**大数据分析技术与应用实验报告**

**实验一:Hadoop环境搭建测试**

专 业： 电子信息

姓 名： 李建辉

学 号： S202075100

指导老师： 李建强、朱娜斐

2021年 04 月 17 日

信息学部软件学院

目 录

[1 实验任务 1](#_Toc476037968)

[2 实验过程 1](#_Toc476037969)

[2.1 Hadoop环境搭建 1](#_Toc476037970)

[2.2 Wordcount实例 5](#_Toc476037971)

[3 实验总结 1](#_Toc476037972)1

# 1 实验任务

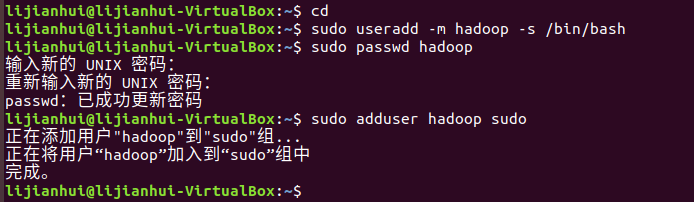
* + 1. 熟悉常用的Hadoop命令
    2. 运行Wordcount实例
    3. 搭建Eclipse编程环境
    4. 编程实现Wordcount程序

# 2 实验过程

## 2.1 Hadoop环境搭建

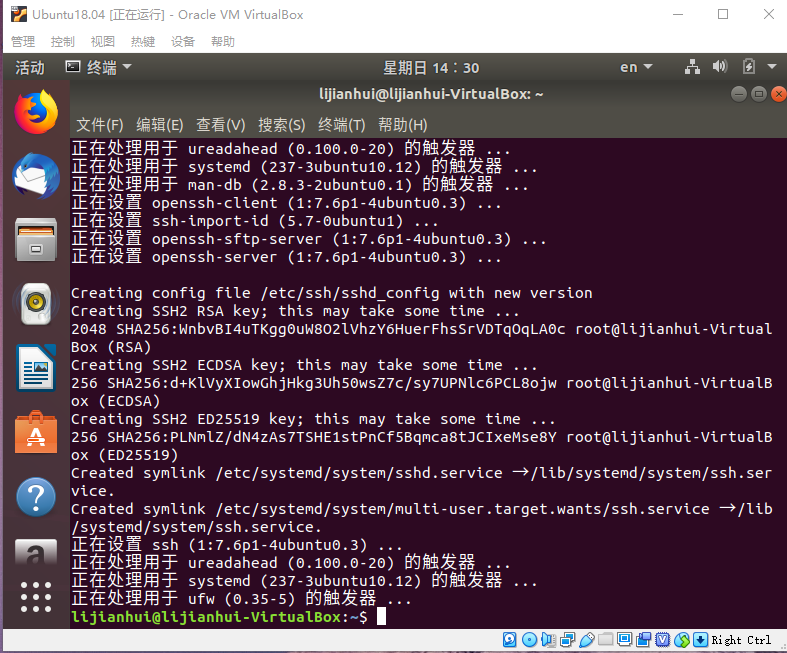
环境搭建说明：VM VirtualBox，Ubuntu 18.04(Linux 64位)，jdk 1.8，Hadoop 2.10.1

