云南大学软件学院

实 验 报 告

课程： 大数据分析及应用实验 任课教师： 何婧

专业： 人工智能 学号： 20211060245 姓名： 陈俊宏 成绩：

专业： 人工智能 学号： 20211120171 姓名： 孔令高 成绩：

# 实验一 Hadoop环境搭建

## 一、 实验目的

1．掌握Ubuntu系统的安装与Linux的基本操作；

2．搭建Hadoop伪分布式平台；

3．掌握Hadoop的常用操作。

4．熟悉常用的HDFS操作。

5．熟悉MapReduce的基本原理和应用。

## 二、 实验内容及步骤

### 2.1 Ubuntu系统的安装

（描述安装的主要步骤和主要命令的使用）

1.下载并安装VMware workstation 17 pro；

2.下载ubuntu-18.04.6-desktop-amd64.iso系统镜像文件；

3.打开VMware，创建新虚拟机；

4.选择对应的镜像文件，选择内存、处理器、外存等硬件配置，安装镜像文件；

5.完成Ubuntu 64-18.04.6的安装



### 2.2 Hadoop的安装及伪分布式配置

1.创建Hadoop用户；

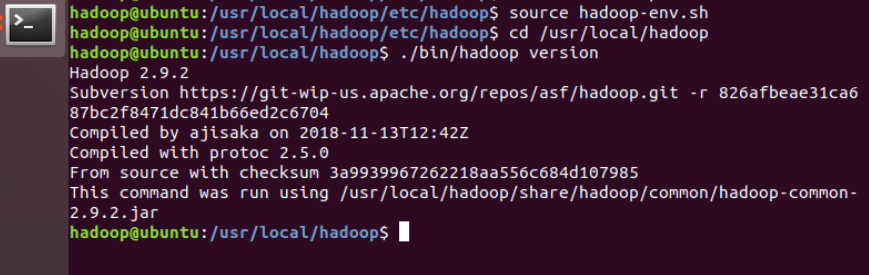
2.更新Ubuntu系统相关组件并安装vim编辑器；

3.配置SSH，并设置无密码登录；

4.安装Java并配置环境变量；

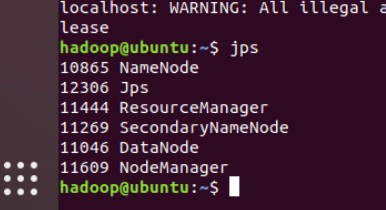
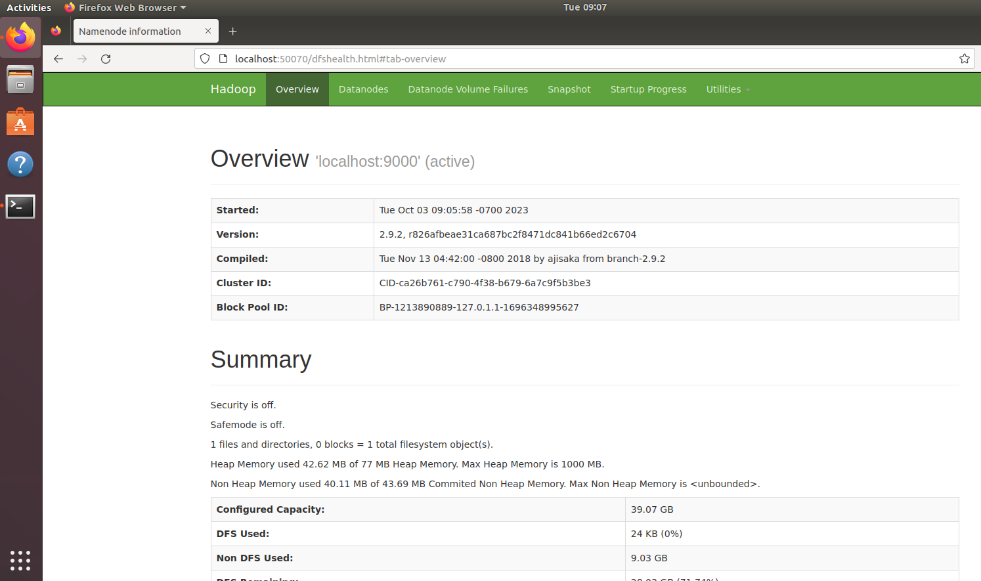
5.到<https://archive.apache.org/dist/hadoop/common/hadoop-2.9.2/>下载编译好的hadoop；

