

（深圳）

实验报告

开课学期： 2022秋季

课程名称： 大数据导论

实验名称：Hadoop环境配置与基本操作

实验性质： 设计型

实验学时： 2 地点： T2608

学生班级： 20级08班

学生学号： 200210231

学生姓名： 王木一

评阅教师：

报告成绩：

实验与创新实践教育中心制

2022年9月

# 实验目的

Hadoop环境配置与基本操作：

1. 掌握Hadoop分布式环境的配置方法；
2. 理解Mapreduce作业的原理和操作方法。

# 实验环境

CentOS 6.10

Hadoop 2.6.4

Java 1.8

# 实验内容

本次实验在“大数据教学管理平台”进行，对应“Hadoop大数据技术基础”课程实训1-4

具体内容：

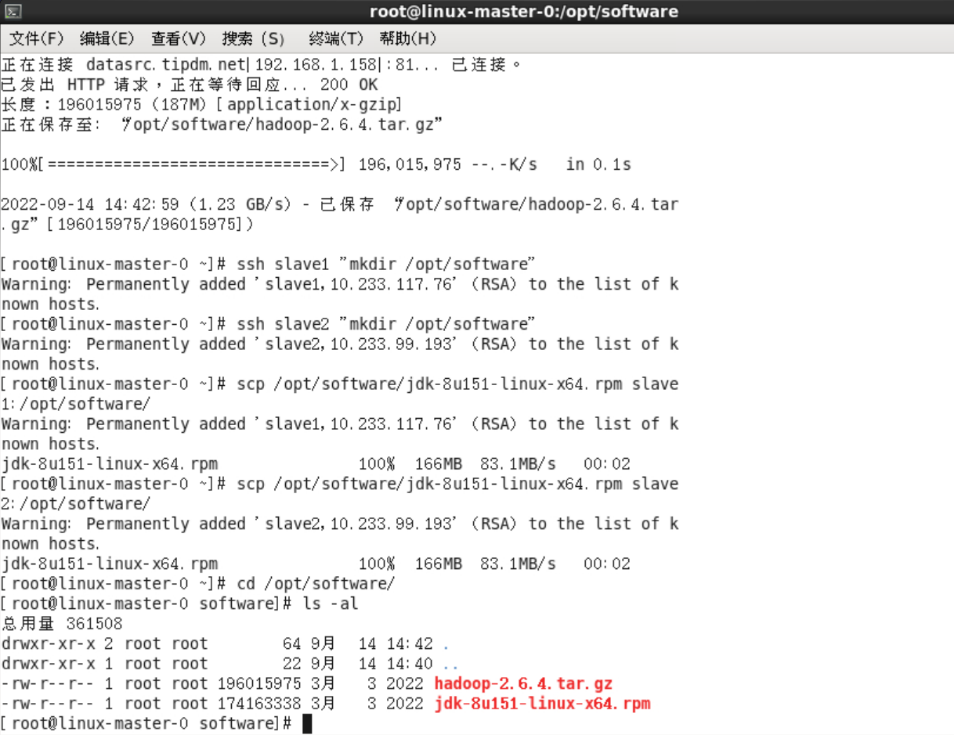
1. 独立搭建Hadoop分布式环境
2. 上传预处理后的实验文本数据至HDFS
3. 使用MapReduce实现WordCount程序任务

