实验1 Spark分布式大数据分析平台安装与配置

1.实验目的

通过本实验，搭建基于spark的大数据分析与设计平台，为（可能的）后续的大数据实验做准备。能够在Linux操作系统下pyspark交互式方式下运行python spark语句。能够使用spark-submit提交数据处理任务。

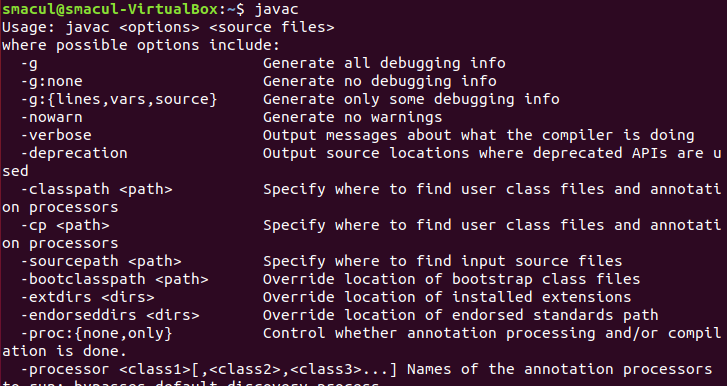
2.实验步骤

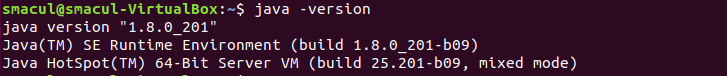
第0步，准备工作

0.1准备Linux操作系统，ubuntu 16.04

0.2安装JDK1.8，配置JAVA\_HOME

0.3下载spark2.4，解压缩

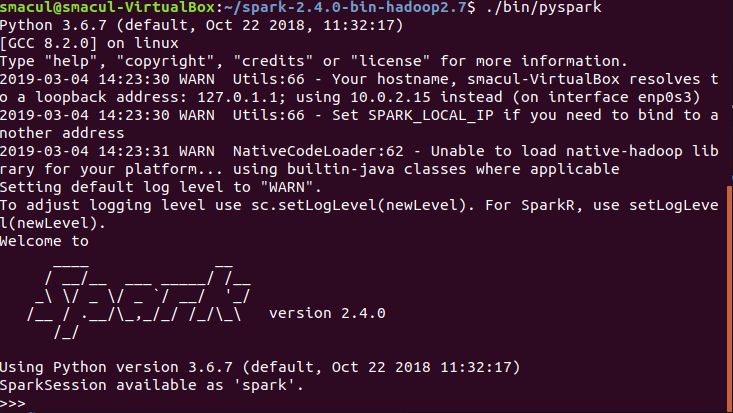




第1步 配置Spark standalone模式

1.1 开启terminal 进入解压缩后的spark文件夹

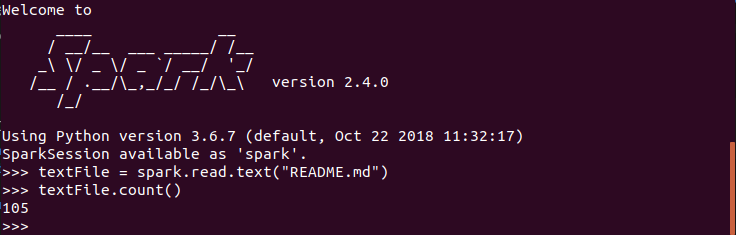
1.2 运行 ./bin/pyspark



1.3在线运行

textFile = spark.read.text("README.md")

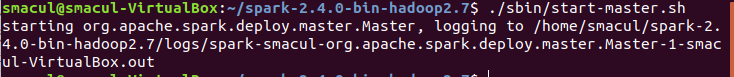
textFile.count()

