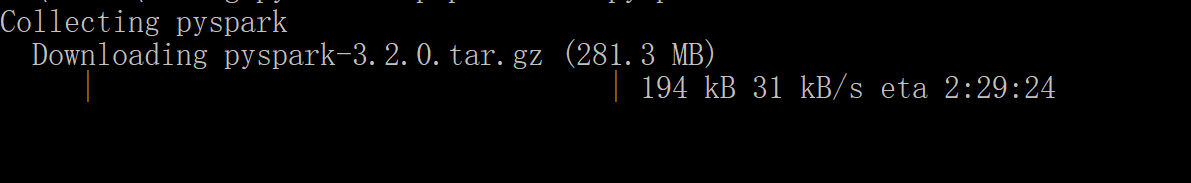
**Lab6\_Spark\_Cluster\_Setup**

**薛劭杰 1930026143**

# Individual Part

Install the by the python pip by the command:

-



After waiting some minutes, take the command again and if you can see “Requirement already satisfied”, then it means that you install successfully.



Then add the to your system environment:

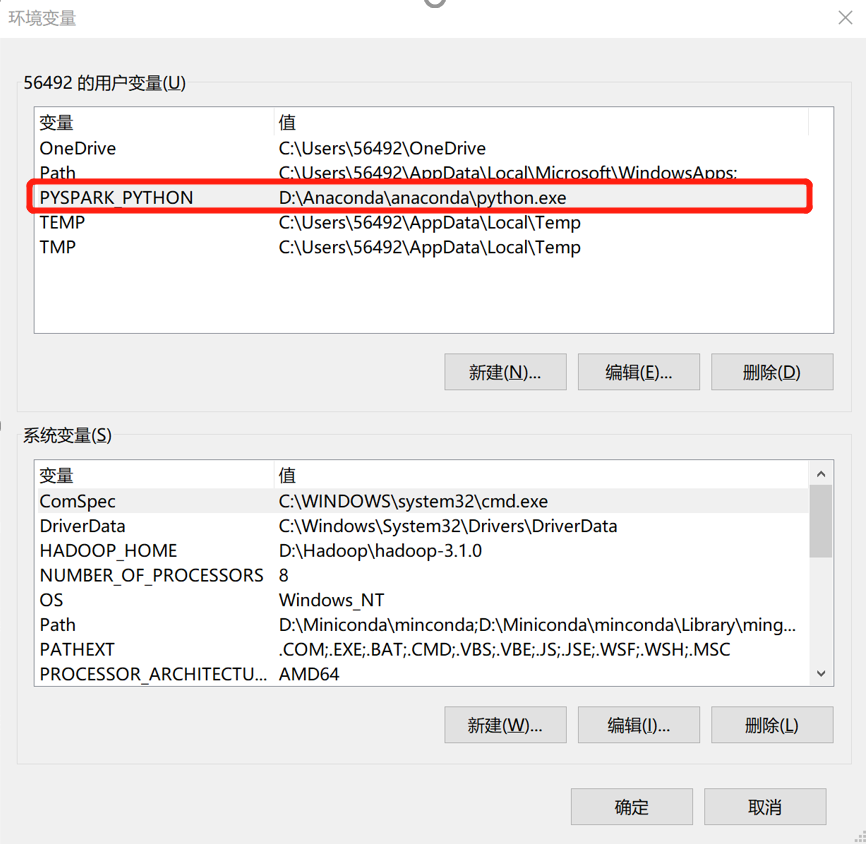
Step1: Entering your “Desktop” and click the “Attribute”, you can this page and click the “Advanced System Settings”.



Then click the “Environment Variable” in the bottom.

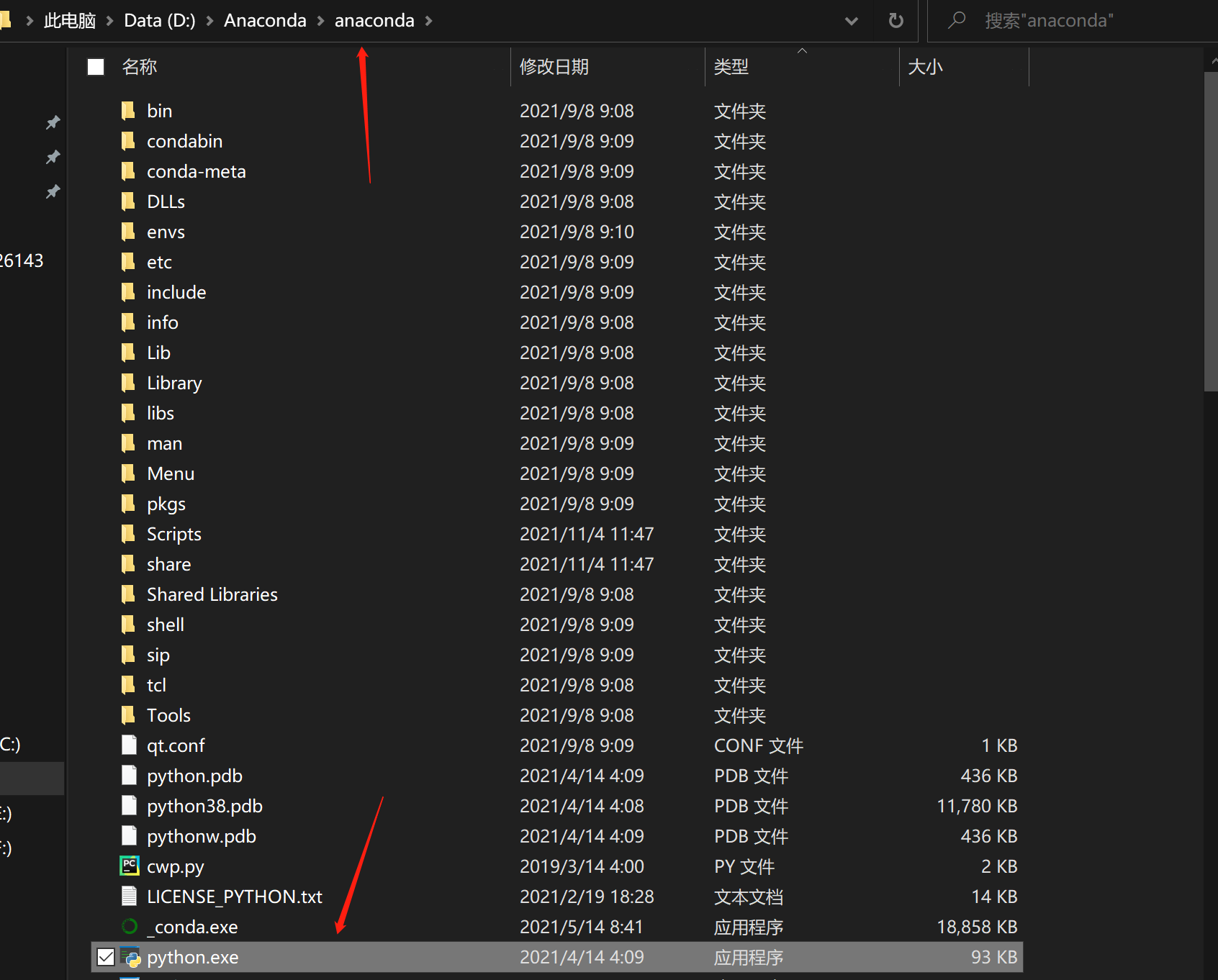


The add the information into the Use Variables and click “Ensure”.