# 实验过程

*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。*

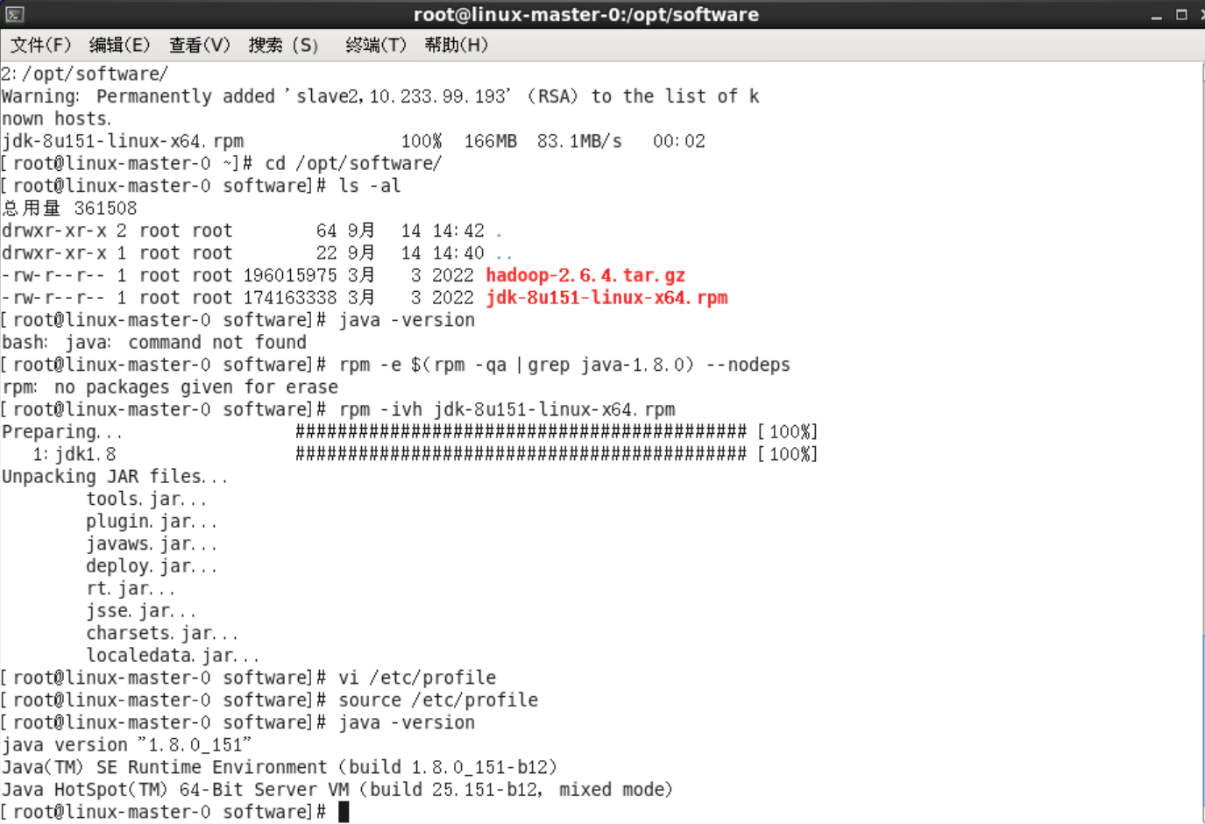
1. Hadoop分布式环境的搭建

1.1 Hadoop&Java的下载和Java的分发

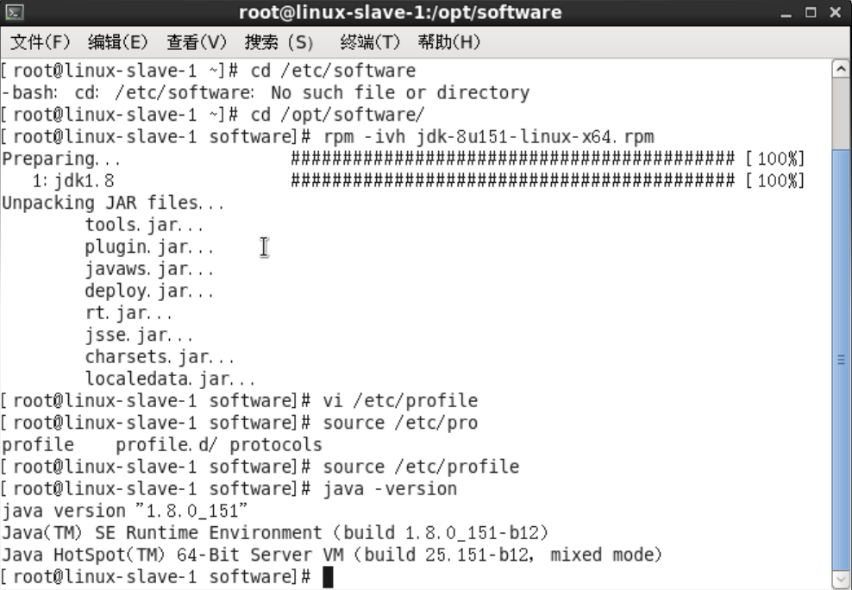


/opt/software/目录下已有hadoop与java的安装包，同时java安装包通过ssh发送给slave1，slave2

1.2 Java安装



在master节点安装java1.8，同时添加环境变量并使其生效



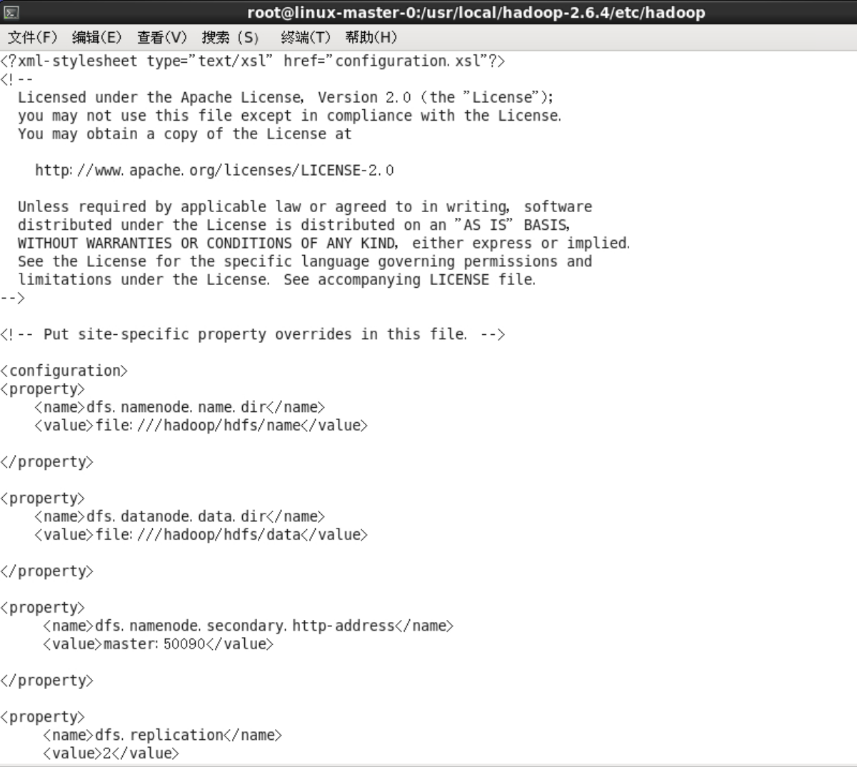
在slave节点安装java，也需添加环境变量并让其生效

1.3 修改Hadoop配置文件

在master节点安装hadoop， 同时进入安装地址目录下修改有关配置文件

|  |  |
| --- | --- |
| hadoop-env.sh | Hadoop运行环境配置 |
| yarn-env.sh | YARN运行环境配置 |
| core-site.xml | 核心配置 |
| hdfs-site.xml | HDFS配置 |
| yarn-site.xml | YARN配置 |
| Mapred-site.xml | MapReduce配置 |
| slaves | datanode与datamanager启动节点配置 |
| /etc/profile | Hadoop环境变量添加 |