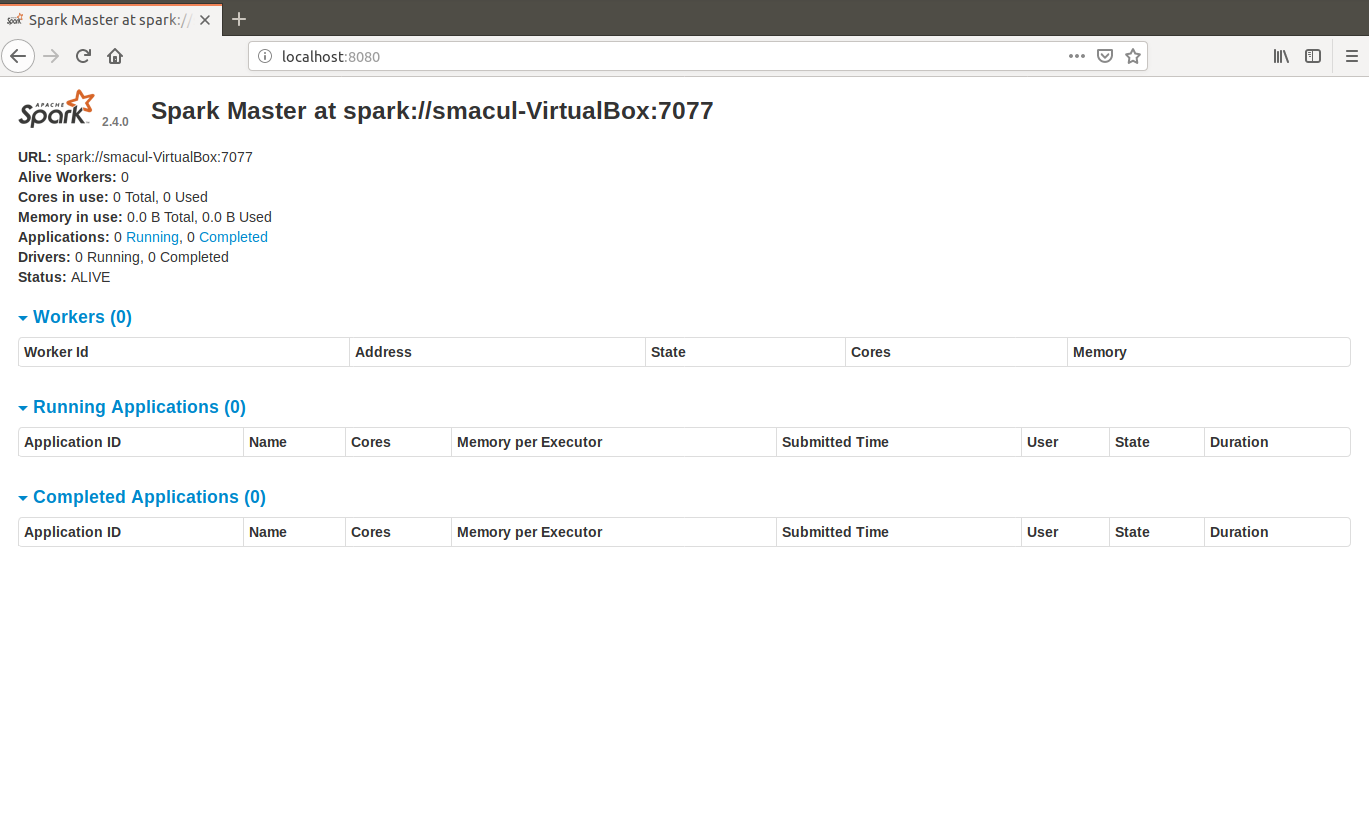


第2步 配置singlenode cluster模式

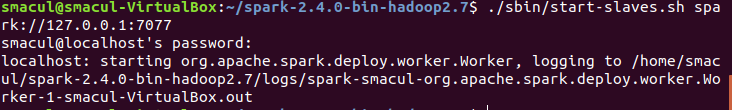