6.配置hadoop环境（可以调用版本号查看版本信息）；



7.配置core-site.xml，hdfs-site.xml，mapred-site.xml，yarn-site.xml四个文件，以达到伪分布式的方式；

8.格式化并启动，通过jps查看六个进程

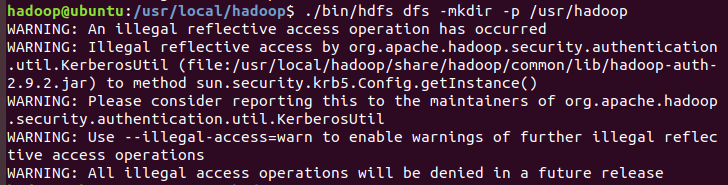


（简述安装及环境配置过程）

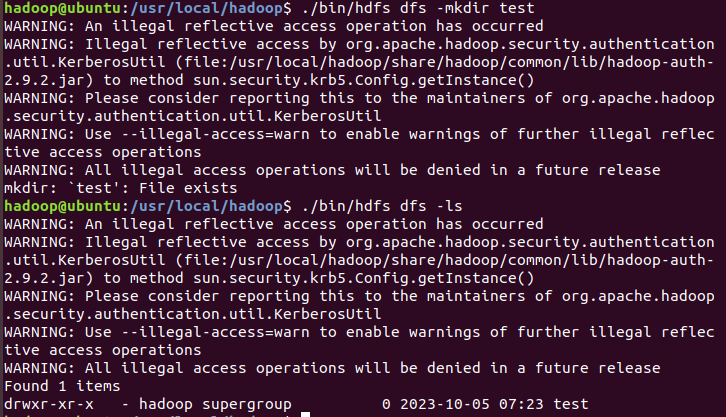
### 2.3、熟悉常用的Hadoop命令

**（按题目要求写出相应的操作命令，或给出运行截图）**

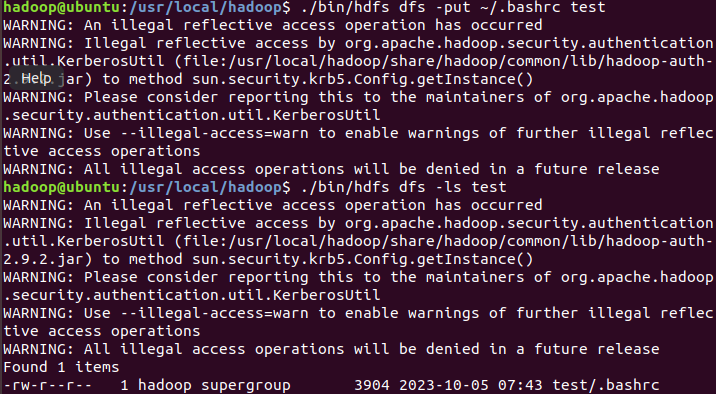
1.使用 hadoop 用户登录 Linux 系统， 启动 Hadoop，为 hadoop 用户在 HDFS 中创建用户目录“/user/hadoop”；



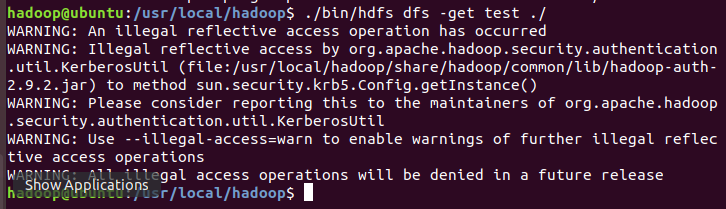
2.接着在 HDFS 的目录“/user/hadoop”下， 创建 test 文件夹，并查看文件列表；



3.将 Linux 系统本地的“～/.bashrc” 文件上传到 HDFS 的 test 文件夹中，并查看 test



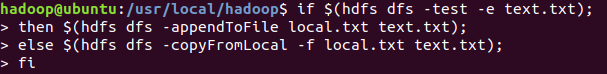
4.将 HDFS 文件夹 test 复制到 Linux 系统本地文件系统的“/usr/local/hadoop”目录下