1）创建hadoop用户，并配置密码，赋予hadoop用户管理员权限

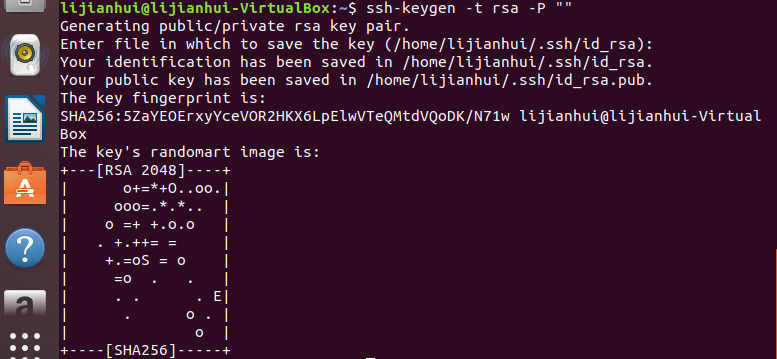


2）SSH登录权限设置，完成配置无密码登录

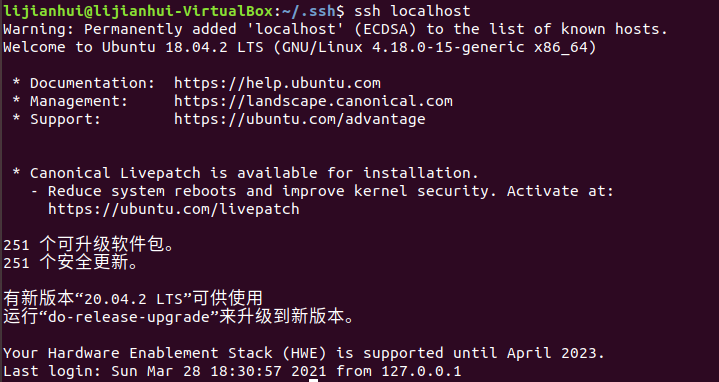
首先，安装SSH服务；



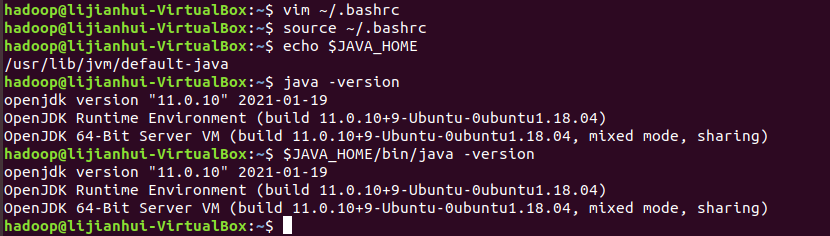
生成图形密码：



加入授权，配置ssh无密码登录成功



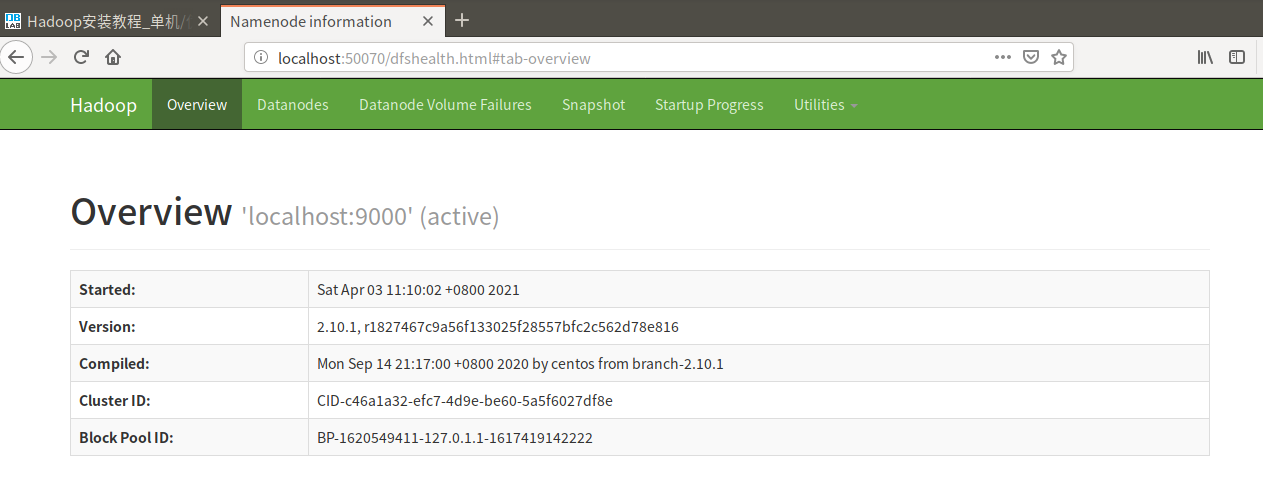
3）安装Java环境  
由于我安装的Linux系统是Ubuntu 18.04，使用sudo apt-get install default-jre default-jdk 命令，配置环境变量，并用 java -version来测试是否安装成功



