实验三： Spark 的安装和使用

实验基本信息：

时间： ： 2024 年 9 月 14 日

实验类型： □验证性 □设计性 ☑综合性

班级： 计算机科学与技术专业（中外合作）4 班 班级代码:
理论老师： 邱开金 实验指导： 邱开金

实验报告提交说明：

本次实验需要撰写实验报告，实验报告填写时以完成实验任务为目的，先简要回答完成实验任务需要的步骤或需要执行的命令代码，再配以结果截图予以说明，图片数量不宜过多，能说明问题即可。最后配上总结，并提交到教师指定位置。

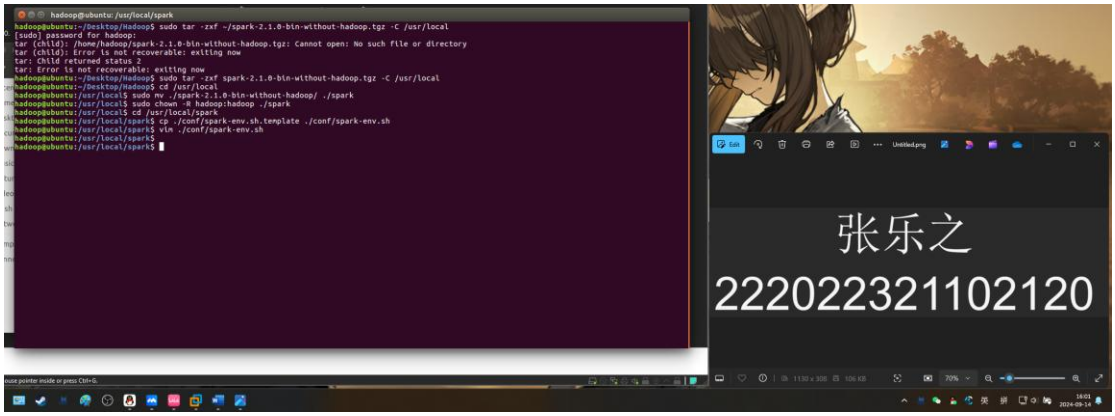
实验目的：

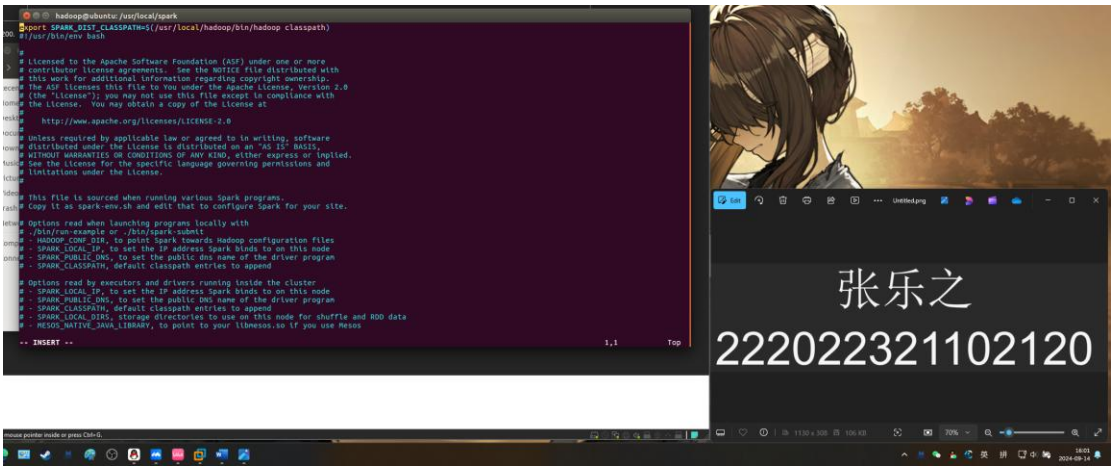
1. 掌握在已经有 hadoop 环境中安装 Spark 的方法。
2. 掌握使用 Spark 访问本地文件和 HDFS 文件的方法。
3. 编写独立的分布式程序，并提交（非 pyspark 交互式）。