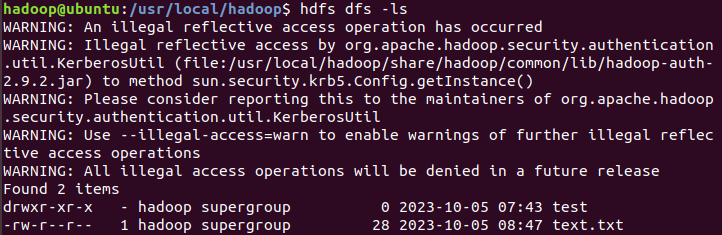


### 2.4熟悉常用的HDFS命令

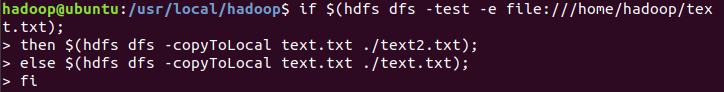
**（按题目要求写出相应的操作命令，或给出运行截图）**

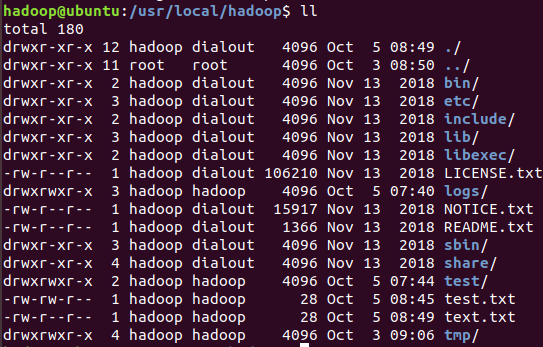
1. 向 HDFS 中上传任意文本文件，如果指定的文件在 HDFS 中已经存在， 则由用户来指定是追加到原有文件末尾还是覆盖原有的文件；



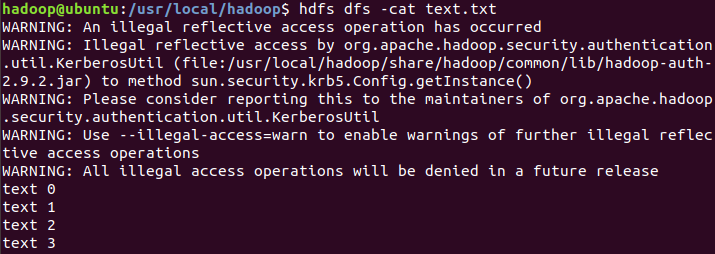


2. 从 HDFS 中下载文件，如果本地文件与要下载的文件名称相同，则自动对下载的文件重命名；

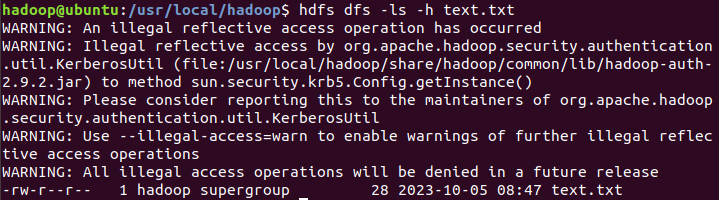




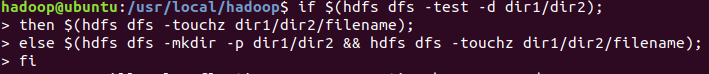
3. 将 HDFS 中指定文件的内容输出到终端中；

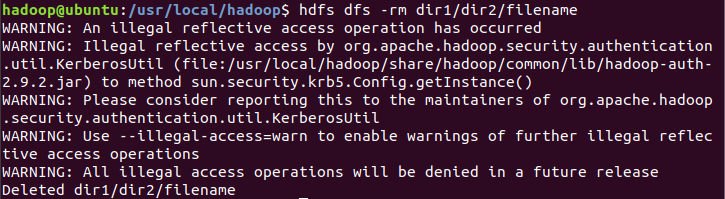


4. 显示 HDFS 中指定的文件的读写权限、大小、创建时间、路径等信息；



5. 提供一个 HDFS 内的文件的路径，对该文件进行创建和删除操作。如果文件所在目录不存在，则自动创建目录；





6. 提供一个 HDFS 的目录的路径，对该目录进行创建和删除操作。创建目录时，如果目录文件所在目录不存在， 则自动创建相应目录；删除目录时，由用户指定当该目录不为空时是否还删除该目录；