2.1修改conf/slaves 文件，增加localhost

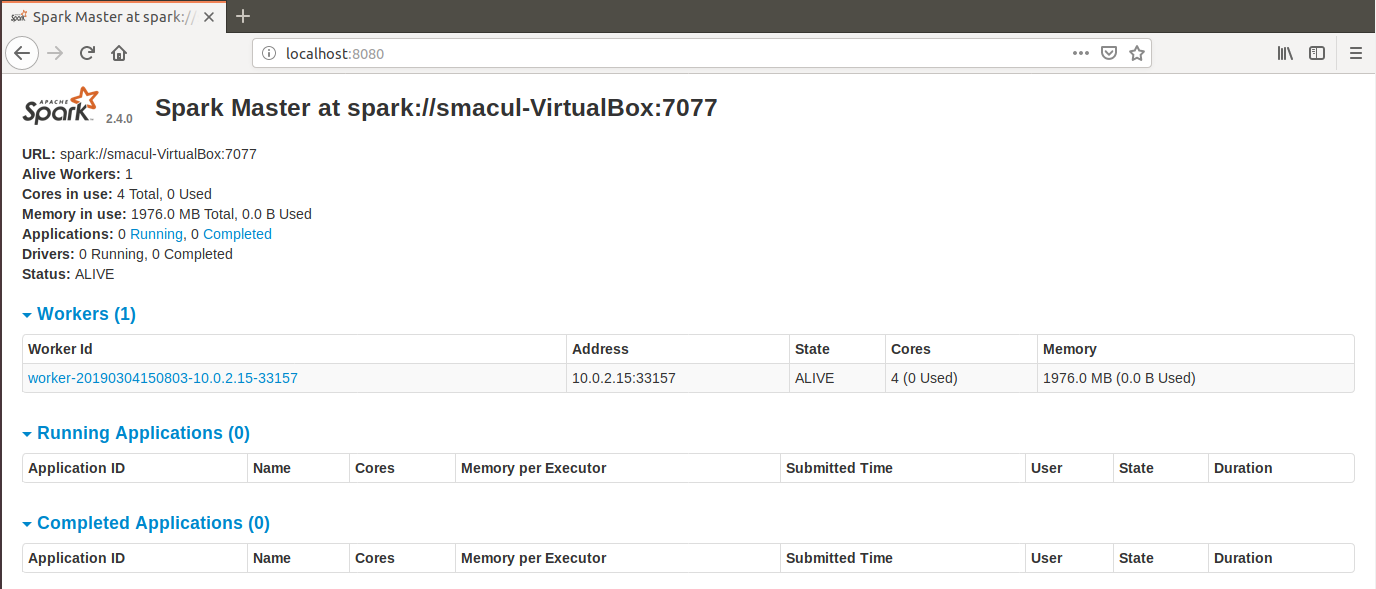
2.2 sbin/start-master.sh查看<http://localhost:8080>



