Spark Cluster User Manual For Python

To get account, contact Efe Çiftci(efeciftci@cankaya.edu.tr)

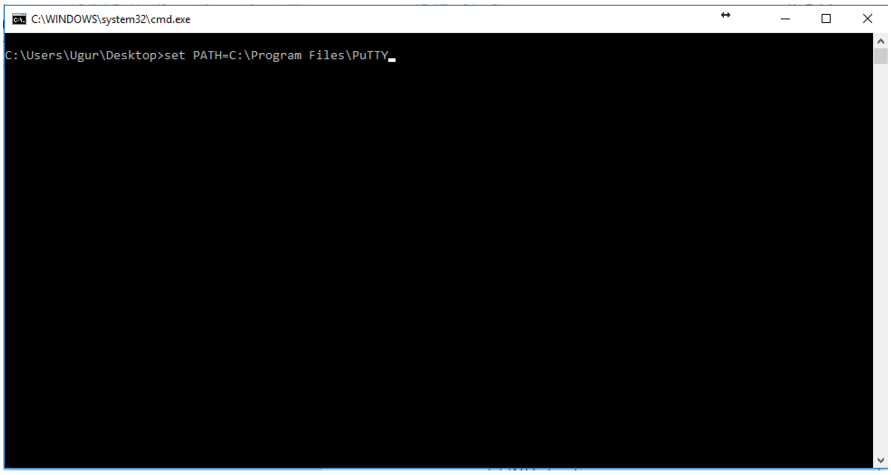
First of all you need to create jar file of your project (for scala and java). To do this you need to follow steps below.

1- To access the server at Çankaya University, you must also intall and run VPN program. You have to follow the steps documented in http://bim.cankaya.edu.tr/wp-content/uploads/sites/2/2018/01/VPN.pdf

2- You should transfer your .py file to the server. To do this in Windows, you should install Putty. After completing the installation, open “command prompt screen”
*If you use linux os, you should use terminal.

3- You need to set the path variable of pcsp which is installed with Putty. Run the following command:

```
set PATH=C:\Program Files\PuTTY
```



```
C:\Users\Ugur\Desktop>set PATH=C:\Program Files\PuTTY_
```

4- To copy the dataset used in your study and the .py file from local machine to server, you should run following command

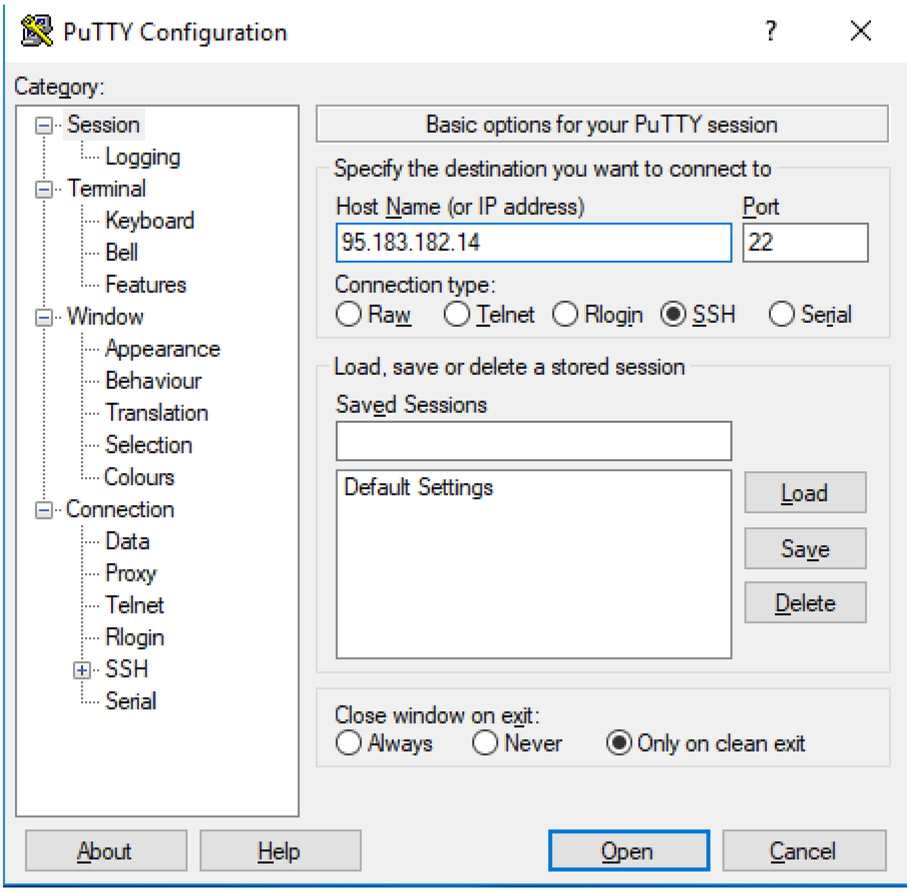
```
pscp <file> <username>@95.183.182.14:home/<username>/
```