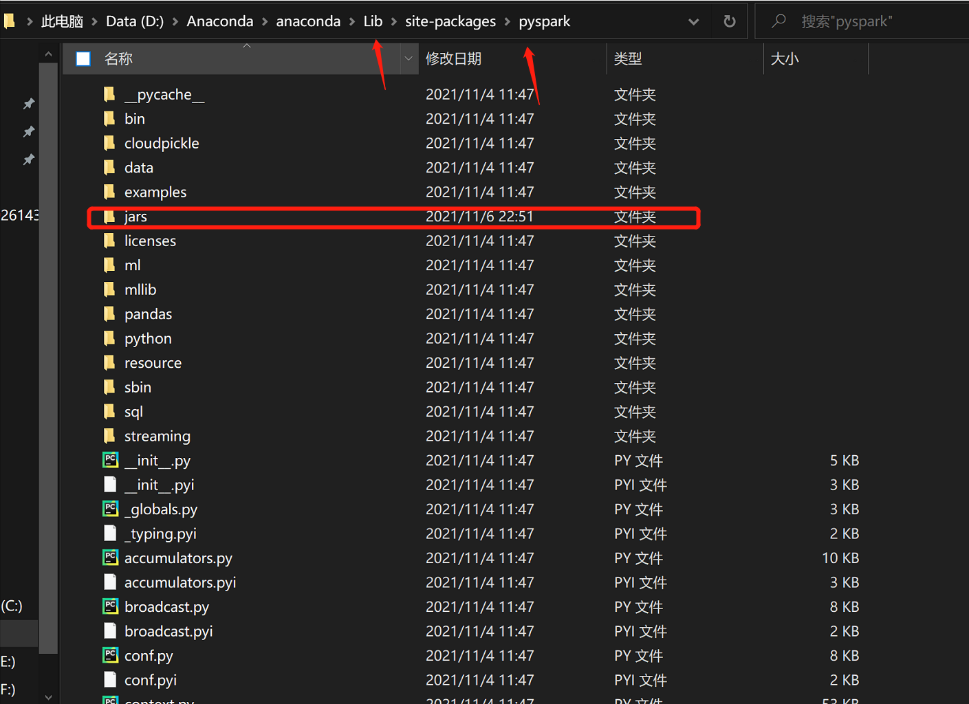


The value of the is the path of the python.exe file of the anaconda.

You can find it in the path you install the and .

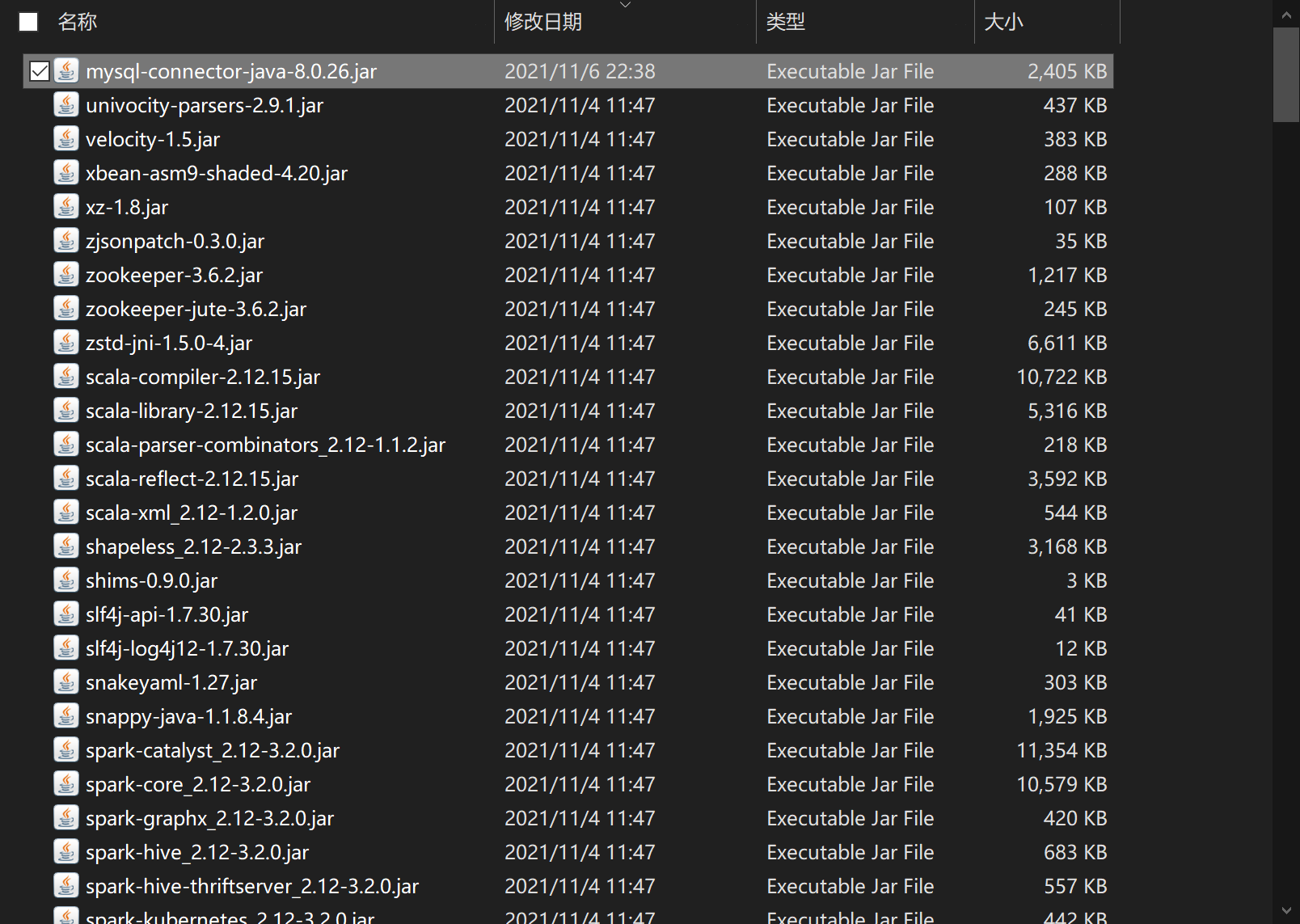


Then download the lasting java .jar package (mysql-connector-java-8.0.26) to the jar library of the anaconda, you can find it in the sub-path in following picture: (Pay attention to the Lib)