4）Hadoop 单机安装配置，将Hadoop解压到/usr/local中，并查看是否安装成功：



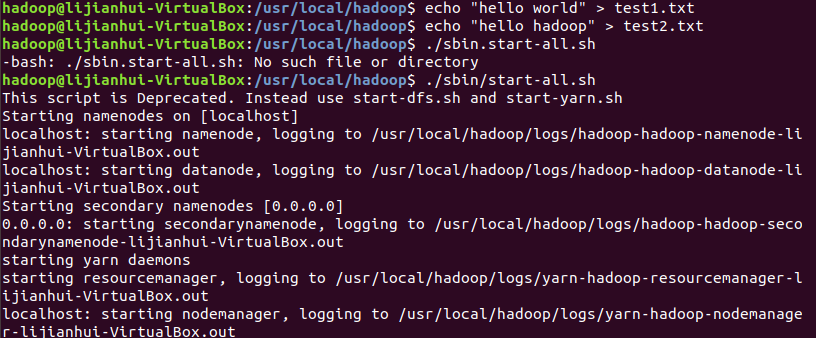
5）web界面，运行成功界面，可对Hadoop进行文件管理：



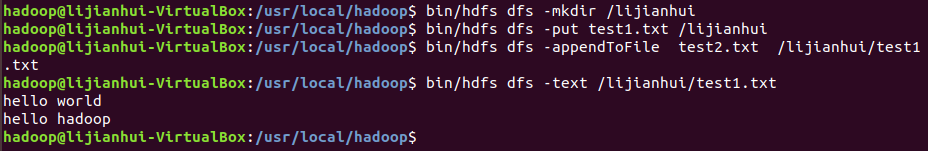
熟悉常用的Hadoop命令

完成要求的Hadoop命令：

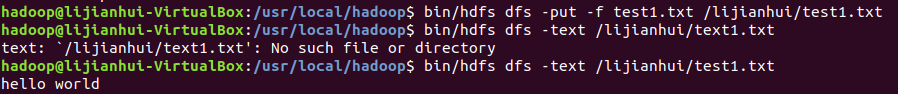
1. 向HDFS中上传任意文本文件，如果指定的文件在HDFS中已经存在，由用户指定是追加到原有文件末尾还是覆盖原有的文件；



