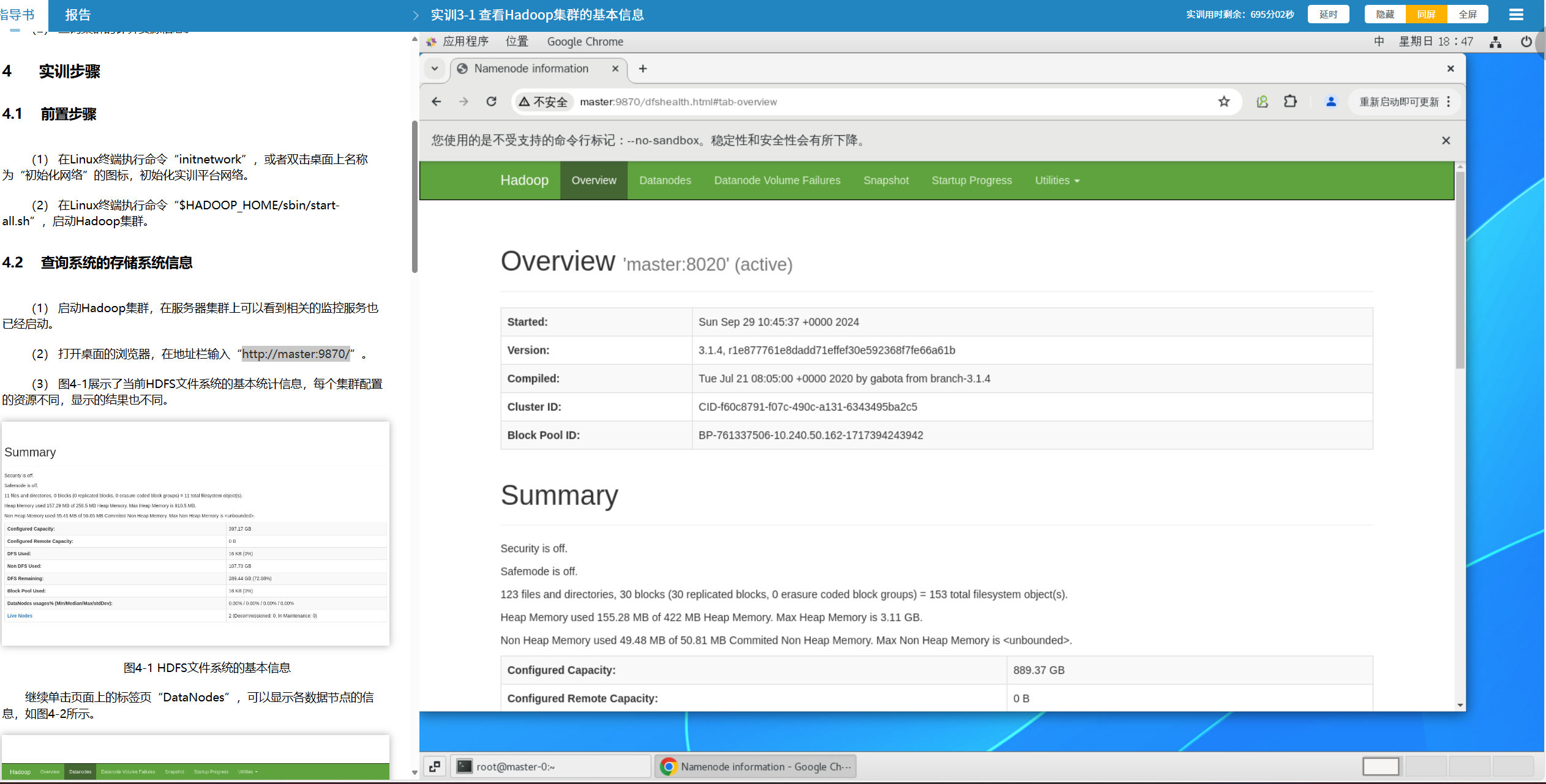


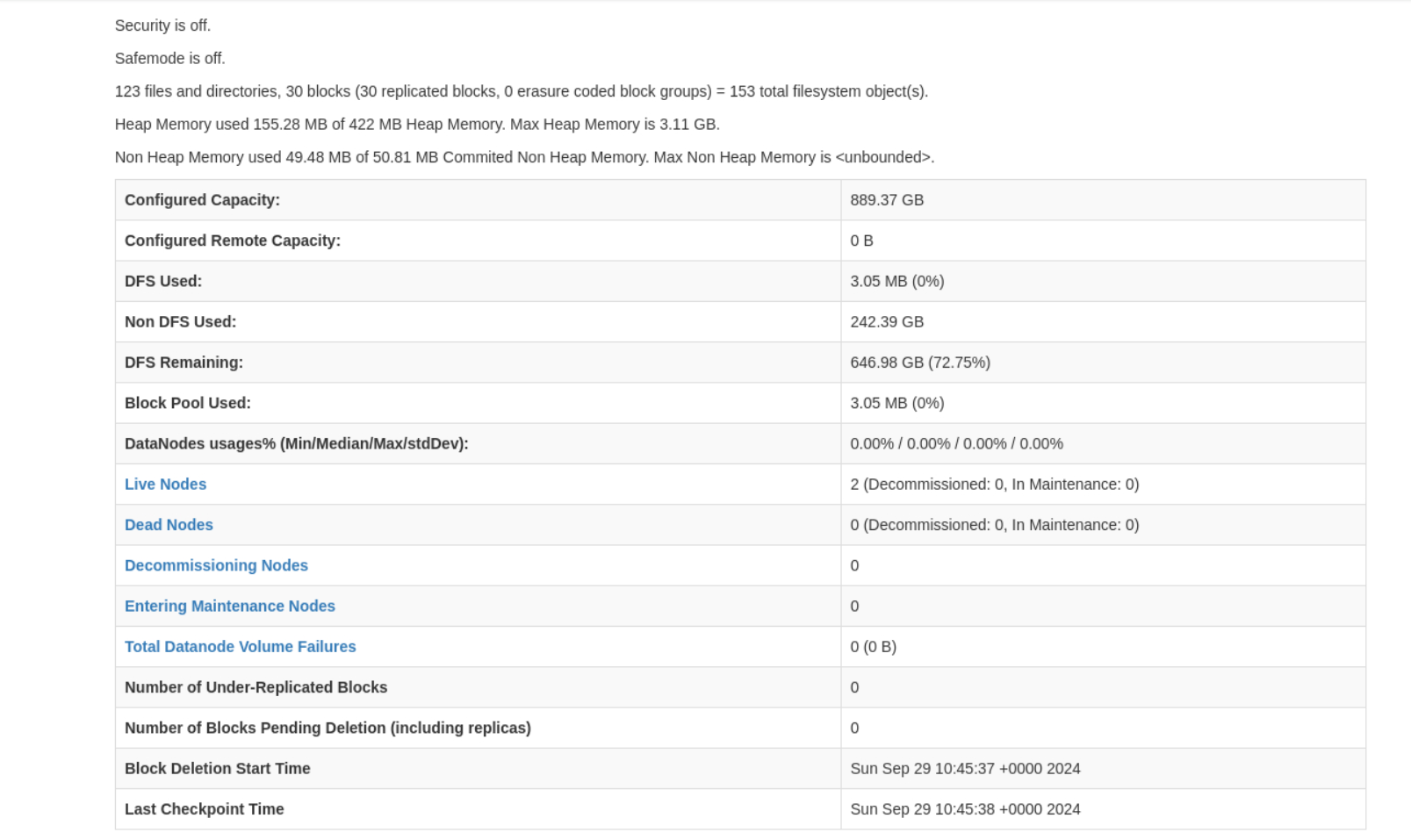
关闭集群：



