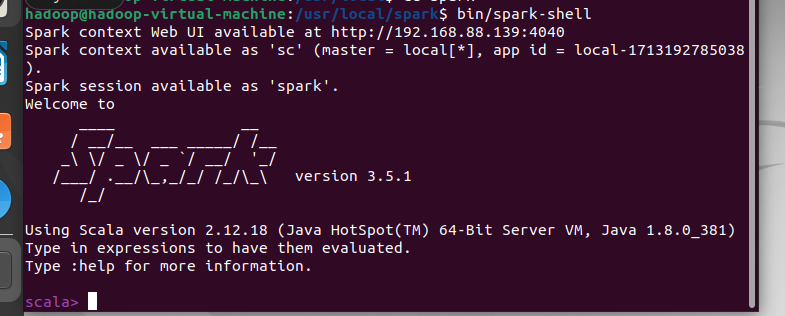
Lab2 Spark实验

第二部分实验流程：

1.使用命令行来实现wordcount

实际操作的文件路径为：**/usr/local/hadoop/README.txt**

首先先在**/usr/local/spark**文件夹打开终端，输入**bin/spark-shell**启动spark

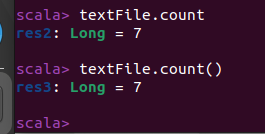


加载本地文件：这里使用file://是本地文件，使用hdfs://是hdfs结点上的文件

**val textFile = sc.textFile("file:///usr/local/hadoop/README.txt")**

使用

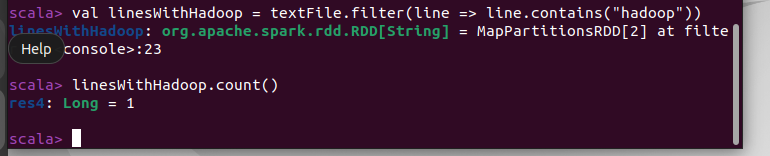
**textFile.count()**



然后接着来查询有hadoop的行数量：（这里请注意，word试验任务书中引号为中文的，记得自己修改为英文）

**val linesWithHadoop = textFile.filter(line => line.contains("hadoop"))**

**linesWithHadoop.count()**

****

2.scala语言实现wordcount

在 ./sparkapp/src/main/scala 下建立一个名为 WordCountApp.scala 的文件（`**vim ./sparkapp/src/main/scala/WordCountApp.scala**`）