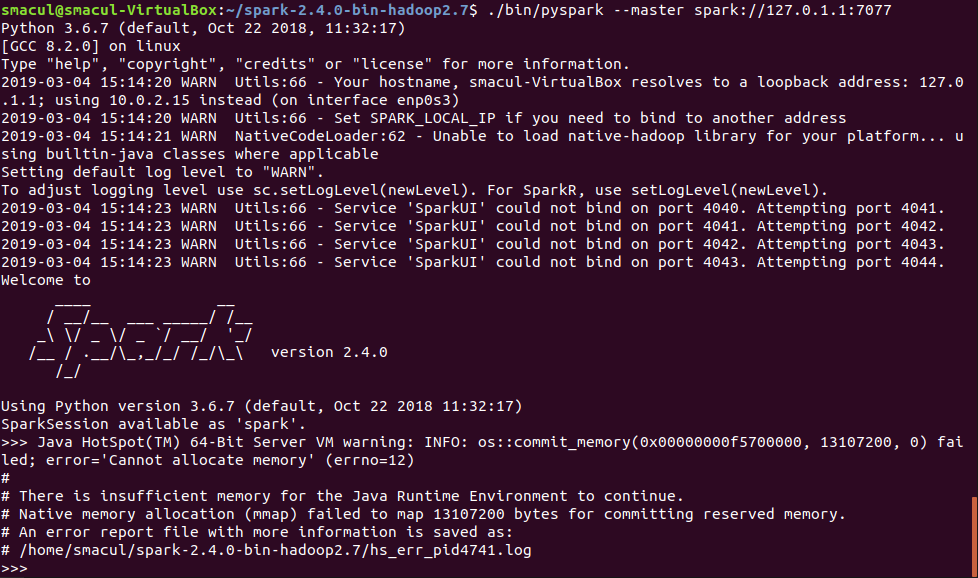


2.3sbin/start-slaves.sh spark://127.0.0.1:7077 查看<http://localhost:8080>





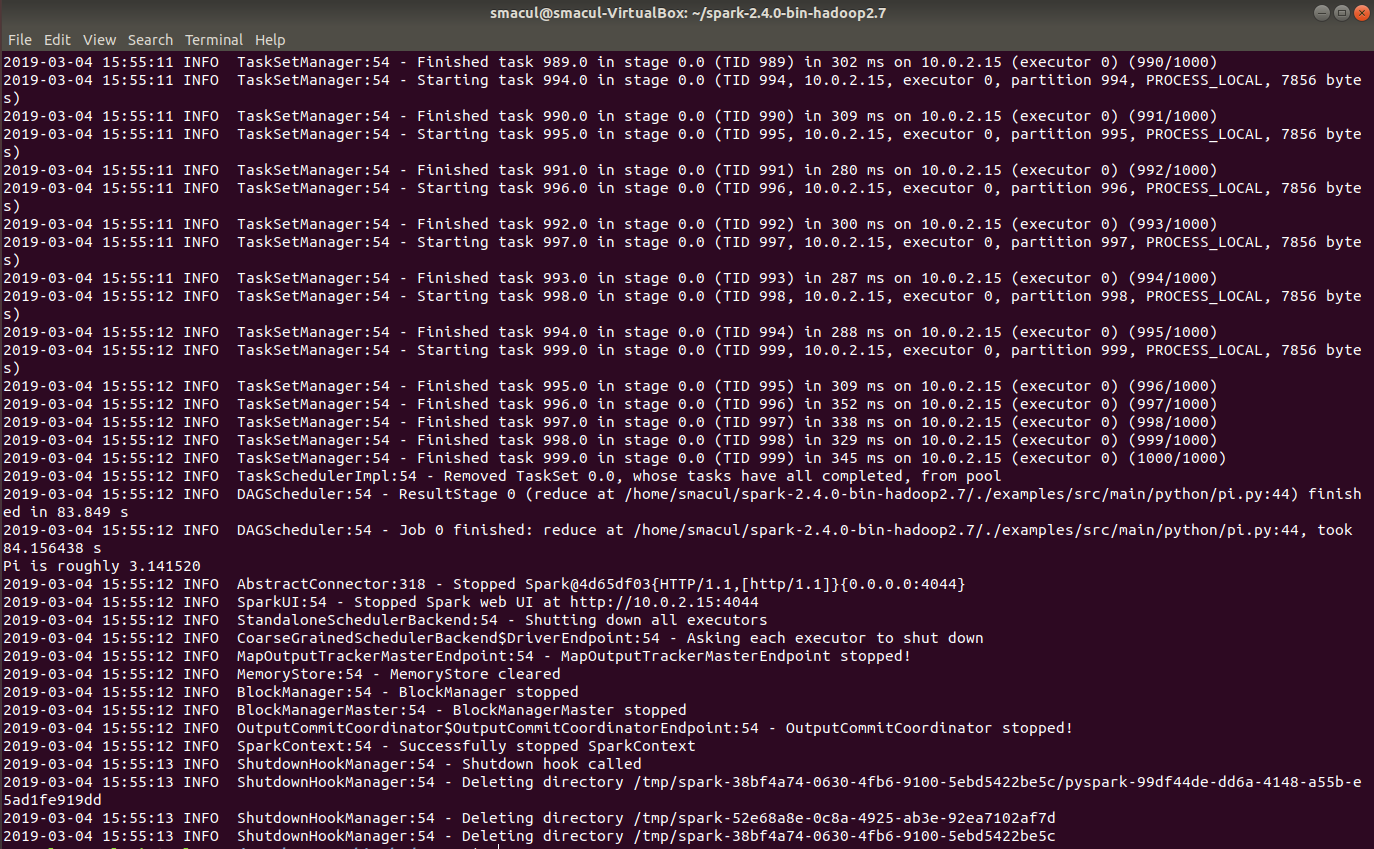
2.4 运行/bin/pyspark --master spark://127.0.0.1:7077