下图为hdfs-site.xml配置结果



1.4 启动集群

包括以下步骤：

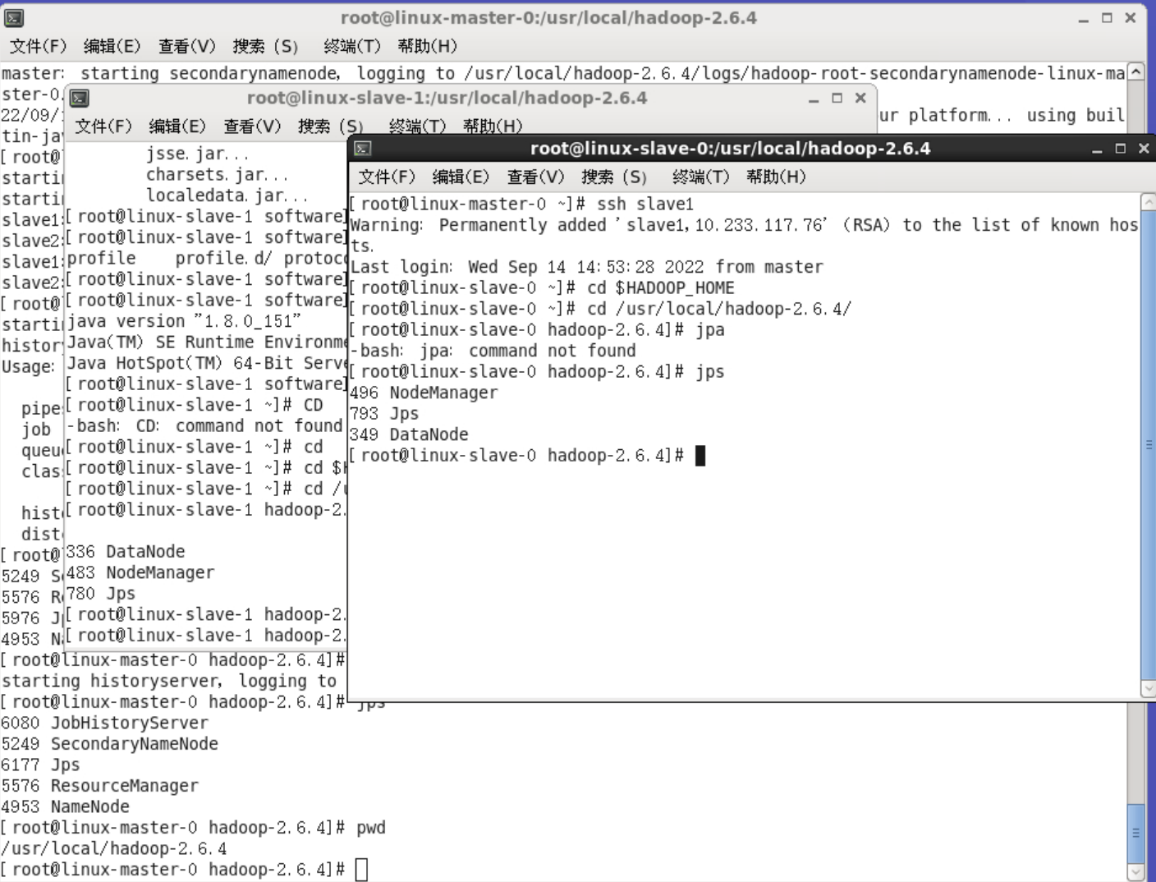
·配置master环境变量

·配置slaves环境变量

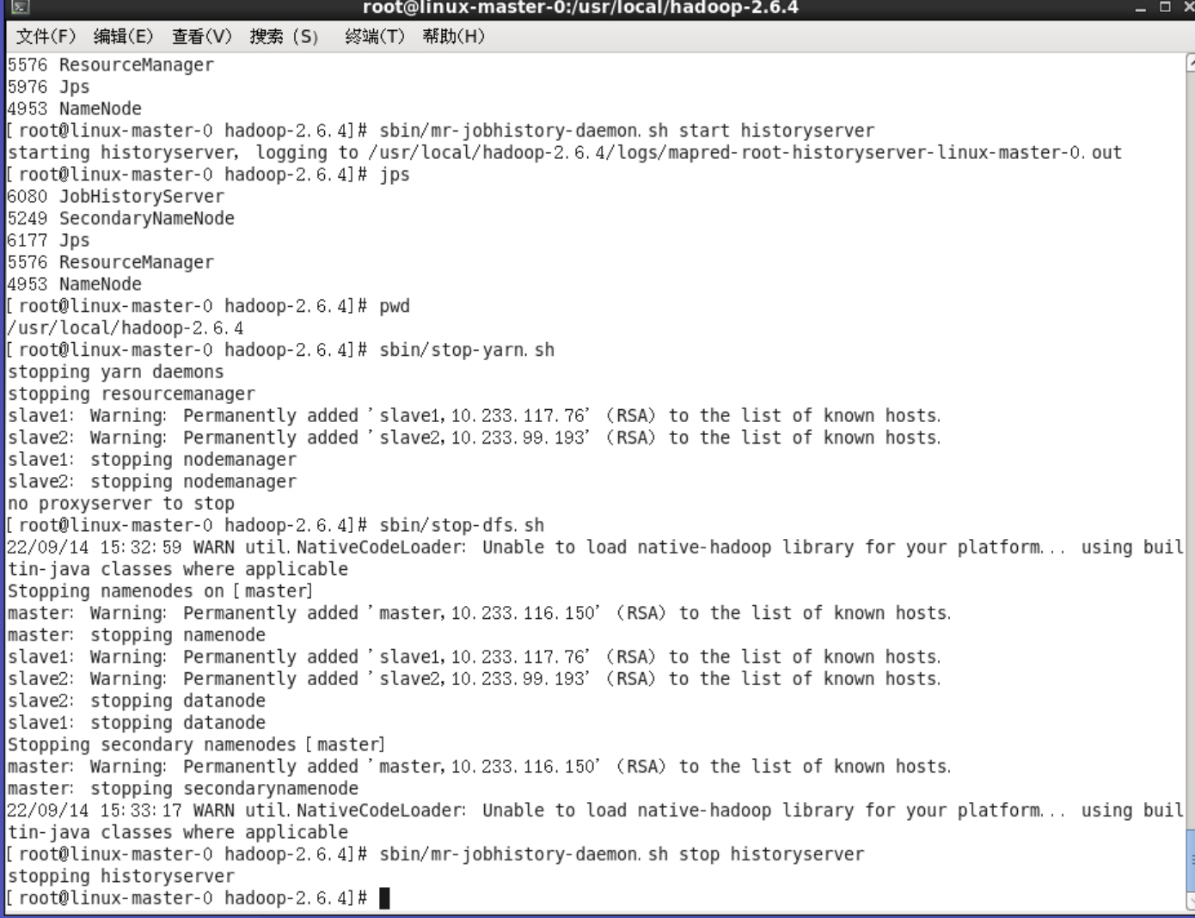
