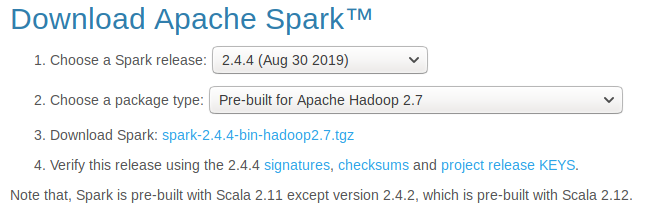


首先是进入实验二时装好java和ssh的环境中，心中已经暗自决定使用python，但是却有些不确定，因为还没有在linux上用过python



下载

sudo tar -zxvf spark-2.4.4-bin-hadoop2.7.tgz spark-2.4.4-bin-hadoop2.7/ -C /usr/local

不知道为什么还是解压到了原处

sudo mv spark-2.4.4-bin-hadoop2.7 /usr/local

cd /usr/local

sudo mv spark-2.4.4-bin-hadoop2.7 spark

配置环境

vi /etc/profile

#SPARK\_HOME

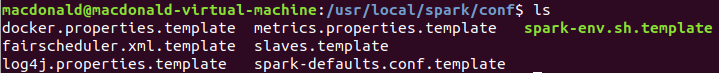
export SPARK\_HOME=/home/hadoop/spark-2.2.0-bin-hadoop2.6

export PATH=$SPARK\_HOME/bin:$PATH

cd spark/

cd conf/

ls



对template文件复制一份并重命名位spark-env.sh



查看java路径echo $JAVA\_HOME

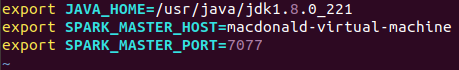


修改spark-env.sh文件，在尾部添加

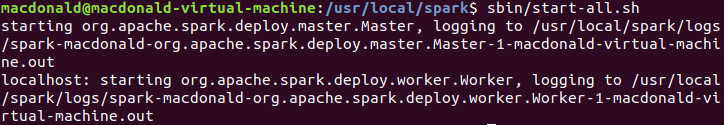
export JAVA\_HOME=/usr/java/jdk1.8.0\_221

export SPARK\_MASTER\_HOST=macdonald-virtual-machine

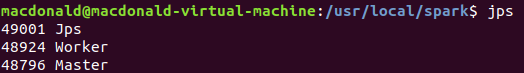
export SPARK\_MASTER\_PORT=7077



回到spark的主目录中，找到sbin目录，然后启动spark,使用命令sbin/start-all.sh



使用jps查看是否启动成功



显示了正确的内容

接下来试着使用它自带的example中求圆周率的示例程序，参数设置为100

bin/spark-submit --master spark://macdonald-virtual-machine:7077 --class org.apache.spark.examples.SparkPi examples/jars/spark-examples\_2.11-2.4.4.jar 100

经历了半分钟的等待（比网上慢了20+s，可能是机器性能差异的问题），终于得到结果



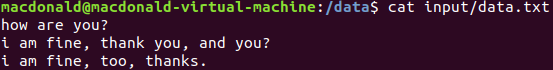
成功

另外创建两个文件夹分别用作输入输出

sudo mkdir -p /data/input

sudo mkdir -p /data/output

在input中创建一个data.txt并进行输入以下内容



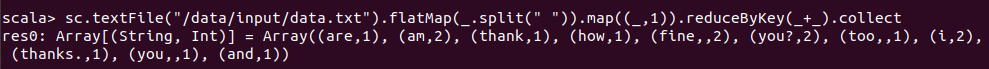
启动spark-shell

bin/spark-shell --master spark://macdonald-virtual-machine:7077



sc.textFile("/data/input/data.txt").flatMap(\_.split(" ")).map((\_,1)).reduceByKey(\_+\_).collect