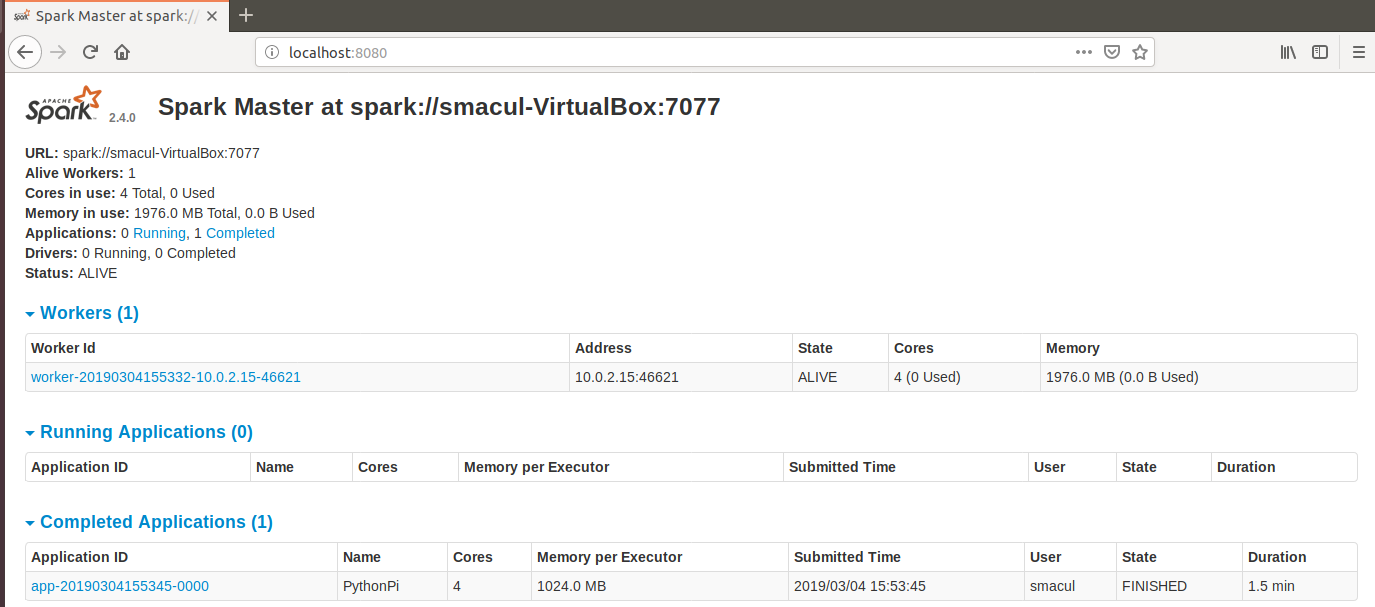


2.5（退出上一步的交互式环境）运行Spark例子：

./bin/spark-submit --master spark://127.0.0.1:7077 ./examples/src/main/python/pi.py 1000

查看http://localhost:8080中的信息





Pi is roughly 3.141520

第3步 愉快的开始spark Python编程

3. 实验结果展示

见报告 2 部分截图

4. 实验总结与收获

通过这次实验，我初步了解了 Spark 的功能，熟悉了Spark的各项配置。