·格式化namenode

·启动集群

集群启动成功后，使用在不同节点使用jps命令可得以下信息：

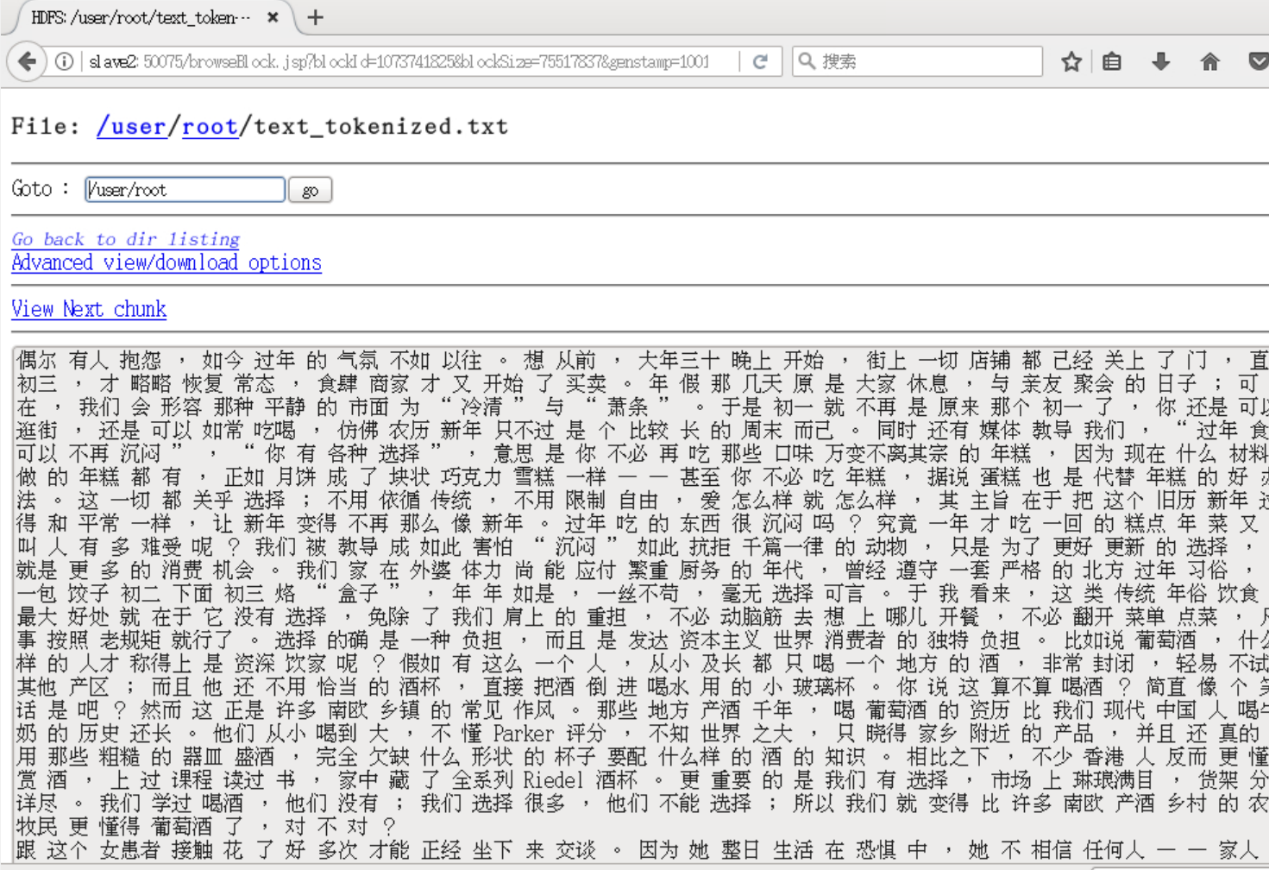


1.5 关闭集群



1. 使用MapReduce实现WordCount，对作业1爬取的网页进行词频统计
   1. 启动hadoop集群
   2. 上传文件至HDFS

文件上传成功后，可在浏览器进入<http://master:50070/nn_browsedfscontent.jsp>（具体地址由之前步骤配置决定）查看



* 1. 执行词频统计

使用hadoop自带的统计功能进行计算

统计结果见“五”