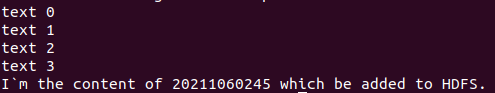


7. 向HDFS中指定的文件追加内容，由用户指定内容追加到原有文件的开头或结尾；

结尾：





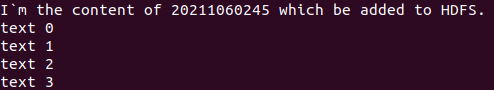


开头：









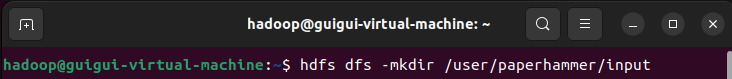
8.**（选做）**Java访问HDFS文件。编程实现一个类“MyFSDataInputStream” ，该类继承“org.apache.hadoop.fs.FSDataInputStream” ，要求如下：实现按行读取 HDFS 中指定文件的方法“readLine()” ，如果读到文件末尾，则返回空，否则返回文件一行的文本。

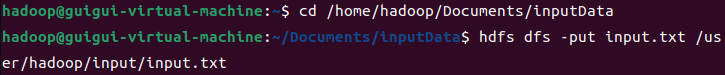
### 2.5 MapReduce应用实例

1.理解MapReduce的基本原理，实现Wordcount，说明实现流程，给出程序源代码和数据测试结果。

基本原理：MapReduce是一个分布式运算程序的编程框架，是用户开发“基于 hadoop 的数据分析应用”的核心框架。MapReduce采用“分而治之”的策略，一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的分片（split），这些分片可以被多个Map任务并行处理。

将输入数据上传到 hdfs：





编写Java程序： 创建一个Java程序，包括Mapper和Reducer类，以执行Word Count任务：

Mapper 文件：

文本

描述已自动生成

Reduce 文件：

文本

描述已自动生成

执行任务：

主函数调用：

文本

描述已自动生成

文本

描述已自动生成

将 .java 文件编译为 .class 文件，交给 Hadoop运行：

3

测试结果：

文本

中度可信度描述已自动生成

2.(选做)基于MapReduce实现并行化K-means聚类算法，说明实现流程，给出程序源代码和数据测试结果。

## 三、 实验总结

（总结本实验主要学习到的知识，遇到的问题及解决办法）

HDFS部分：

1. 环境设置：此部分较为简单，网络上有大量的教程，只要仔细基本不会出现问题，需要注意的是在Ubuntu版本的选择中，最新版Ubuntu可能存在未知的错误，故选择了18.04.6的版本，Hadoop版本采用2.9.2。
2. 伪分布式配置：Hadoop的伪分布式配置在网络上也有许多教程，但是部分存在问题，在进行这个阶段时进行了多次尝试，同时发现了不能多次执行namenode -format的命令，此举动会导致Namenode和Datanode的ClusterID不一致，从而导致Hadoop启动失败。
3. 熟悉常用的Hadoop命令：此部分也较为简单，只要稍加理解，不难实现。
4. 熟悉常用的HDFS命令：此部分需要基于上一部分的基础，同时需要注意要将HDFS设置为环境变量，否则HDFS命令需要在/bin目录下运行，稍有难度，但若仔细也可较快完成。
5. HDFS的编程部分：此部分未实现。主要原因为在我配置完jdk，Intelli IDEA，VScode，maven并配置了hadoop编程环境后，代码仍然在出现错误，最后不报错了却无法实现功能，遂放弃。

MapReduce部分：

1. 环境设置：在开始实验前，根据操作系统和版本的要求，下载适当的Hadoop发行版并安装它。配置Hadoop集群的各个组件，例如HDFS等。
2. 编写MapReduce程序：编写了一个包括Mapper和Reducer的MapReduce程序，用于执行Word Count任务。这需要定义Map和Reduce阶段的逻辑，以便对文本数据进行单词计数。
3. 数据准备：在Hadoop集群中，需要创建一个输入目录，并将要进行单词计数的文本文件上传到该目录。这个步骤是为了让MapReduce任务能够访问数据。
4. 运行任务：使用Hadoop命令提交MapReduce任务。确保提供正确的输入和输出路径，以便任务能够读取数据并将结果写入指定目录。
5. 错误处理：在编写和运行MapReduce任务时，可能会出现各种问题，如导入错误、依赖项问题或Hadoop集群配置问题。在遇到错误时，仔细阅读错误消息，并根据错误消息提供的信息来解决问题。
6. 结果分析：完成任务后，您可以查看输出目录中的结果文件，用于了解MapReduce如何处理数据。

总之，我们学习了如何使用Hadoop和MapReduce框架。接下来可以进一步学习和优化分布式计算任务，以处理更大规模的数据和更复杂的计算任务。