Sample Command:

```
pcsp Sample.txt sparkuser@95.183.182.14:home/sparkuser/
```

Linux : scp <file> <username>@95.183.182.14:home/<username>/

5- Then you should connect the server with Putty. Write IP address of server(95.183.182.14) and port number then click the “Open” button
Linux : ssh <username>@95.183.182.14



Specify the destination you want to connect to

Host Name (or IP address): 95.183.182.14

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

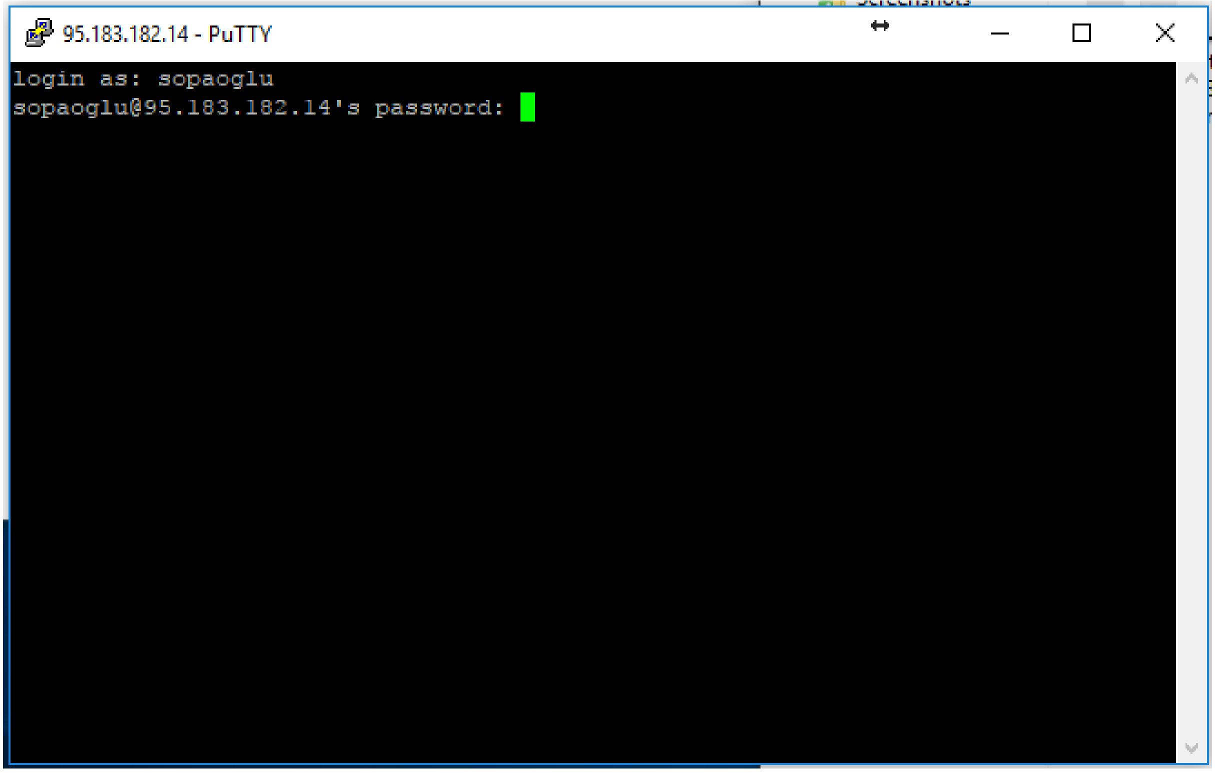
Load, save or delete a stored session

Saved Sessions

Default Settings

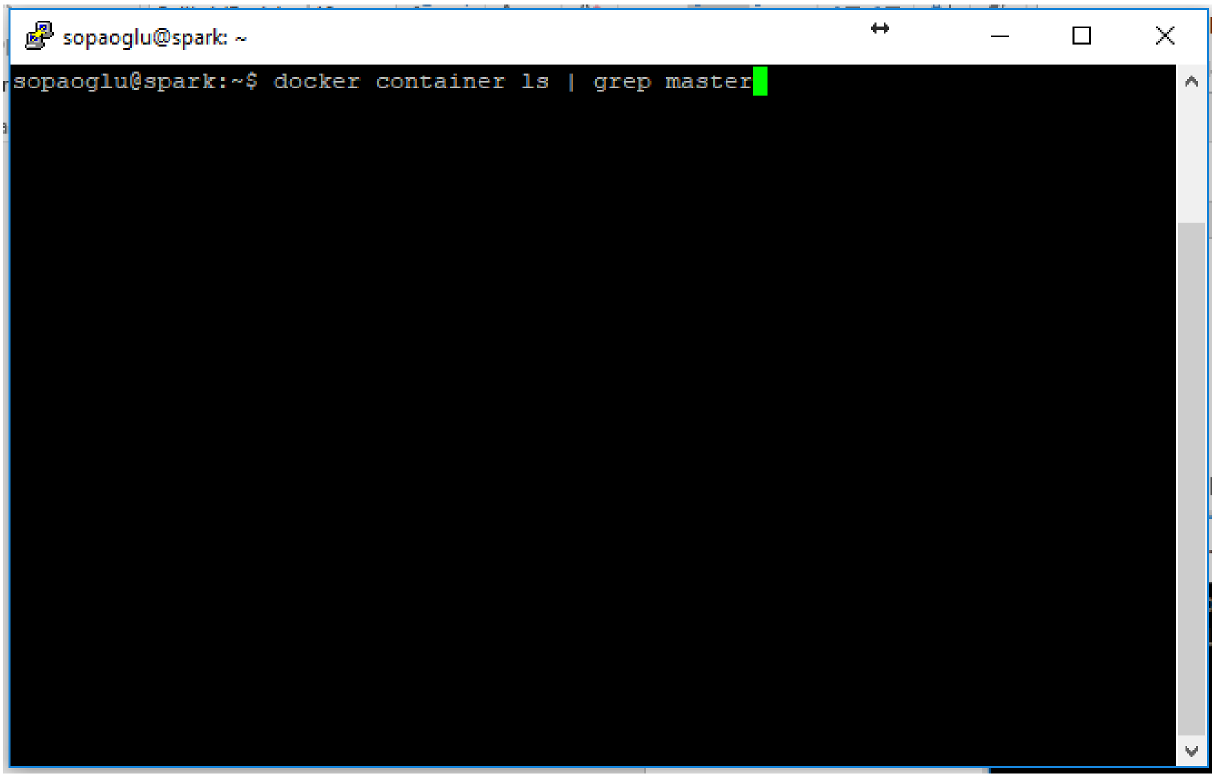
Close window on exit: ○ Always ○ Never ● Only on clean exit