# 实验结果

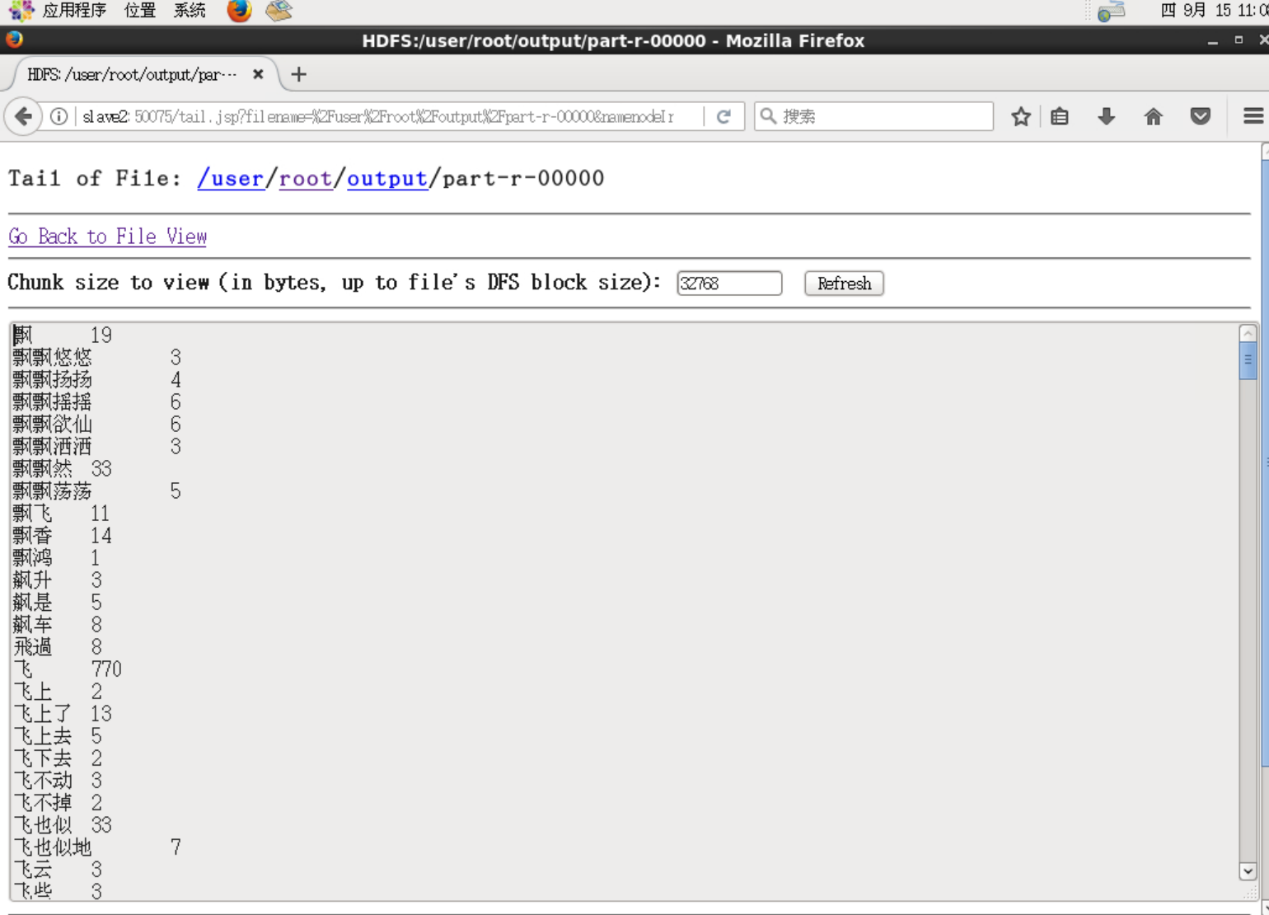
*执行WordCount程序任务的运行结果，可截图运行命令、部分词频统计结果。*

任务一：Hadoop环境搭建

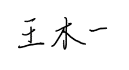
具体结果已在上方“四.1.4”体现，在不同节点使用jps均可显示相关配置信息

任务二：词频统计

具体结果也可在浏览器中查看（或使用命令行）， 如下图



（分词时并未去除标点符号，故统计结果的前面部分为符号的统计，若只是执行词频统计任务，将不可用数据提出即可）

个人签名 ： 

2022年 9 月 19日