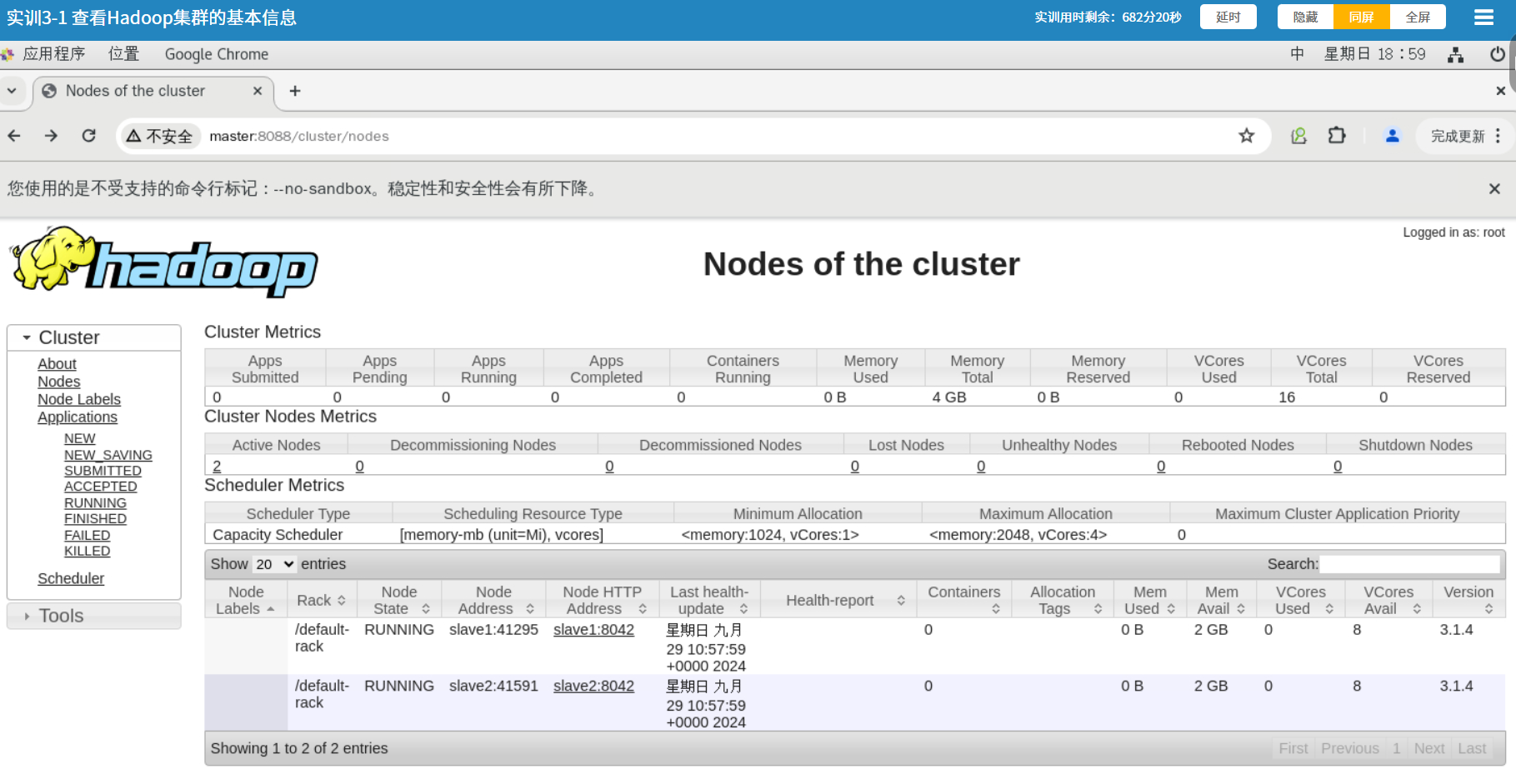


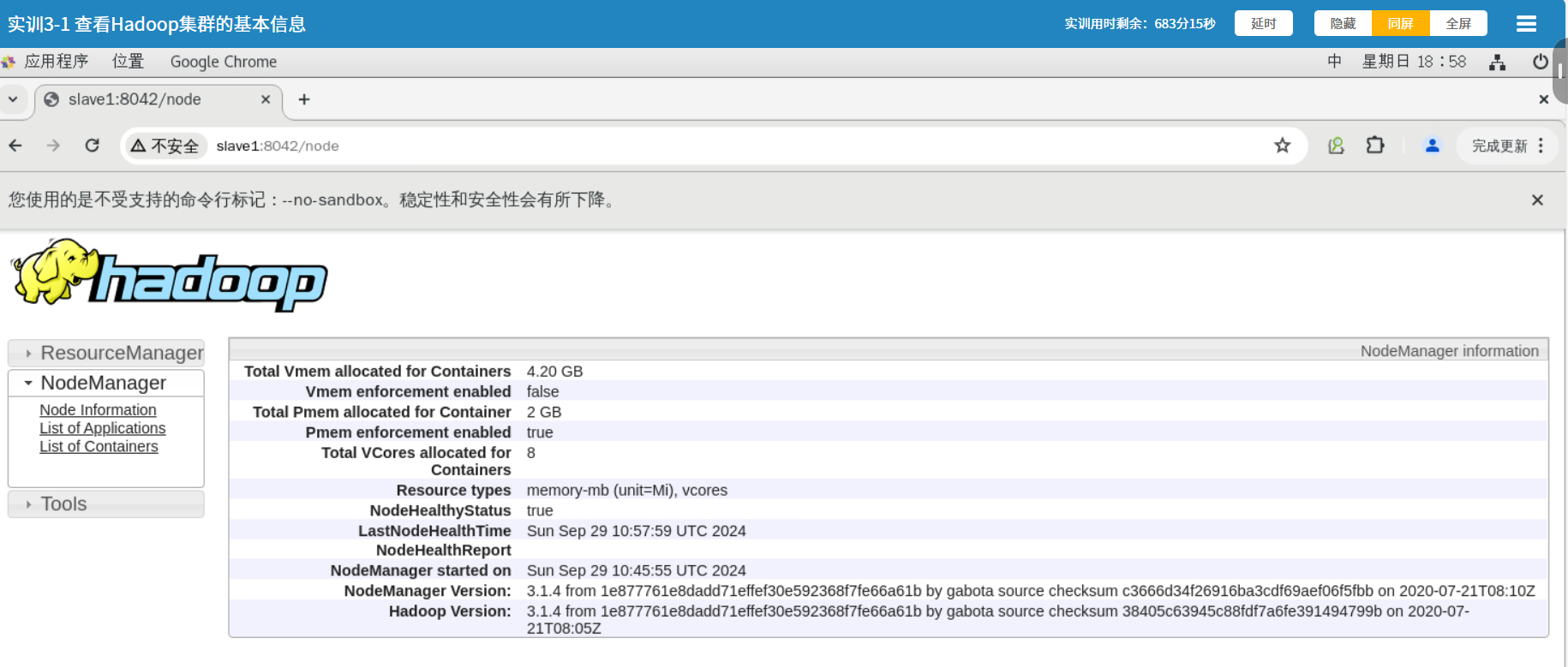
查看集群信息：





查看hadoop节点的计算资源：