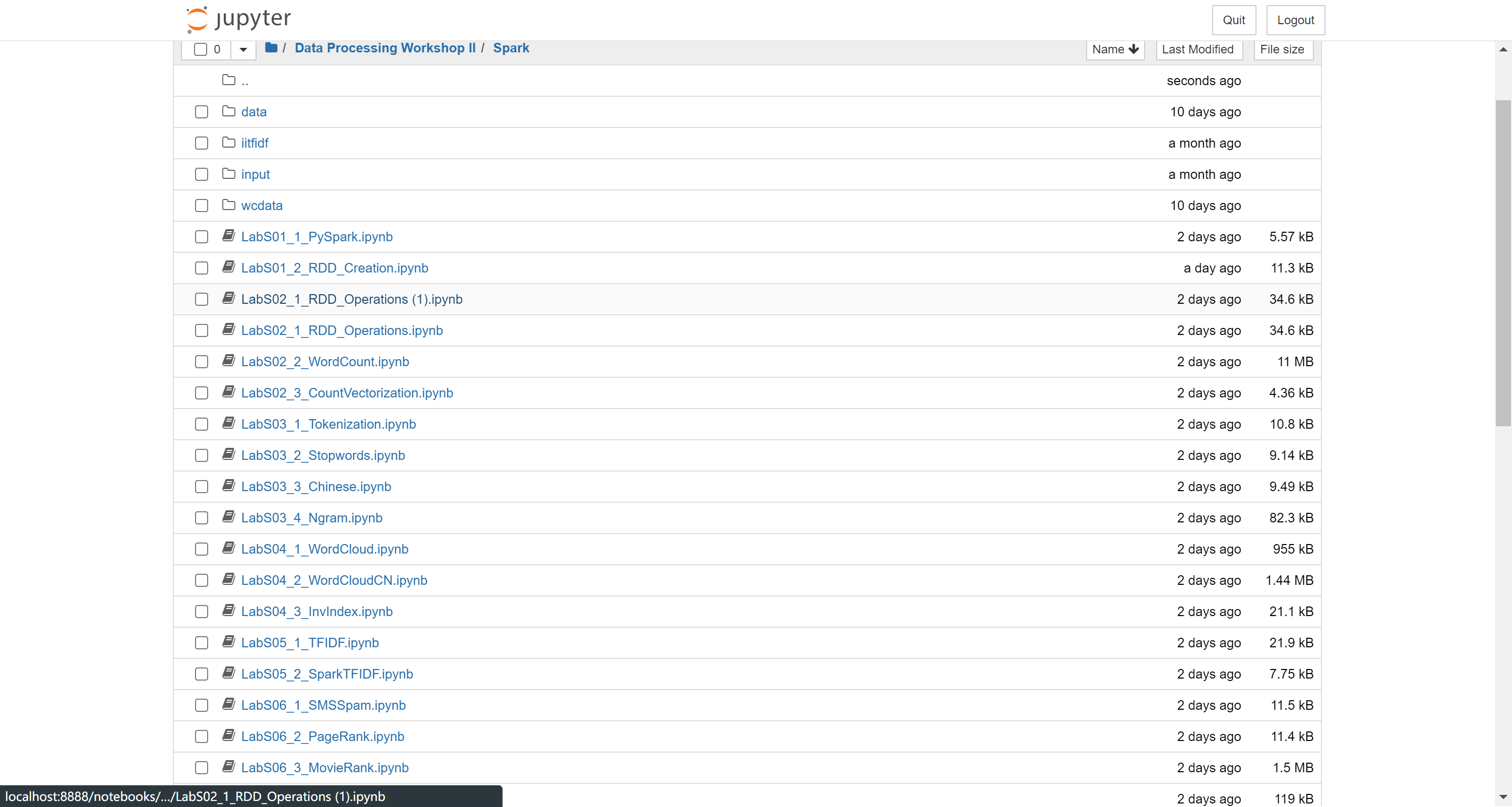


Get into the path you can see there are many .jar file and you can just download the

mysql-connector-java-8.0.26 in this folder.

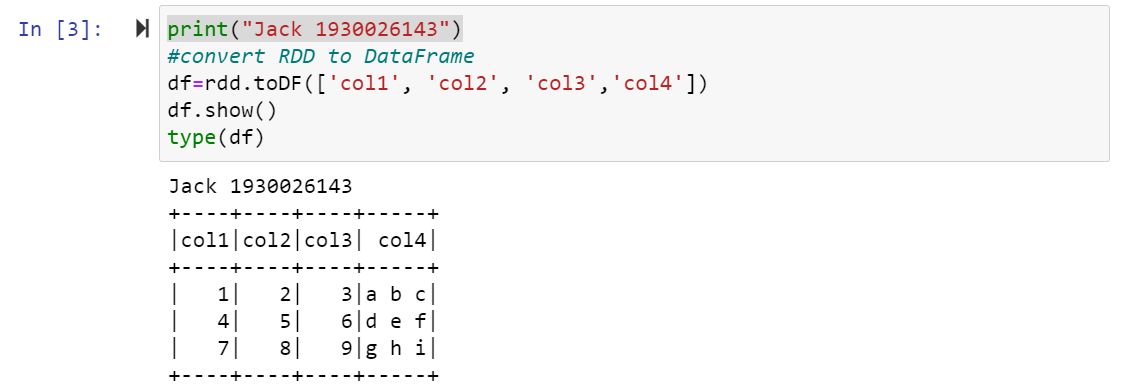


Then you can download the LabS01\_1\_PySpark.ipynb to your home directory and open your.



Open the LabS01\_1\_PySpark.ipynb:

Firstly, create RDD by using parallelize( ) function



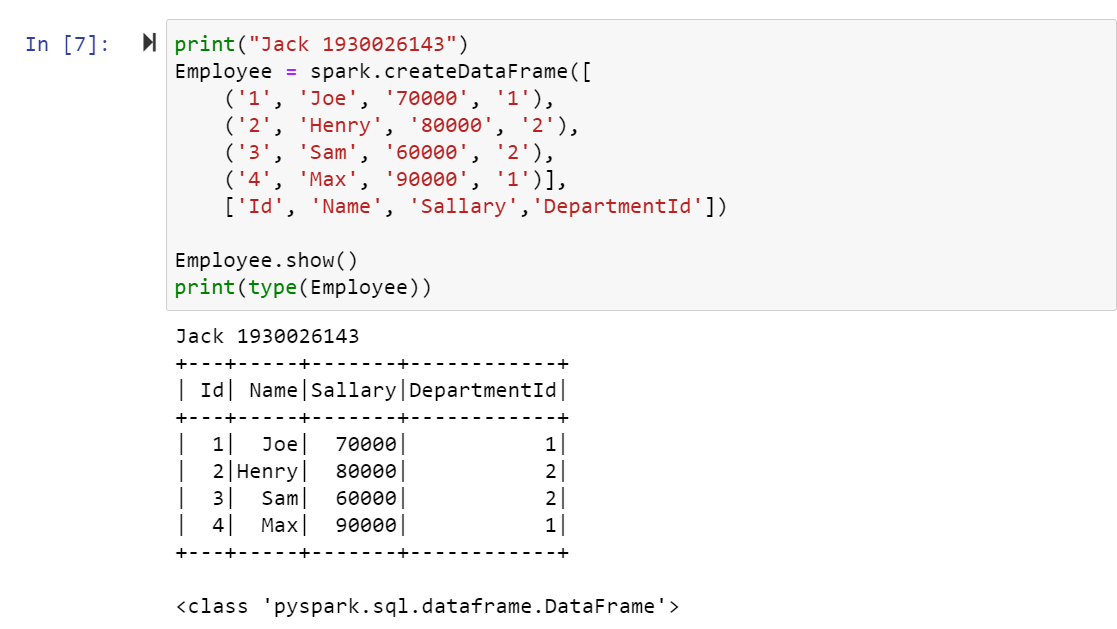
Convert an RDD to a list data type



convert DataFrame to RDD

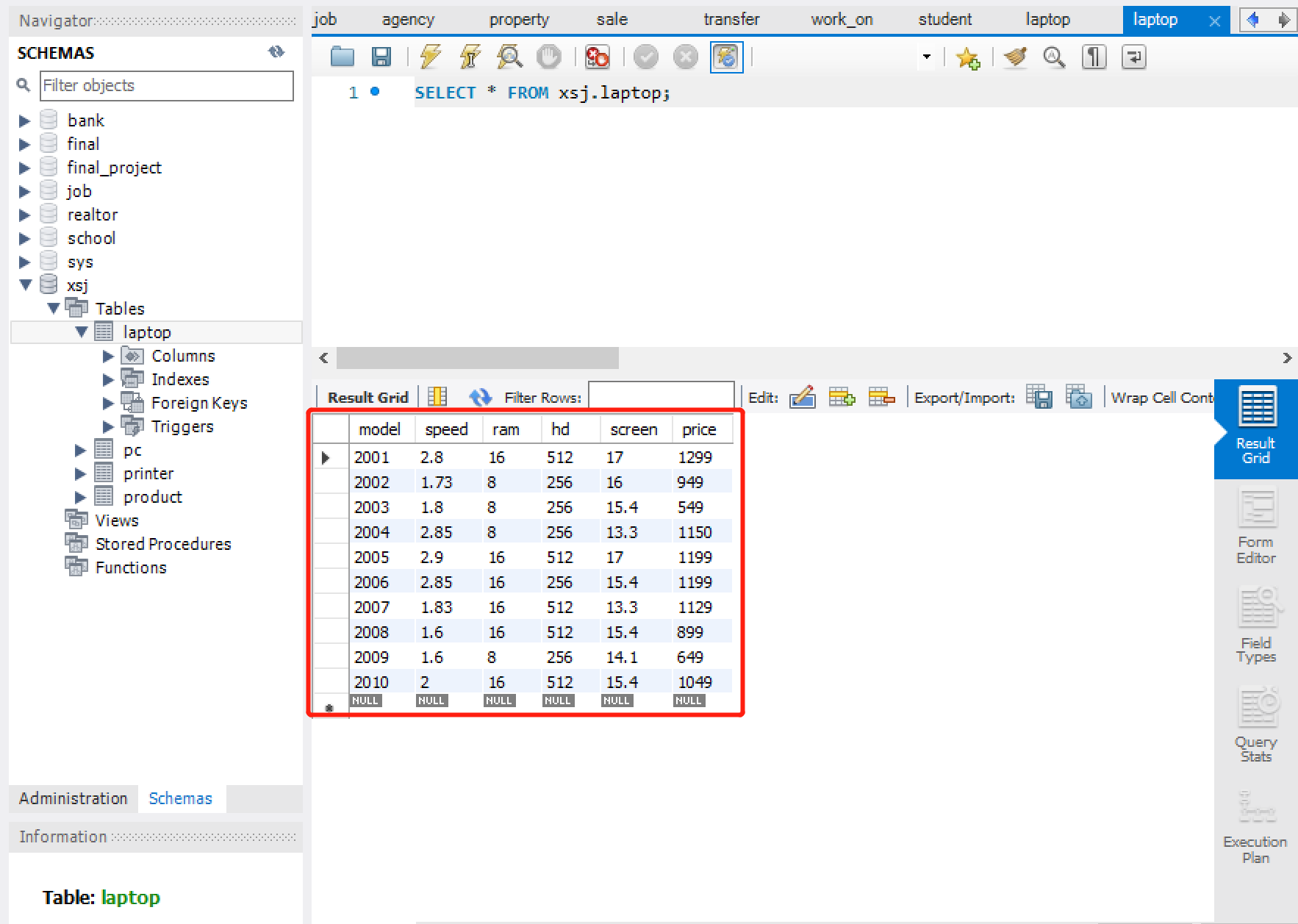


Then by using ( ) function

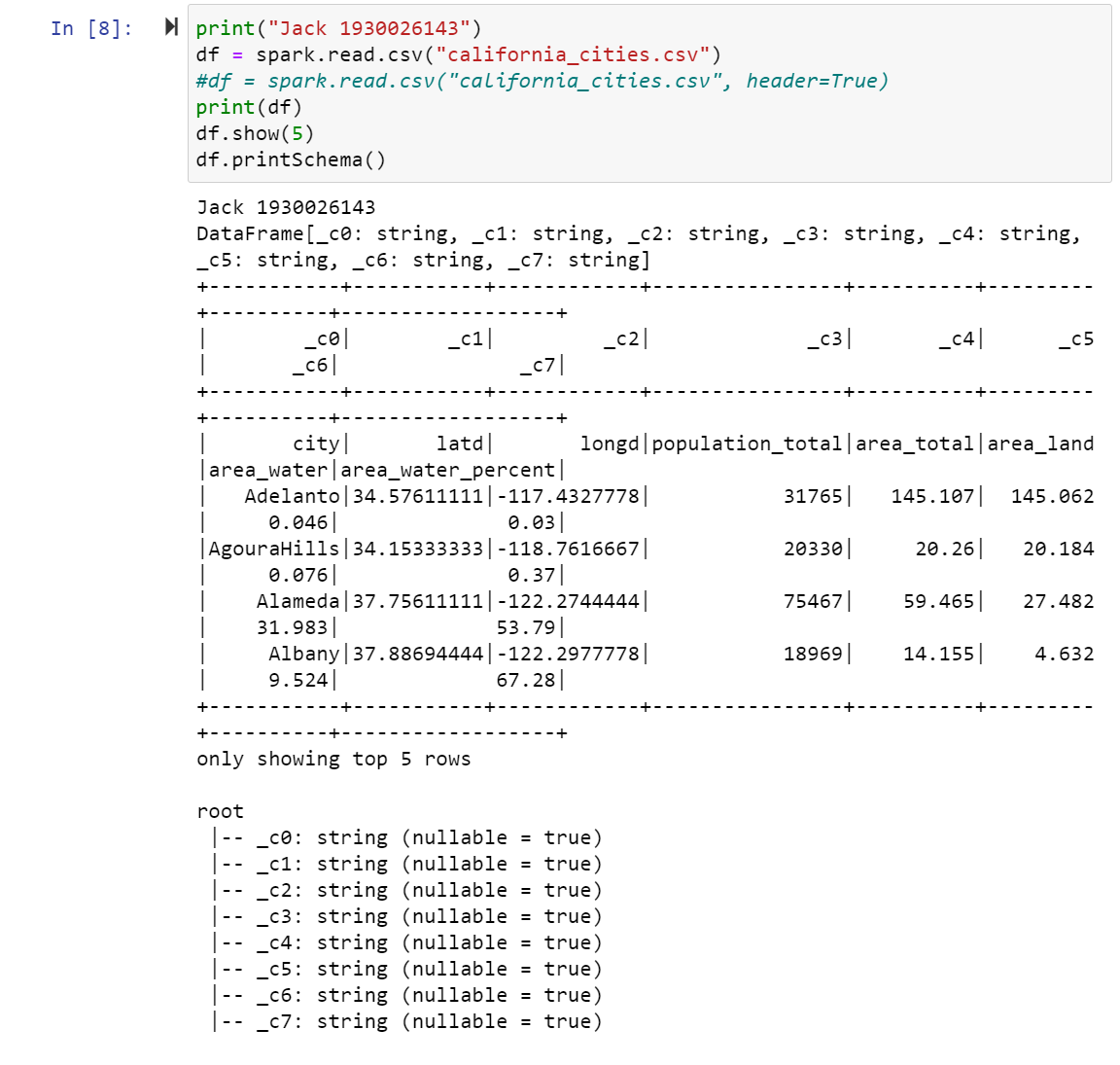


By using read and load functions, then verify that you can connect to your database.