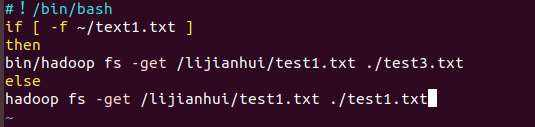
追加的情况：



覆盖的情况：

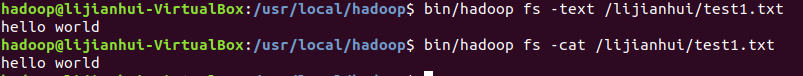


1. 从HDFS中下载指定文件，如果本地文件与要下载的文件名称相同，则自动对下载的文件重命名；





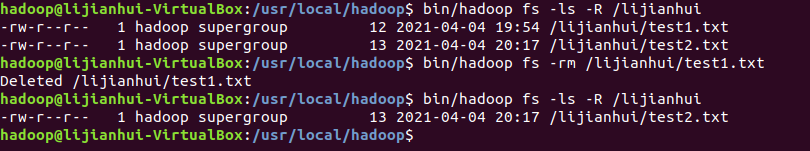
1. 将HDFS中指定文件的内容输出到终端中；



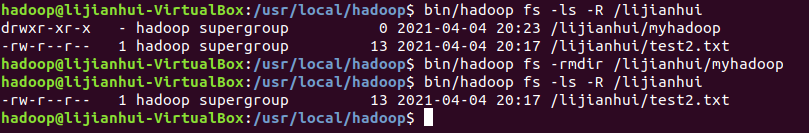
1. 显示HDFS中指定的文件的读写权限、大小、创建时间、路径等信息；

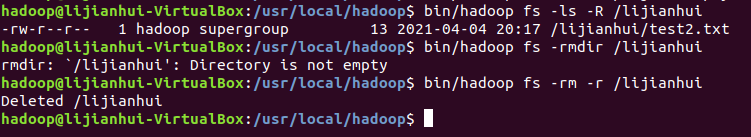


1. 删除HDFS中指定的文件；

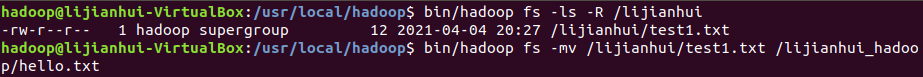


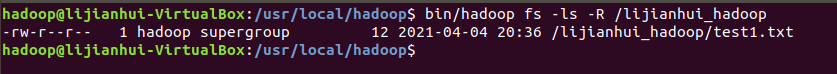
1. 删除HDFS中指定的目录，由用户指定目录中如果存在文件时是否删除目录；





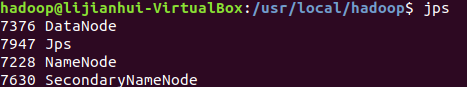
1. 在HDFS中，将文件从源路径移动到目的路径。





## 2.2 Wordcount实例

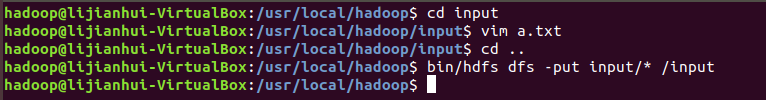
启动Hadoop，查看所有进程：



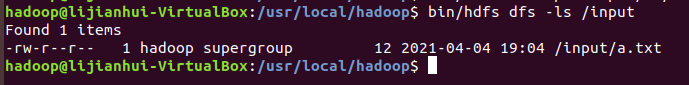
创建hdfs系统中input文件夹：



在本地建立input文件夹并编辑一个文件a.txt，将该文件上传到hdfs系统中的input文件夹：



查看hdfs文件系统中是否上传成功，成功截图如下：



查看文件内容：



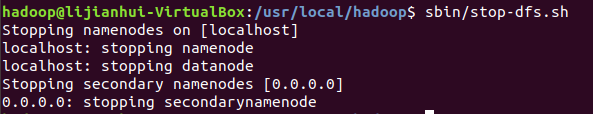
在hadoop中使用wordcount实例，input文件夹作为输入，并将结果放入到output文件夹下：



查看wordcount结果，正确，实验结果无误：



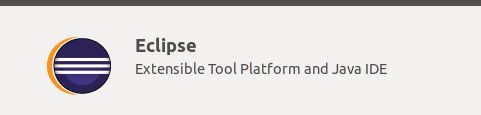
关闭进程：



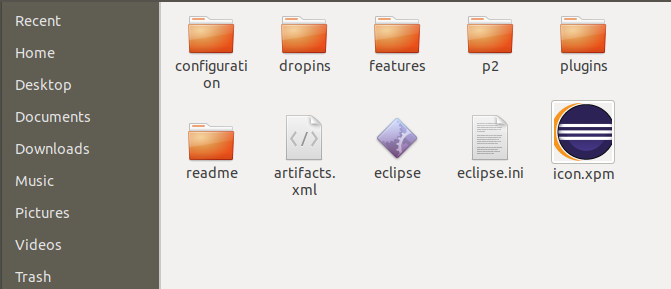
最后，动手编程实现Wordcount程序，运行验证。

1. 安装eclipse Java编程环境；

在软件中心找到eclipse，并安装



Eclipse安装路径/usr/lib/，并将hadoop-eclipse-kepler-plugin-2.10.1.jar（我自己用ant编译成功hadoop-eclipse-kepler-plugin-2.10.1.jar）放到eclipse解压文件夹的plugins文件夹下，以便顺利启动。