**import java.io.File**

**import scala.io.Source**

**object WordCountApp {**

**def main(args: Array[String]): Unit = {**

**//文件路径**

**val filePath = "file:///usr/local/hadoop/README.txt "**

**val codec = "utf-8"**

**//打开文件**

**val file = Source.fromFile(filePath, codec)**

**val wc = file.getLines().flatMap(\_.split("\t")).toList.map((\_, 1)).groupBy((\_.\_1)).mapValues(\_.size)**

**println(wc)**

**// 关闭文件**

**file.close()**

**}**

**}**

./sparkapp 中新建文件 simple.sbt（`vim ./sparkapp/simple.sbt`）

**name := "Simple Project"**

**version := "1.0"**

**scalaVersion := "2.12.18"**

**libraryDependencies += "org.apache.spark" %% "spark-core" % "3.5.1"**

```

然后在/sparkapp文件夹中运行以下代码进行打包：

**/usr/local/sbt/sbt package**

打包完成之后使用： 记得需要修改成自己的scala版本

**/usr/local/spark/bin/spark-submit --class "WordCountApp" ~/sparkapp/target/scala-2.12/simple-project\_2.12-1.0.jar**

3.还可以使用java独立编程

### 第三部分实验流程：

这里可以参照word或者是网上博客内容：[Spark2.1.0入门：套接字流(DStream)\_厦大数据库实验室博客 (xmu.edu.cn)](https://dblab.xmu.edu.cn/blog/1387/)

按照word文档中的提示逐步完成。

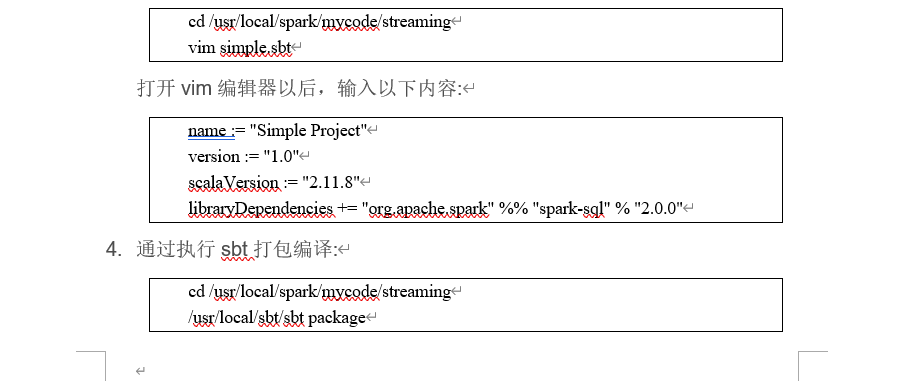
首先在spark中创建文件夹mycode，然后创建streaming

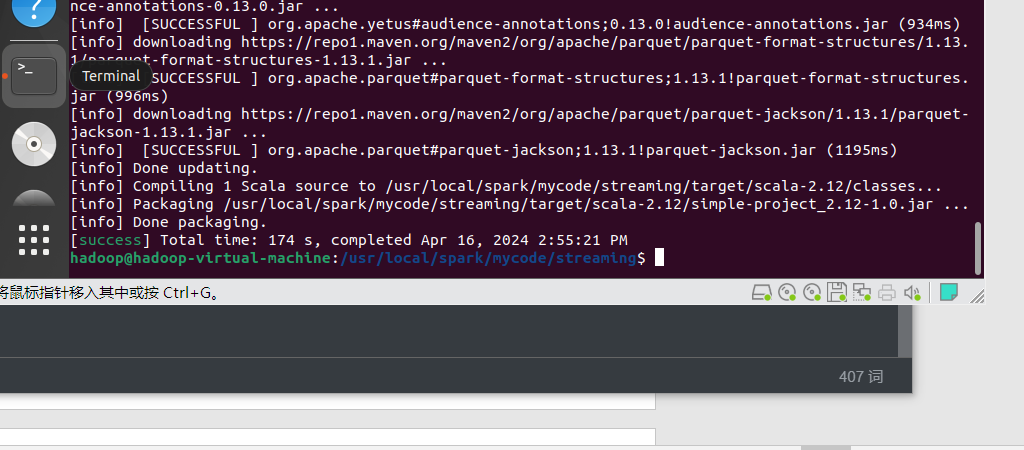
**mkdir /usr/local/spark/mycode**

**mkdir /usr/local/spark/mycode/streaming**

后续需要编写simple.sbt文件的脚本，这里面涉及scala和spark的版本号，在启动spark-shell后，可以直接观察到自己的对应版本号，

本部分首次打包的时候也会出现下载配置，请耐心等待





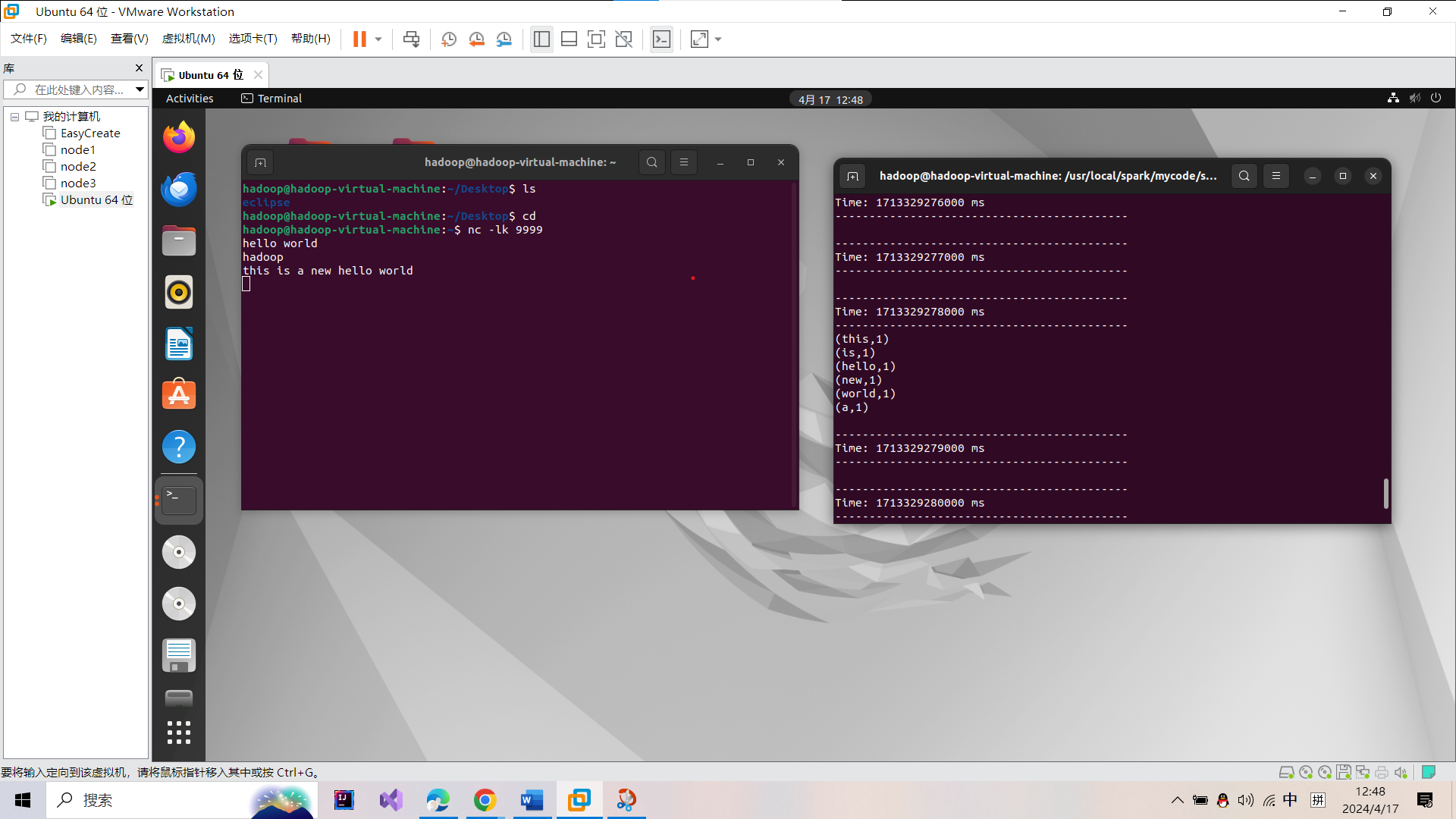
打包完成之后，运行下列代码，记得在target文件夹中查看自己的对应的scala的版本号

**cd /usr/local/spark/mycode/streaming**

**/usr/local/spark/bin/spark-submit --class "org.apache.spark.examples.streaming.NetworkWordCount" /usr/local/spark/mycode/streaming/target/scala-2.12/simple-project\_2.12-1.0.jar localhost 9999**

在运行上述jar包后，当前的终端窗口会报告time以及connection refused的错误，这是由于我们还没有启动nc窗口，再打开一个窗口，输入以下代码即可。

**nc -lk 9999**

****