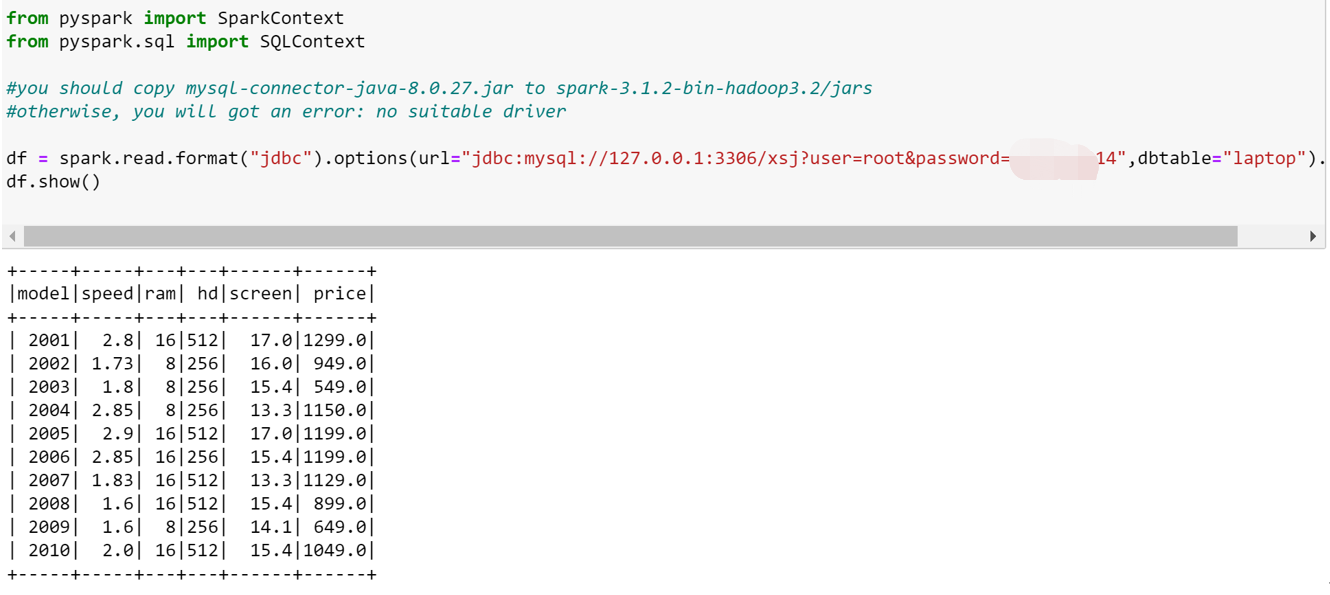
You can check your schema and the table name.



Change code with the username and the password of your MySQL and select the schema and specific table.



Read dataset from MySQL database:

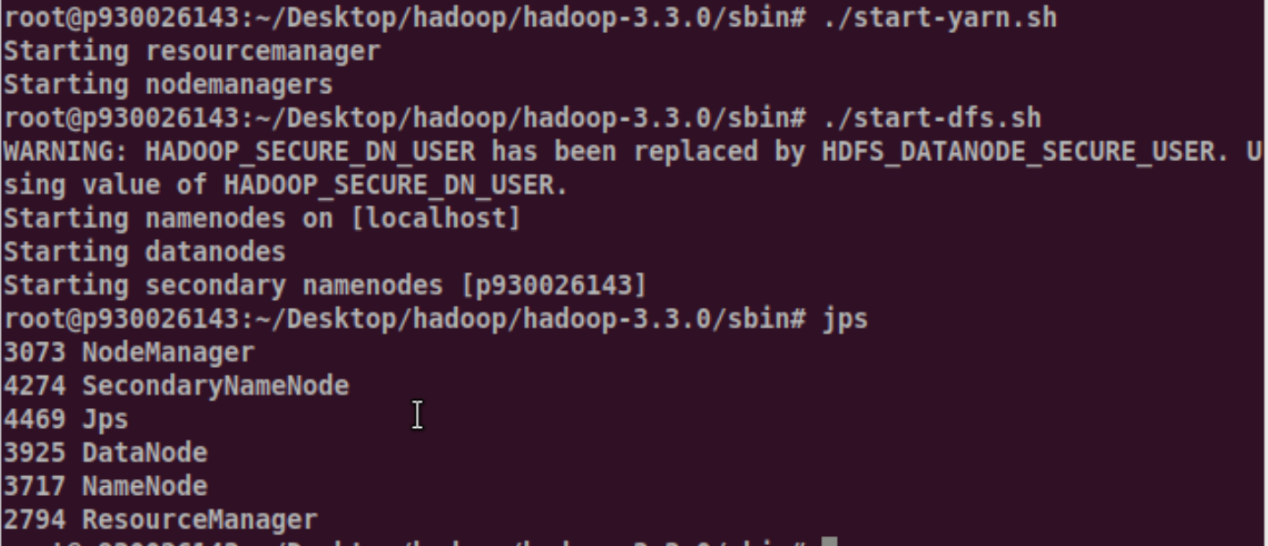


The last part: Read dataset from HDFS:

Enter the folder of :



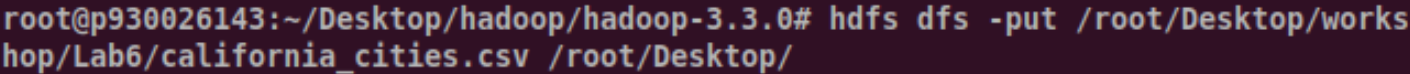
Start the and by the command by start-dfs.sh and start-yarn.sh



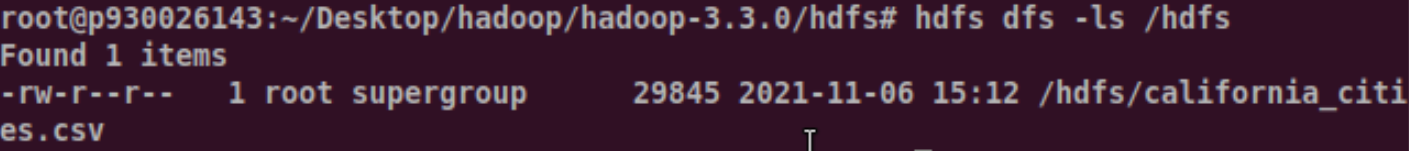
create a directory in hdfs: hdfs dfs -mkdir /hdfs