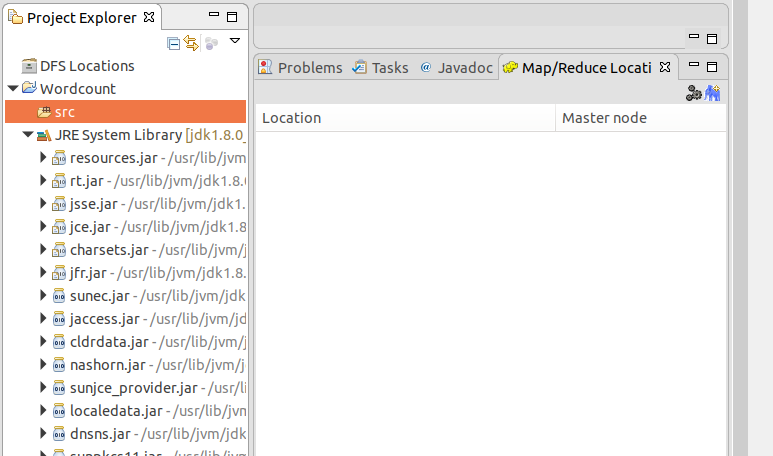


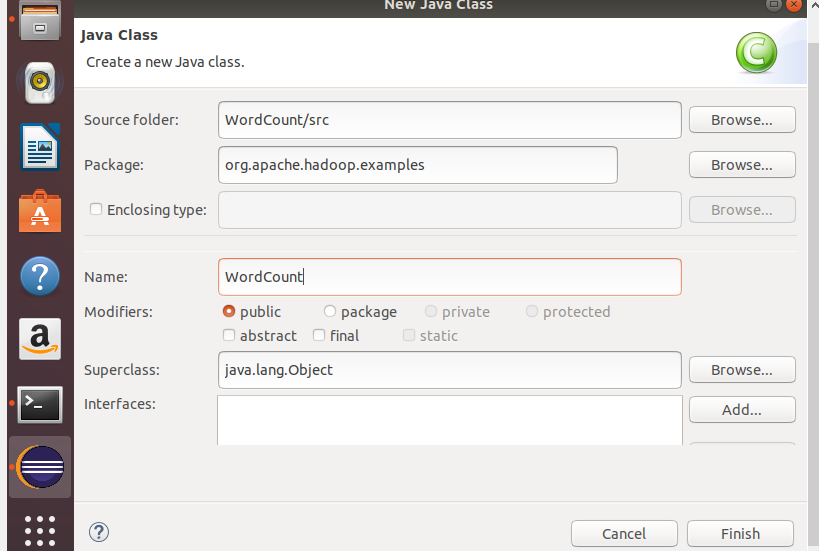
开启hadoop，并将本地文件上传到hdfs文件系统中，其中我的本地文件包括两个txt文本，内容分别为hello world、hello hadoop：