得到正确的单词统计结果

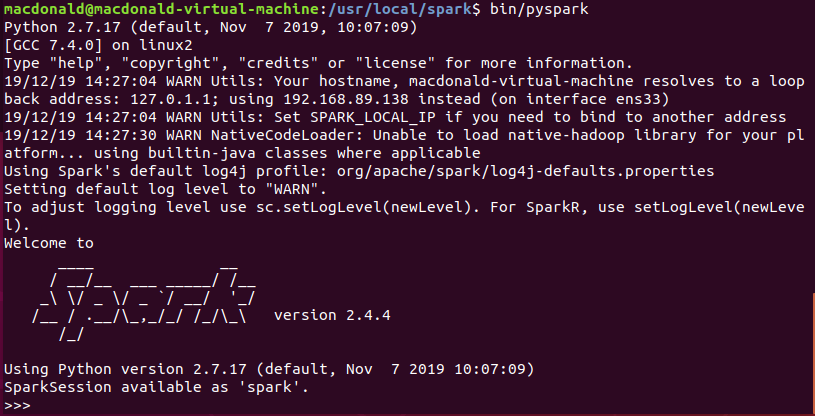


安装python

sudo apt install python



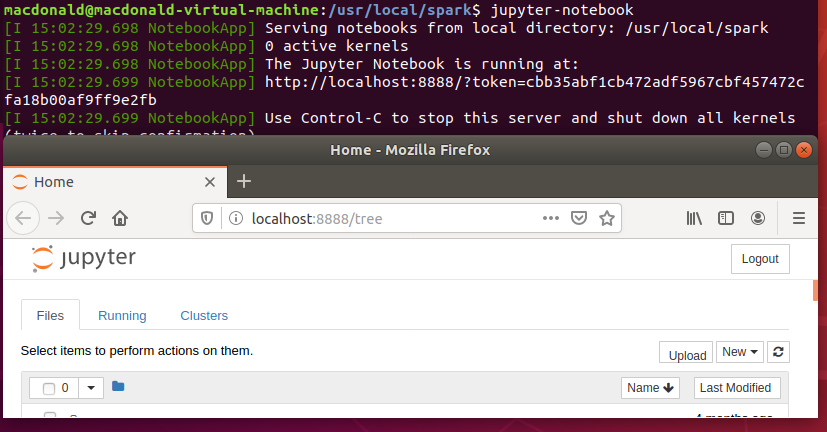
开启pyspark



可以正常运行简单的程序

为了方便，还是安装jupyter notebook

sudo apt install jupyter



PySpark in Jupyter

去~/.bashrc文件最后，添加配置PySpark driver的环境变量

sudo vi ~/.bashrc

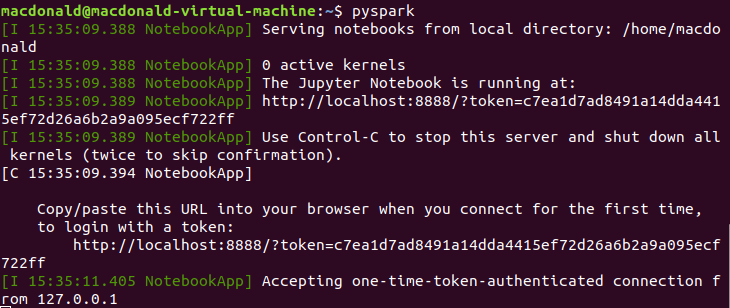


export PYSPARK\_DRIVER\_PYTHON=jupyter

export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook'

source ~/.bashrc

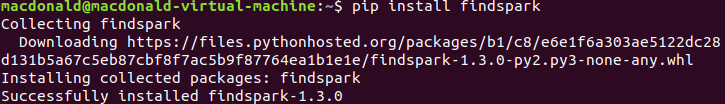
然后pyspark 可以直接打开notebook



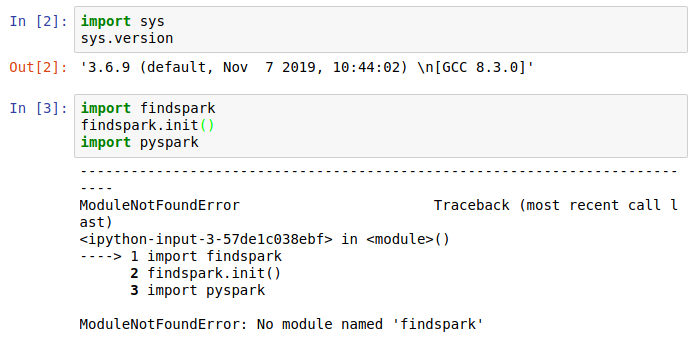
在运行时报错，尝试改用findspark

sudo apt install python-pip

pip install findspark



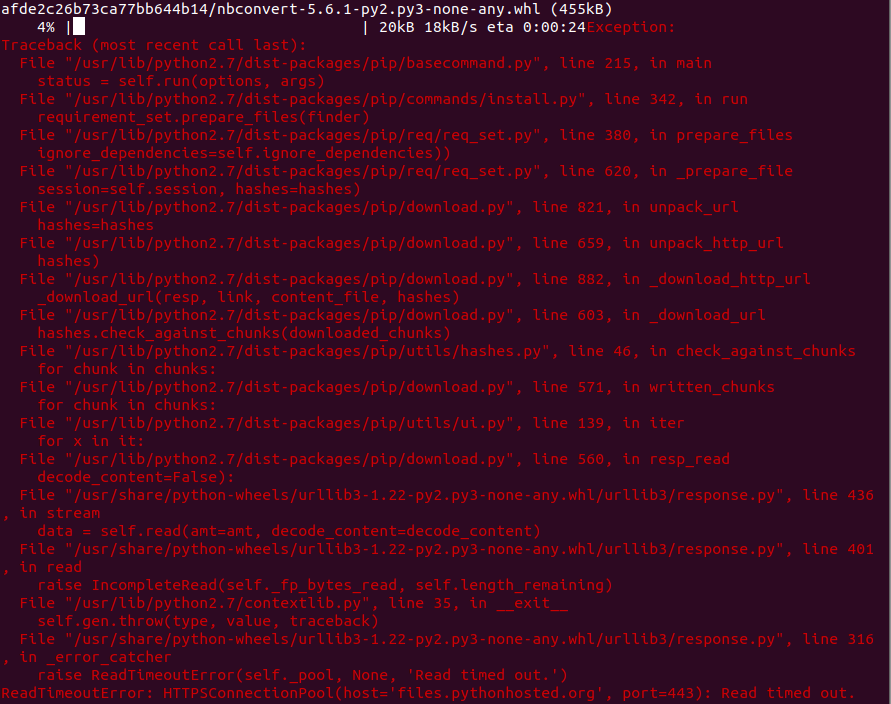
却有



觉得是版本的问题，因为虚拟机上既有python也有python3，查到了jupyter应该安装正确的内核，因为前面尝试过不安装python2就无法启动pyspark，所以必须得使用python2的内核。

想起来，前面因为提示没有pip，所以安装jupyter的时候直接使用了sudo apt install jupyter，在这里要改用pip来安装的话还要卸载之前的版本，所以打算使用快照，回跳。

几乎重做一遍前面的环节，但区别是把sudo apt install python-pip放在juypter的安装前面，用pip install jupyter来安装jupyter，遇到报错。



受限于机器配置，虚拟机处理起来总是让人难以保持耐心，后来听说 Windows 上完成实验也比较容易，于是再次转换策略。

1. 下载解压spark
2. 配置环境变量
3. python -m pip install findspark
4. 补充配置spark驱动的环境变量
5. 启动jupyter notebook跑样例