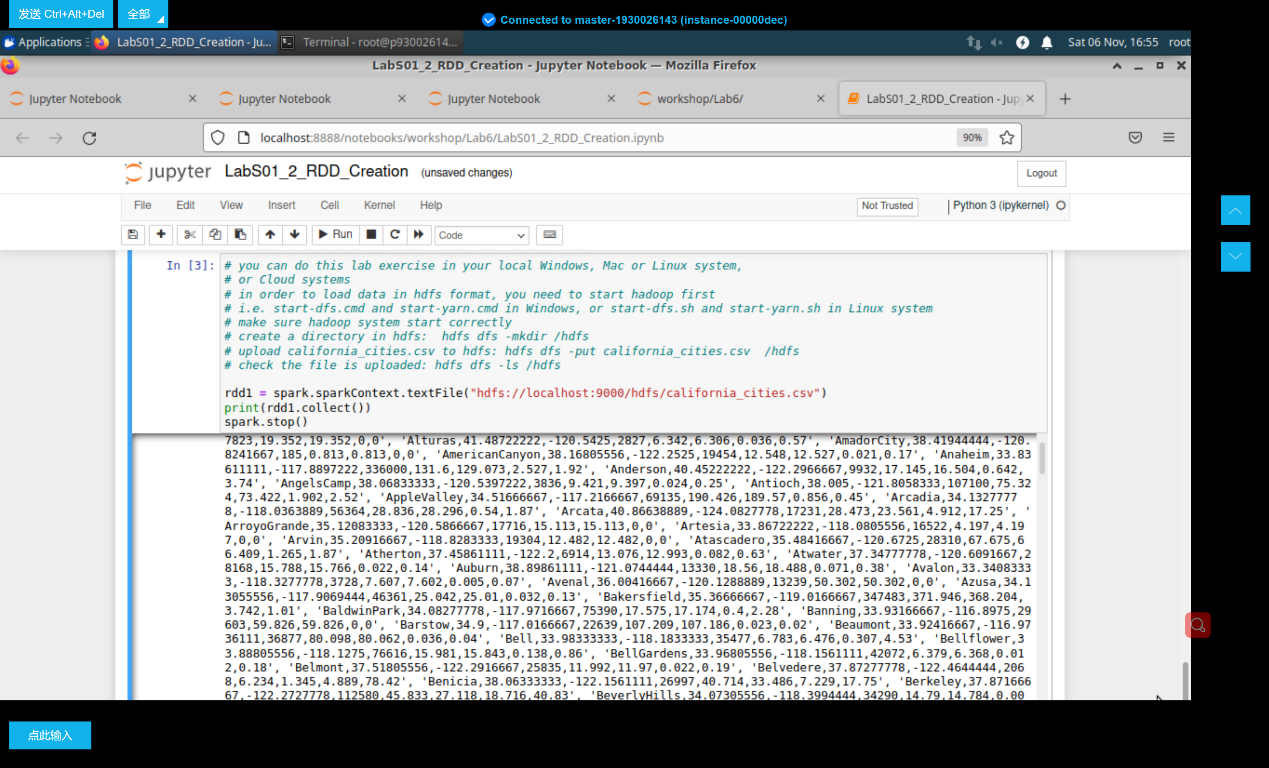
upload california\_cities.csv to hdfs: hdfs dfs -put california\_cities.csv /hdfs



check the file is uploaded: hdfs dfs -ls /hdfs



Run the code and if you can see the result without some errors, you are successful.



# Group bonus

Assume you have finishing login via RDP or VNC, install JDK

Firstly, update the system by typing the following command in Linux command line

**# cd ~/Downloads**

**# apt-get update && apt-get upgrade -y**

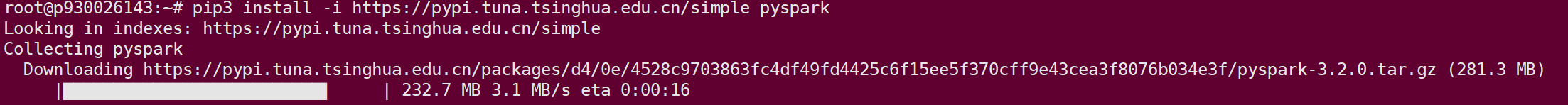
****

Then, install the pyspark, and set the spark-hadoop

**# pip3 install pyspark numpy -i https://pypi.mirrors.ustc.edu.cn/simple**

**# wget https://mirrors.bfsu.edu.cn/apache/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz**

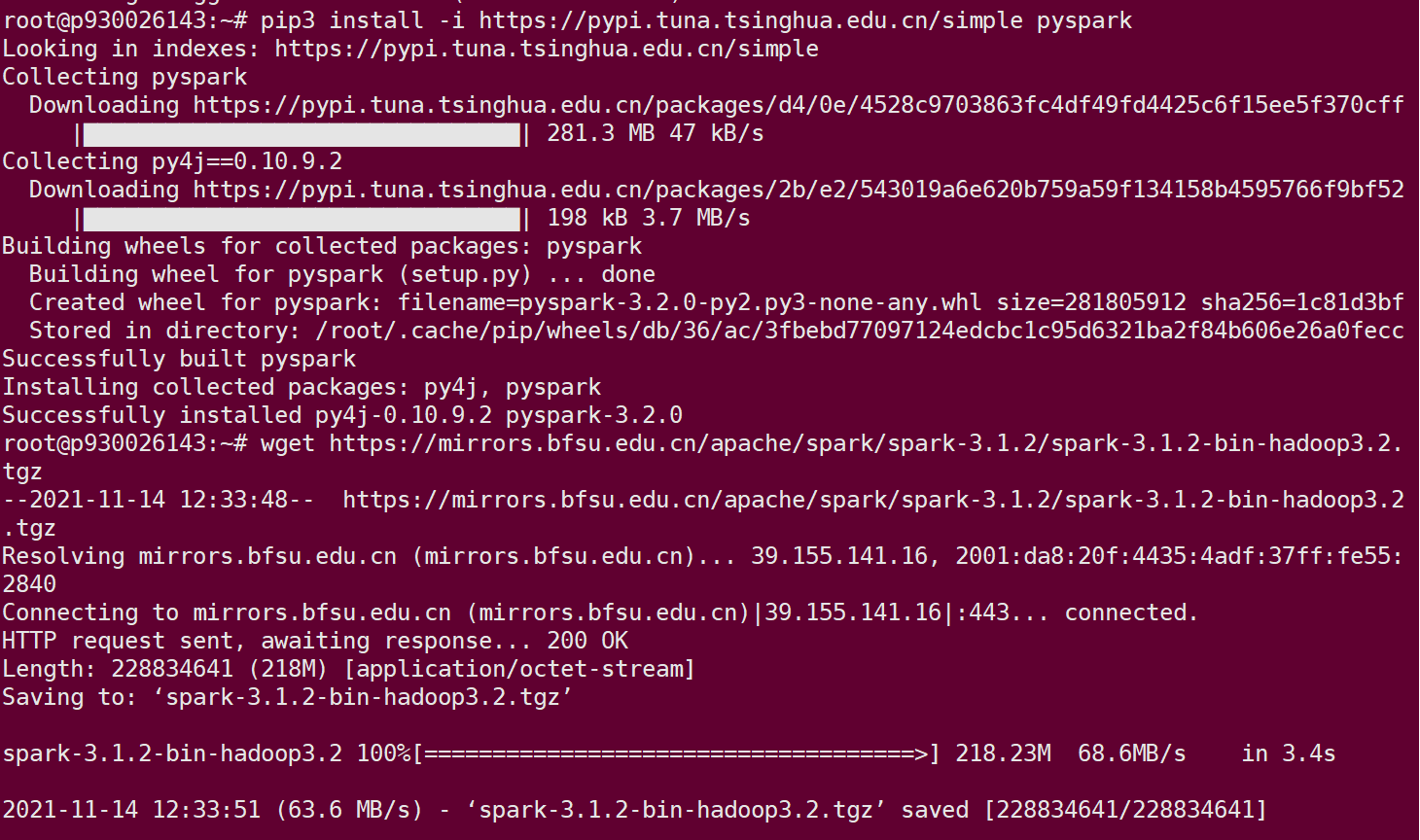
**# tar xfz spark-3.1.2-bin-hadoop3.2.tgz && init 6**