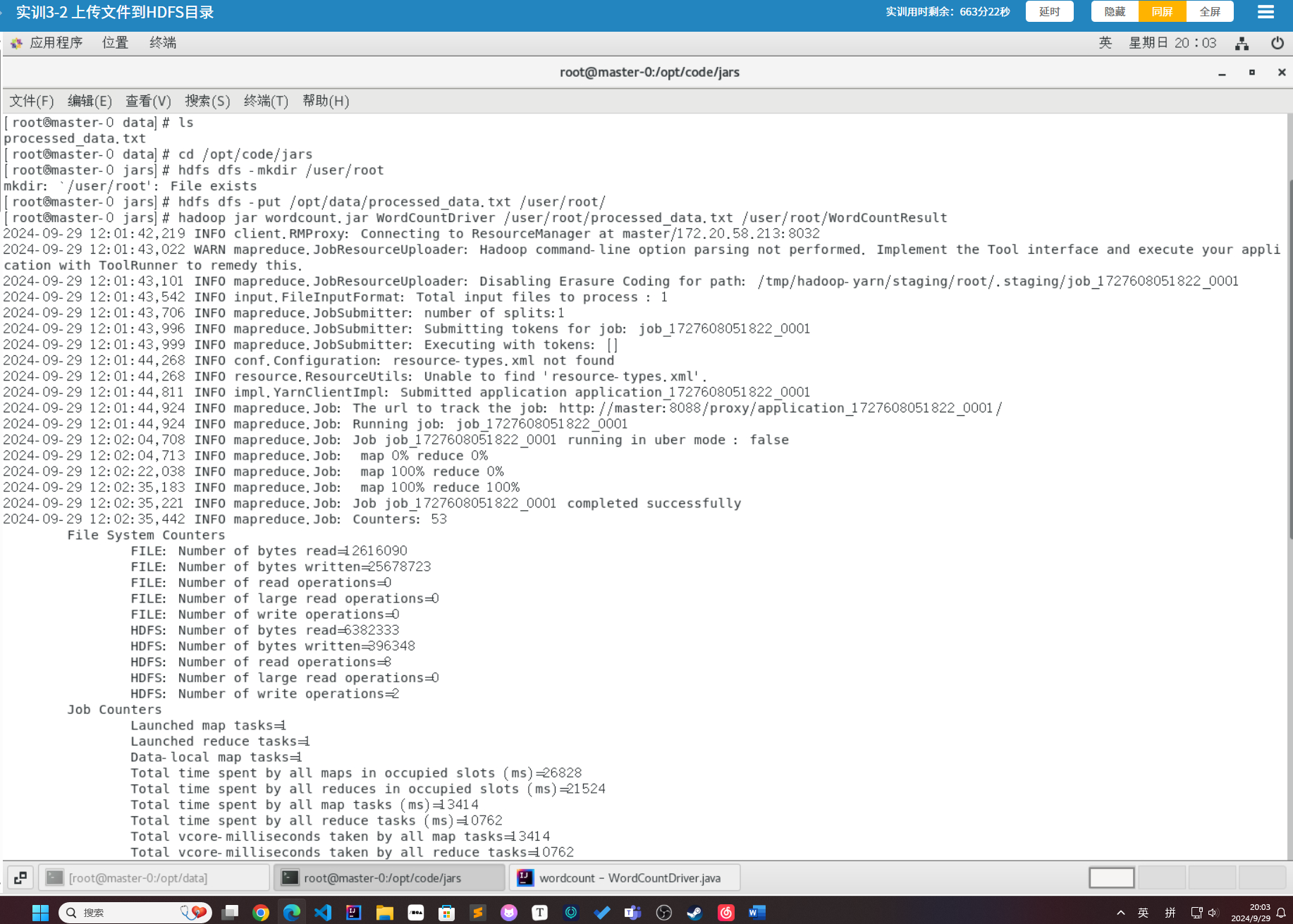


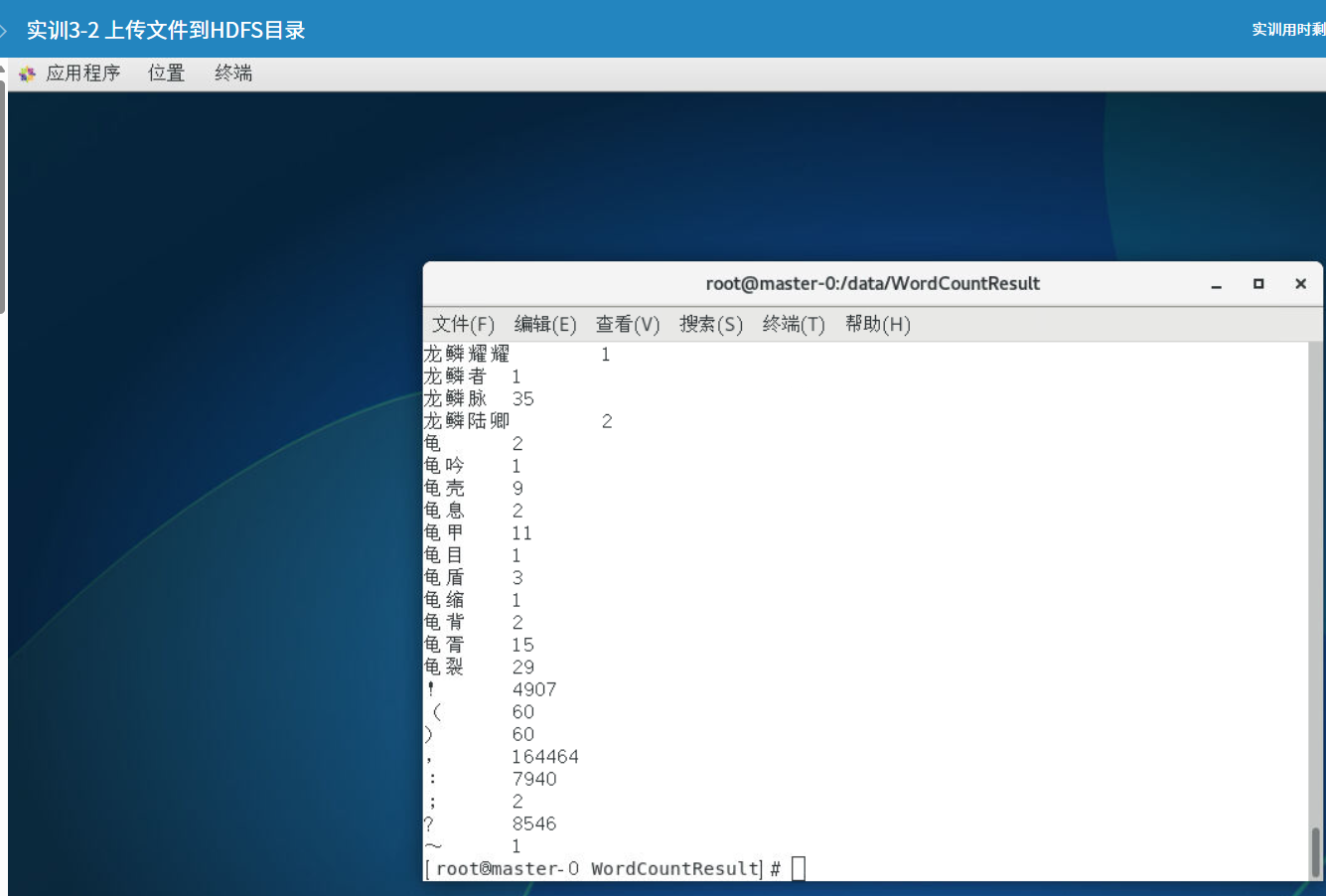


## 4.2 MapReduce编程

*针对以下每个问题，简单描述你的Mapper和Reducer模块的处理逻辑，并截图部分运行结果。每个项目Mapper，Reducer和Driver模块的代码文件（\*.java）以及完整的运行结果也需在作业平台提交。*

统计词频：





1. 获取词频统计Top 20 关键词

Map按行处理每个单词的出现次数。

Reduce按一个key一个key的来处理，以单词作为key值，汇总当个key对应的所有value值求和便是该单词出现的总次数。

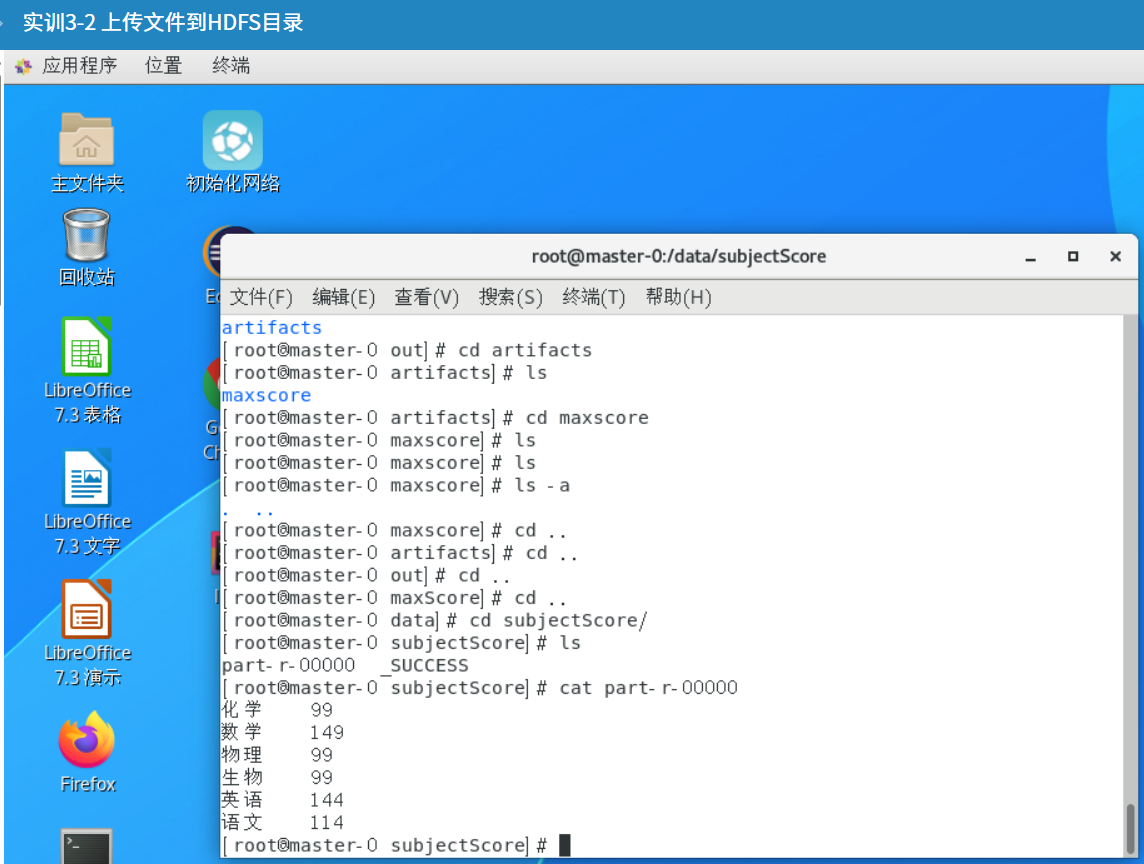
最后为了得到前20个词频最高的单词，在reduce中重写clean\_up,使用优先队列的方式，优先队列从头到尾词频依次递增。若当前队列长度>20，则弹出队首。