6- Write your username and password

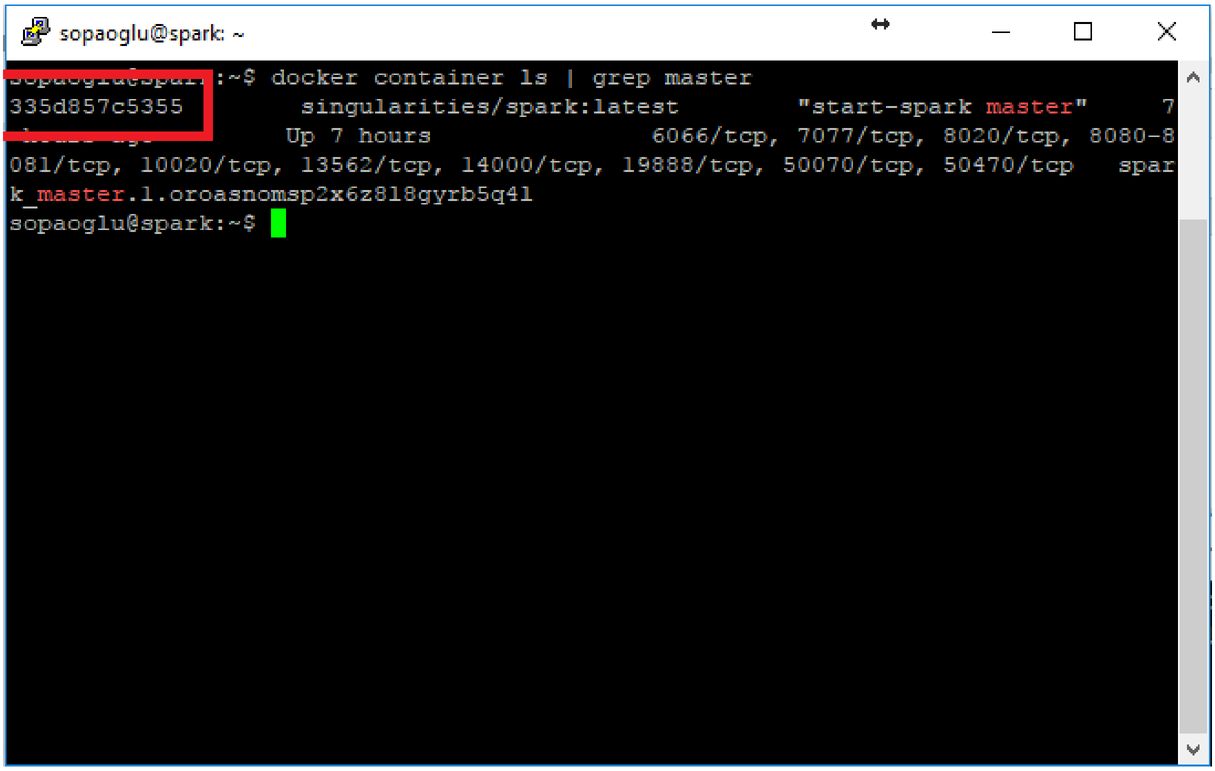


```
login as: sopaoglu
sopaoglu@95.183.182.14's password:
```

7- Now, you should identify the container id of master spark node. To do this, run following command:



```
sopaoglu@spark:~$ docker container ls | grep master
```



```
sopaoglu@spark:~$ docker container ls | grep master
335d857c5355        singularities/spark:latest        "start-spark master"     7
hours ago          Up 7 hours                  6066/tcp, 7077/tcp, 8020/tcp, 8080-8
081/tcp, 10020/tcp, 13562/tcp, 14000/tcp, 19888/tcp, 50070/tcp, 50470/tcp   spar
k_master.1.oroasnomsp2x6z818gyrb5q4l
sopaoglu@spark:~$
```

8- Now you should copy your **.py file** and **dataset** to the master node. You should run following command:

docker cp <your file> <container ID>:/spark

sample commands:
docker cp WordCoun.py 335d857c5355:/spark (**you should put your .py file under “spark” folder**)
docker cp Sample.txt 335d857c5355:/

9- Enviroment is ready to run your spark job. However, you should connect master node with following command

```
docker exec –it <Container ID of master spark node> bash
sample command:
docker exec -it 335d857c5355 bash
```

10- You should create /user/spark directory.

```
hdfs dfs –mkdir /user
hdfs dfs –mkdir /user/spark/
```

11- You should put your dataset into hdfs. You should use following command

```
hdfs dfs –put <your dataset> /user/spark/
sample command:
hdfs dfs –put Sample.txt /user/spark/
```

12- Now you can run your spark app with following command

```
spark-submit --master spark://master:7077 spark/<your .py file>
```

```
sample command:
spark-submit --master spark://master:7077 spark/WordCount.py
```

You can follow your spark app status from link below

95.183.182.14:8080

You can see cluster nodes from link below

95.183.182.14:8081