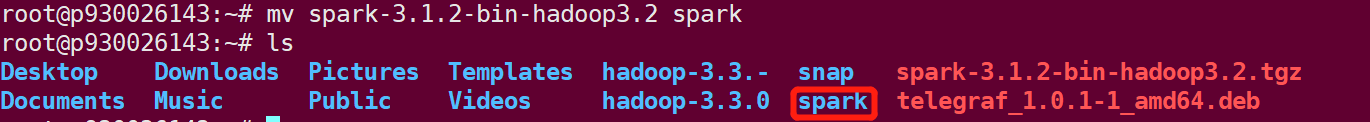
You can use the Domestic mirror https://pypi.tuna.tsinghua.edu.cn/simple

maybe you will meet the problem do not find pip, you can just install it

**# apt install python3-pip**



Change the **spark-3.1.2-bin-hadoop3.2 name to spark**



Assume that the Master named “master”, two slave name “slave1” and “slave2”

As for master

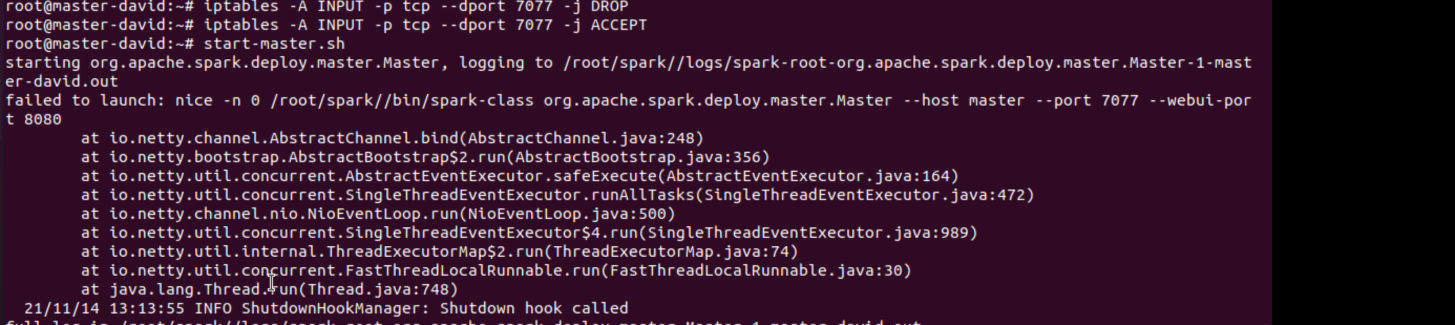
Edit the “~/Downloads/spark-3.1.2-bin-hadoop3.2/conf/spark-env.sh.temp” file and change the name of it to “spark-env.sh”

Then add the line

**export SPARK\_MASTER\_HOST=”** **master”**

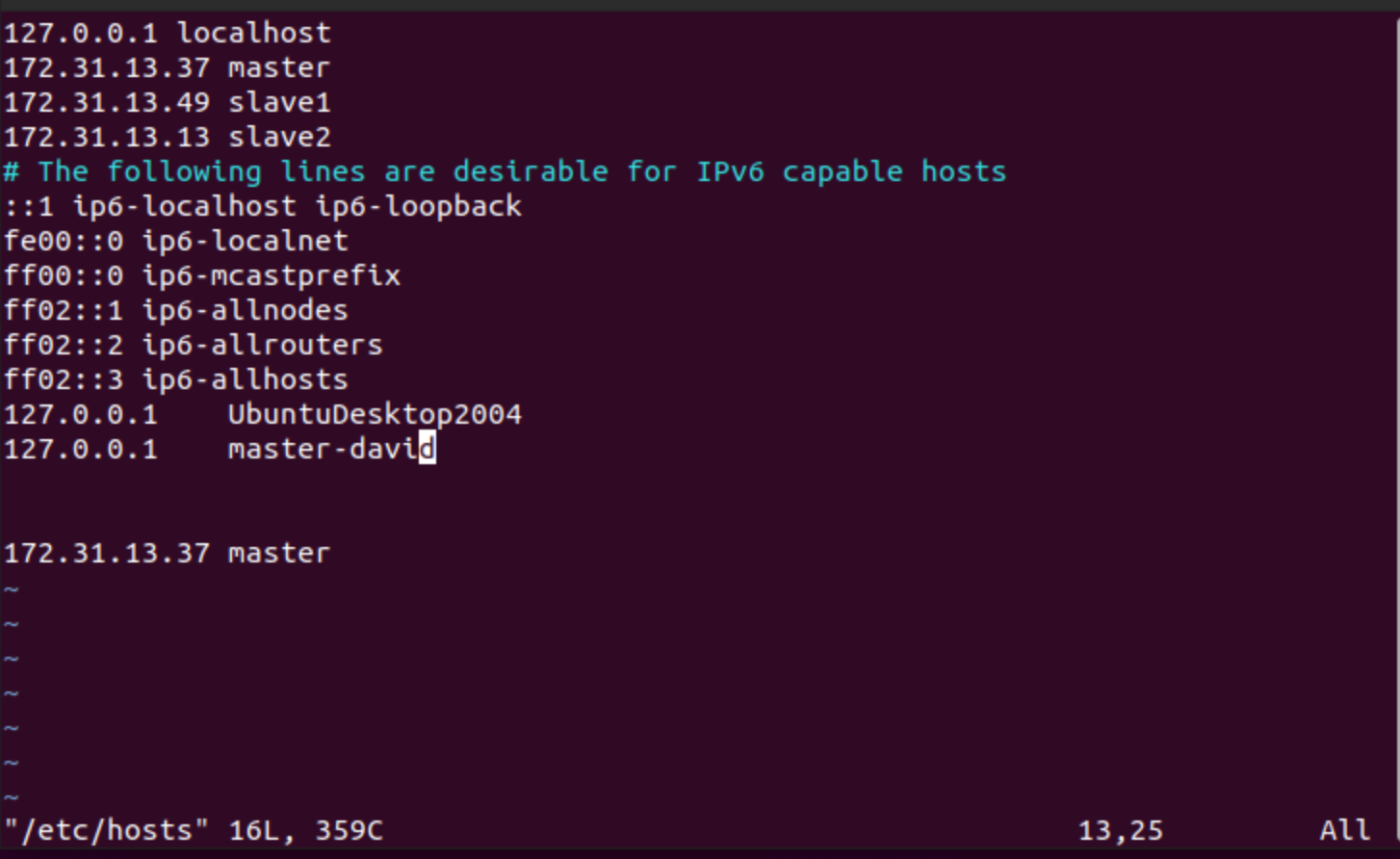
**export SPARK\_LOCAL\_IP=”10.20.3.0”**

You should append the host IP as local IP, not public IP, otherwise, you may meet the problem when execute the start-master.sh like the following