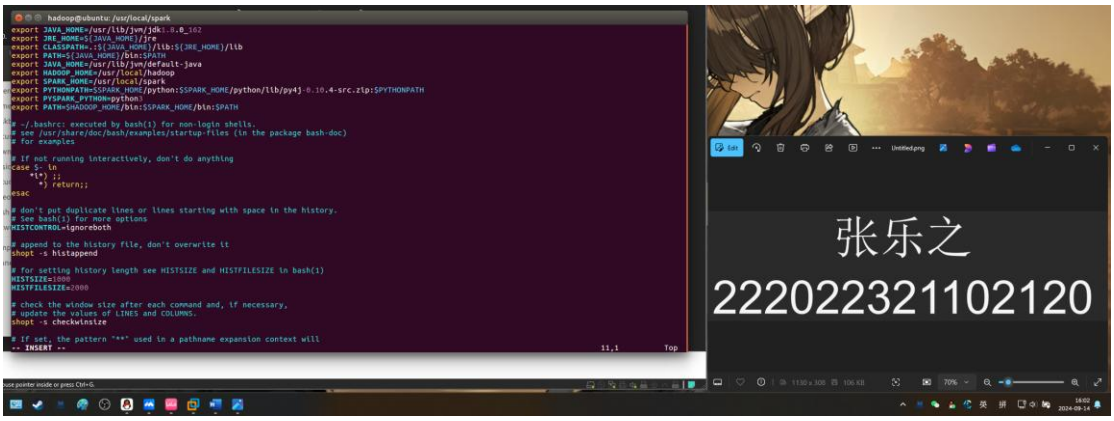
实验任务：

1. 在已有 hadoop 环境中安装 Spark

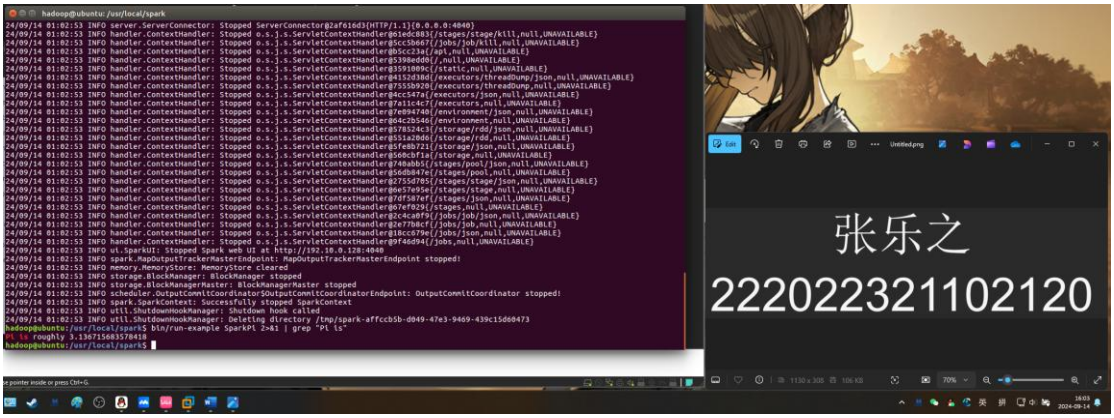
解压文件，修改配置文件并安装 Spark。





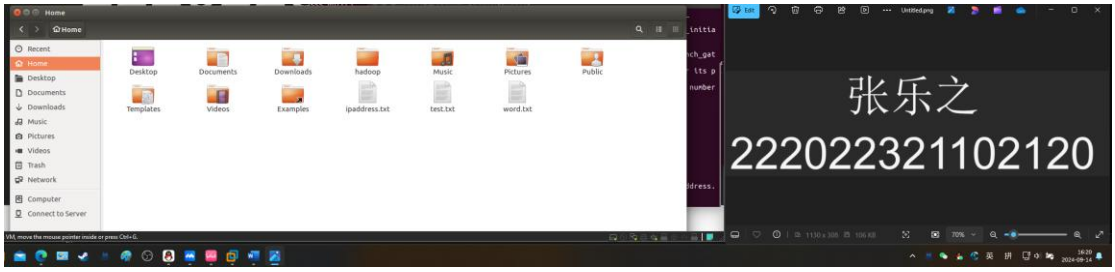


使用 Spark 验证成功安装。



2． 利用 pyspark 操作本地文件

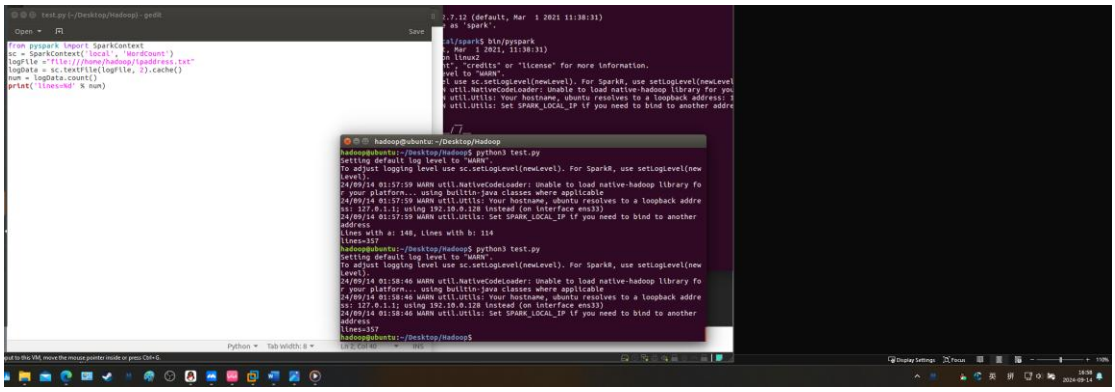
1) 将老师提供的 ipaddress.txt 文件复制到 ubuntu 系统中的 /home/hadoop/ipaddress.txt。



2) 在 pyspark 中读取 Linux 系统本地文件“/home/hadoop/ipaddress.txt"，然后统计出文件的行数。