使用eclipse执行wordcount：

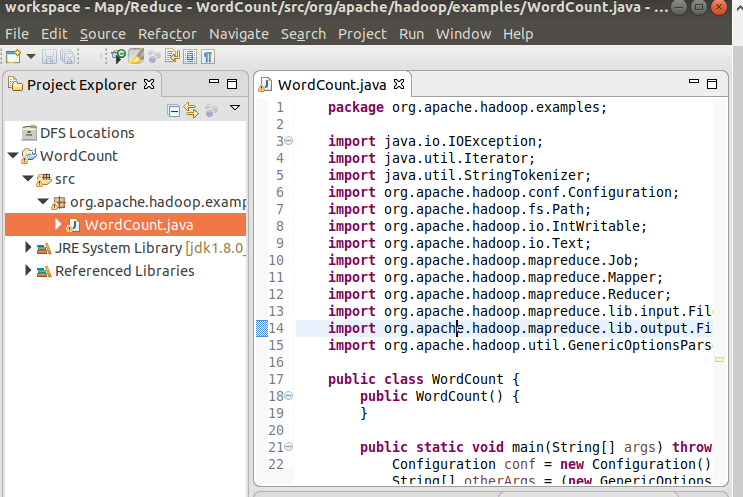
创建wordcount java项目：



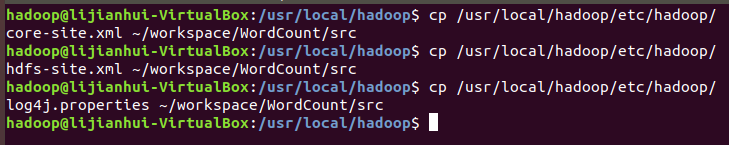


1. 使用Hadoop提供的Java API 进行编程实现；

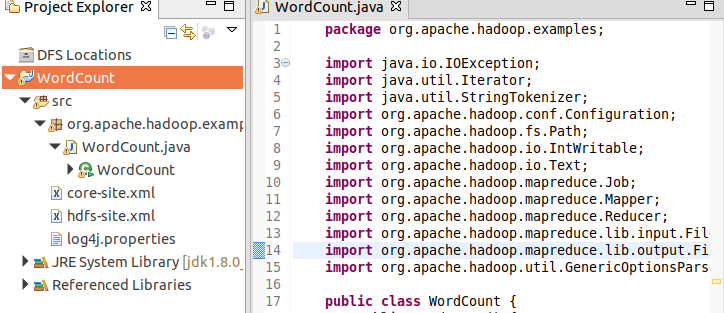
将wordcount代码复制到java文件中：



将Hadoop配置文件复制到java项目中

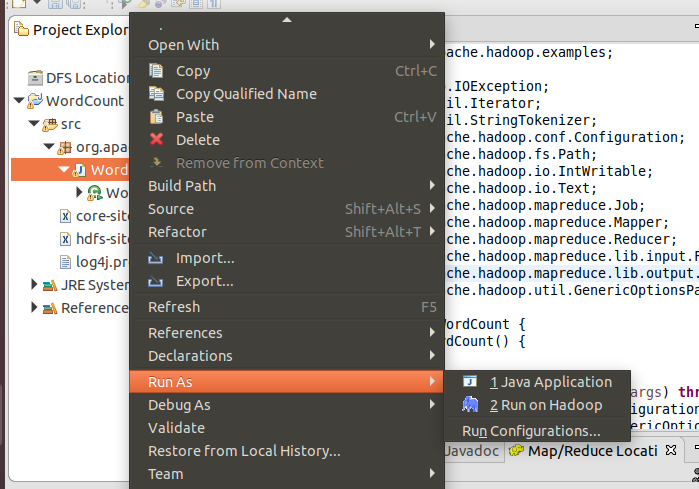


刷新后：

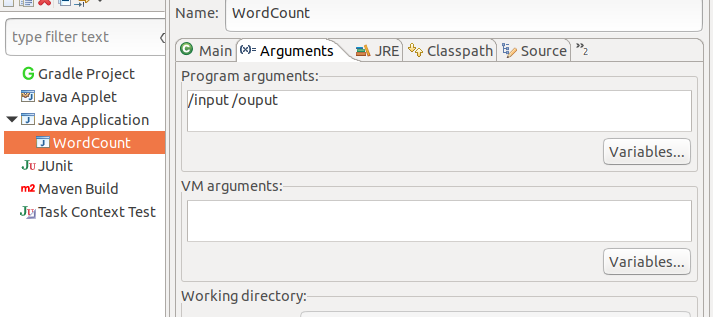


1. 运行程序，验证实验结果。

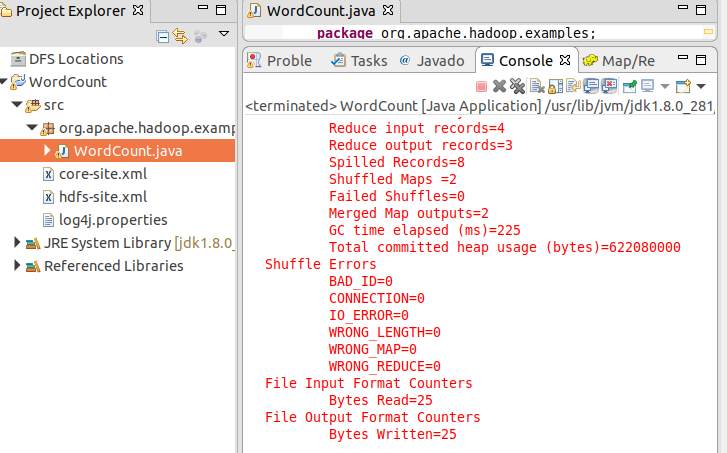
在Hadoop上运行：

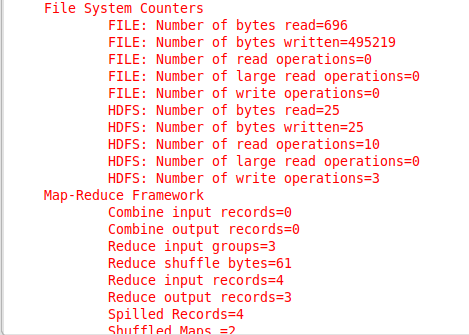


设置变量（要加斜杠，否则会报错）：