1. 获取成绩表的最高分记录

Map按行处理，一次只处理一条记录，以课程作为key，成绩为其value。

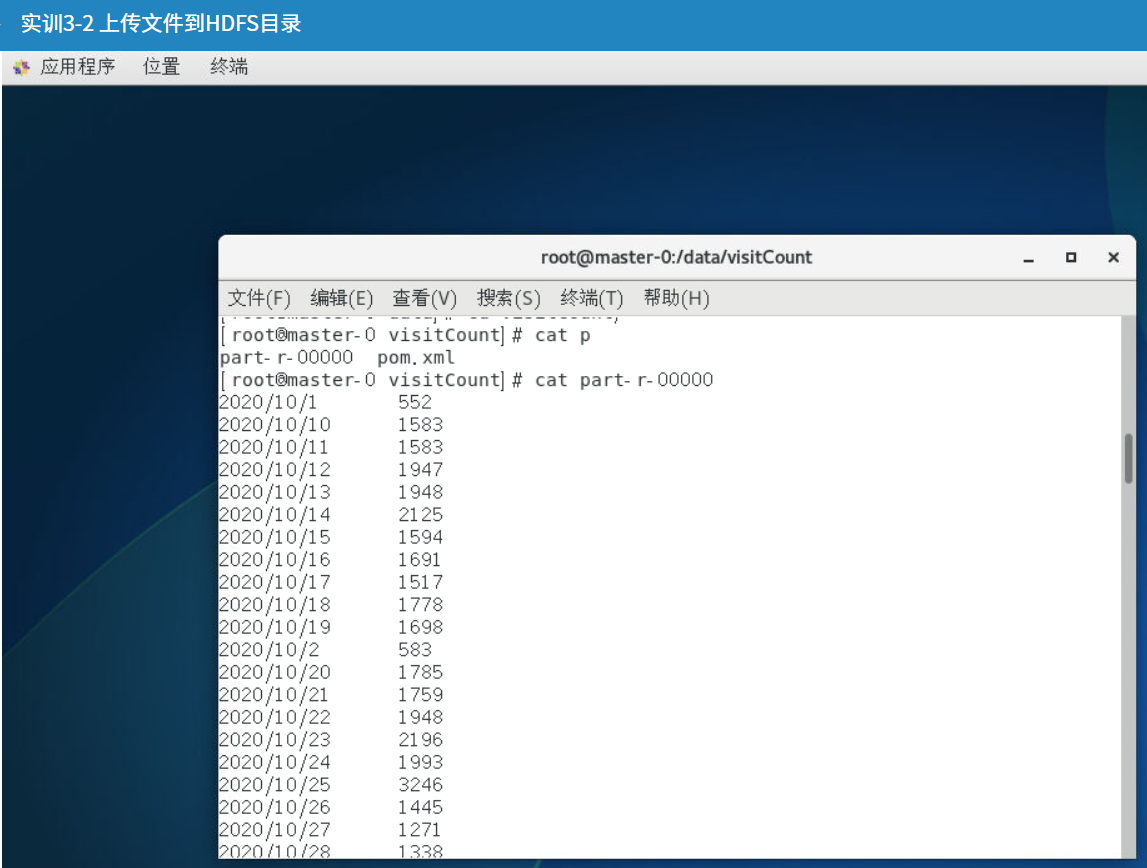
Reduce按key处理，取当前key中value值最高的，则为该门课程的最高成绩。



（3）统计网站每日的访问次数

Map按行处理，一次只处理一条记录，将第五列中的字符串以空格作为分隔符得到访问日期，将此作为key，value为1对应一次访问。

Reduce按key处理，累计value出现的个数，便是当前日期对应的访问次数。



个人签名 ： 邢瑞龙

2024年 9月 29日