(ipaddress.txt 由老师提供，请注意核对自己编写脚本的输出，本题目需要在实验报告中粘贴代码，并运行截图)

```
from pyspark import SparkContext
sc = SparkContext('local', 'WordCount')
logFile ="file:///home/hadoop/ipaddress.txt"
logData = sc.textFile(logFile, 2).cache()
num = logData.count()
print('lines=%d' % num)
```

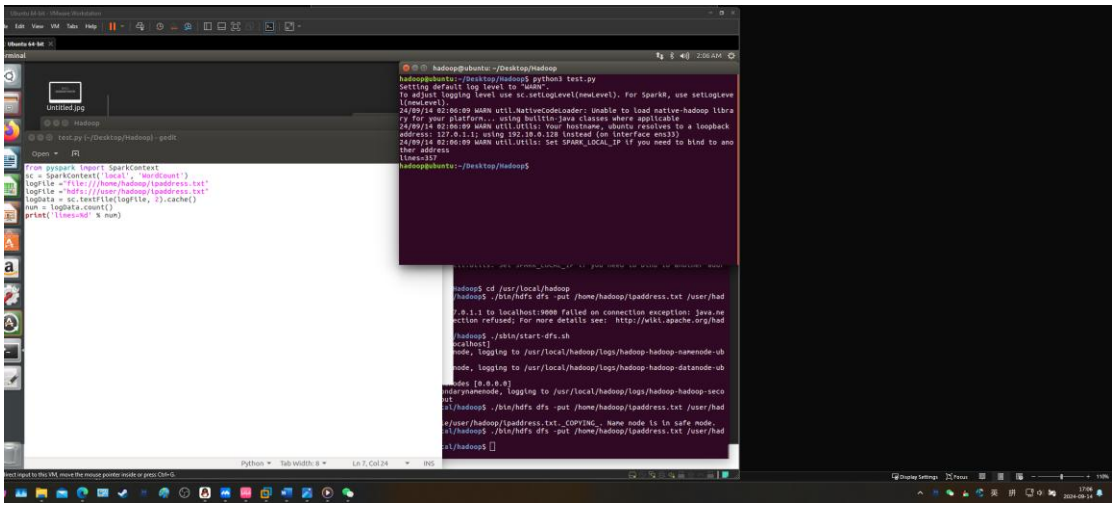


3．利用 pyspark 操作 hdfs 文件

在 pyspark 中读取 HDFS 系统文件“/user/hadoop/ipaddress.txt",然后，统计出文件的行数。(首先将本地的 ipaddress.txt 文件上传到 hdfs 中，路径如题所示。本题目需要在实验报告中粘贴代码，运行截图)