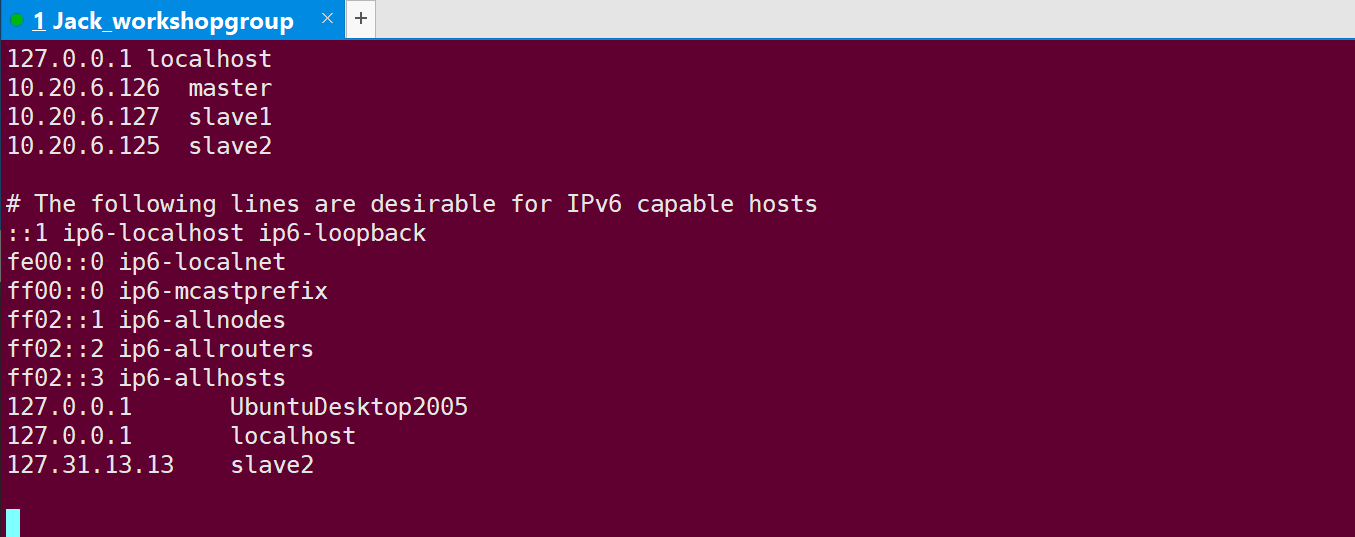


Then, for master and slave, you should modify the “/etc/host”

master:



slave1



Set up environment variable for spark

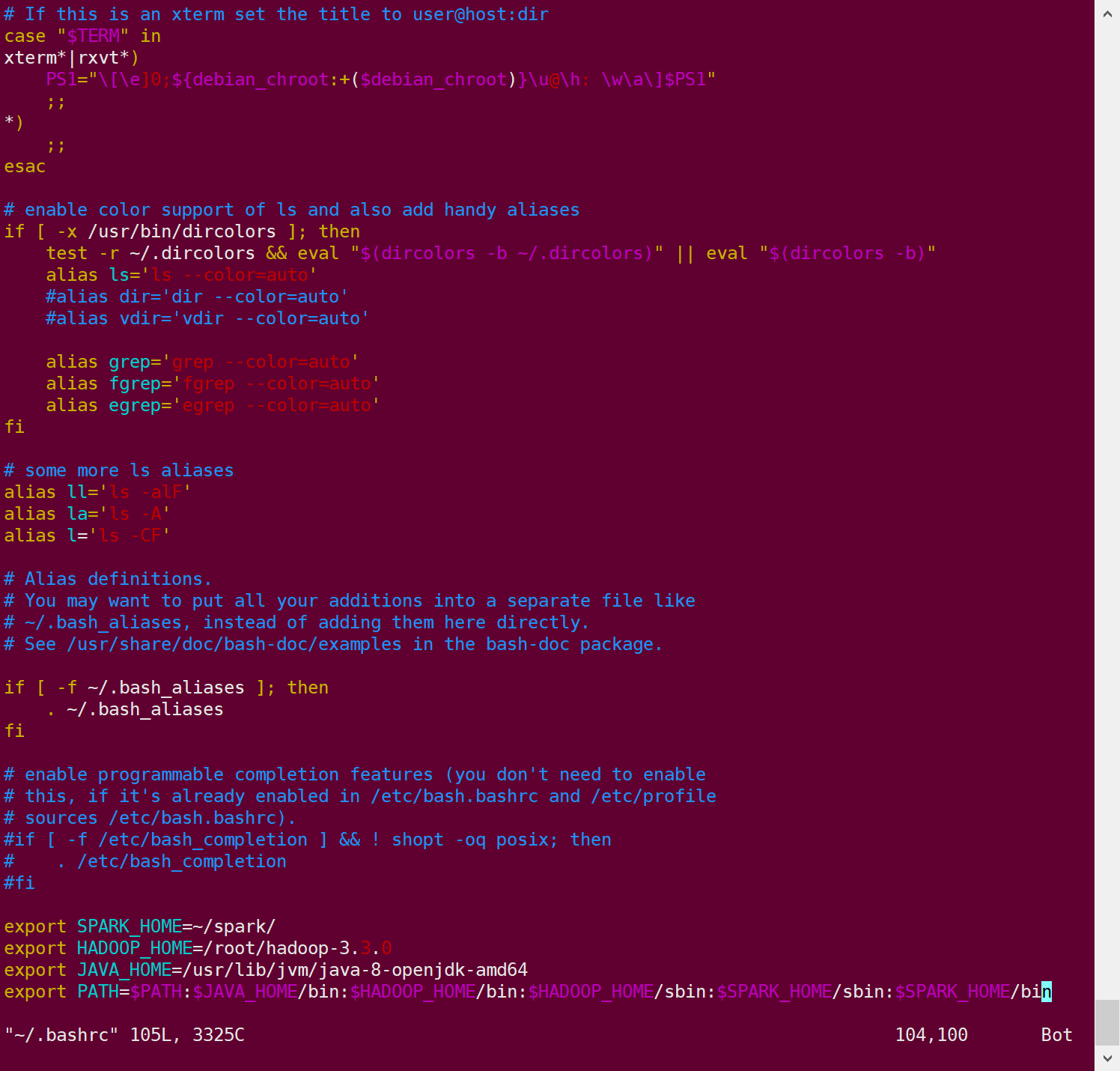
type the following command:

**$ sudo vi ~/.bashrc**

add these two lines

**export SPARK\_HOME=~/spark/**

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



save it and run by the command: source ~/.bashrc

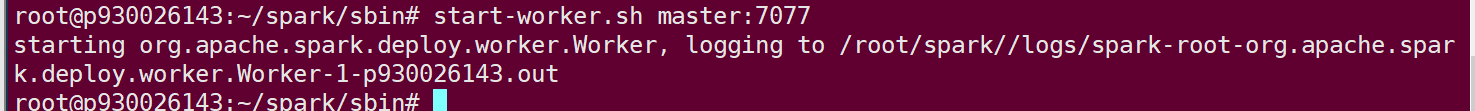
Run the spark master for master

**# start-master.sh**

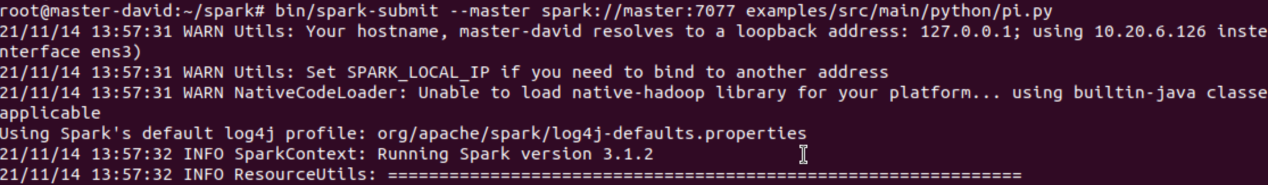


Run the spark worker for worker (yellow part should be the master name)

**# start-worker.sh master:7077**

****

On master



and type the url on web page <http://master:8080>

If you see the following result, then you do it successfully