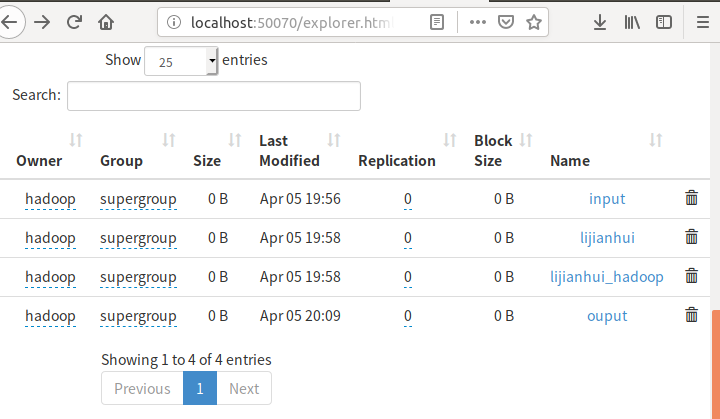


执行成功：

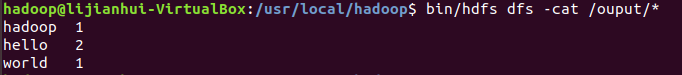




并且在web端可以看到出现了/output,成功：



可以得到我们的词数，准确无误：



## 3.实验总结

通过本次Hadoop实验，我从安装虚拟机，Ubuntu系统安装，再在Ubuntu上部署Java环境，最后完成Hadoop单机配置，这个过程实际上是环境部署。这整个流程中，我遇到了很多困难，包括版本的兼容性问题，由于我自身Ubuntu的版本过高，导致很多内容和教程并不一致。在这种情况下，这些困难锻炼了我的解决问题的能力，面对未知状况自我学习的能力。

在环境部署成功完成后，我查看并使用了常用的Hadoop命令 ，了解到如何上传文件到hdfs文件系统，如何从hdfs文件系统下载文件，如何查看文件等等常用命令；了解了Hadoop的分布式存储，在操作过程中熟悉了各个命令。

在数据作为主流的今天，作为软件工程相关专业的我们有必要去学习数据处理相关的知识。综上而言，此次实验对我意义重大。