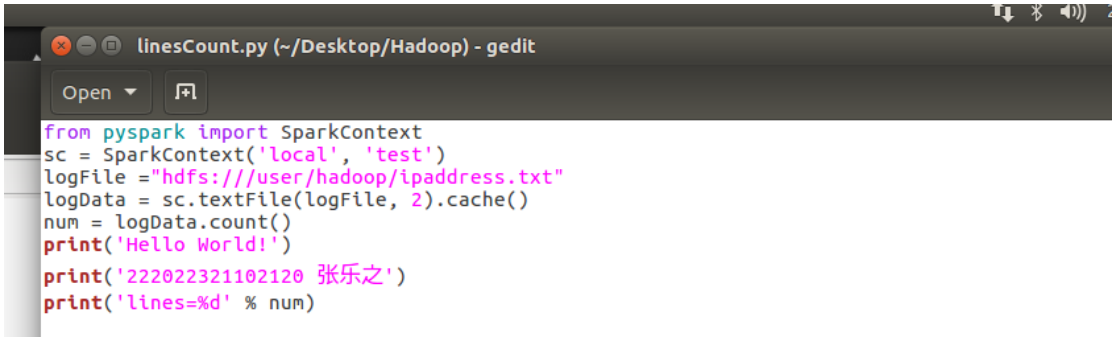
```
cd /usr/local/hadoop
./bin/hdfs dfs -put /home/hadoop/ipaddress.txt /user/hadoop/ipaddress.txt
```

```
from pyspark import SparkContext

sc = SparkContext('local', 'WordCount')

logFile ="file:///home/hadoop/ipaddress.txt"
logFile ="hdfs:///user/hadoop/ipaddress.txt"

logData = sc.textFile(logFile, 2).cache()

num = logData.count()

print('lines=%d' % num)
```

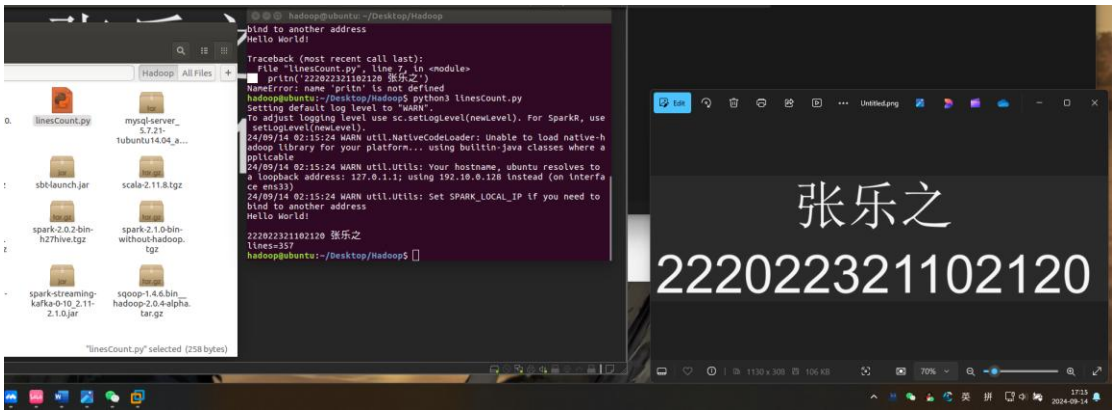


4．编写 spark 独立的应用程序（*.py 文件），实现对 ipaddress.txt 文件统计行数，**要求输出文件行数前先输出一行 hello world 以及本人的学号姓名(拼音亦可)**。

编写 spark 独立应用程序，读取 HDFS 系统文件“/user/hadoop/ipaddress.txt”，然后，统计出文件的行数。（作答要求：利用 vi 或 pycharm 工具编写 linesCount.py 程序，并在本地 调试运行通过）编写独立 python 文件，只需要在文件前面加上 from pyspark import SparkContext sc = SparkContext('local', 'test') 这两句，之后就可以像在 pyspark 中一样的操作。




```
from pyspark import SparkContext
sc = SparkContext('local', 'test')
logFile ="hdfs:///user/hadoop/ipaddress.txt"
logData = sc.textFile(logFile, 2).cache()
num = logData.count()
print('Hello World!')
print('222022321102120 张乐之')
print('lines=%d' % num)
```



实验总结：

本次实验让我更加深刻认识到了分布式系统的组成和相应开发工具的使用方法。我现在学会了熟练在主机和分布式系统之间互相传输文件，使用开发工具的 Api 编写使用分布式系统文件的 Spark 应用程序，以及使用 Linux 的 vim 进行 Python 编程。