



How to use our cluster

Author: Haodong Zhao

Read in the browser:

<https://ubiquitous-bladder-8a5.notion.site/How-to-use-our-cluster-4e00f7230f5f43fe8455376f2aa0970c>

[0. Preparation](#)

[1. On your machine](#)

[2. On the VM](#)

[3. Test](#)

[Reference](#)

0. Preparation

1. Create a VM and associate a floating IP
2. Set Security Group `default`

1. On your machine

It's similar to what we did in A3.

Configure the `~/.ssh/config` on your local laptop/desktop/lab computer like this.

This is for the university lab machines or other unix-like systems and (Windows Subsystem for Linux) WSL, you may have to modify the instructions if you are using some other system):

- Replace the `130.238.x.y` to your VM's floating IP.
- Change the `IdentityFile` to the path where you put your key.

```
Host 130.238.x.y
  KexAlgorithms +diffie-hellman-group1-sha1
```

```
User ubuntu
# modify this to match the name of your key
IdentityFile ~/.ssh/id_rsa
# Spark master web GUI
LocalForward 8080 192.168.2.156:8080
# HDFS namenode web gui
LocalForward 9870 192.168.2.156:9870
# python notebook
LocalForward 8888 localhost:8888
# spark applications
LocalForward 4040 localhost:4040
LocalForward 4041 localhost:4041
LocalForward 4042 localhost:4042
LocalForward 4043 localhost:4043
LocalForward 4044 localhost:4044
LocalForward 4045 localhost:4045
LocalForward 4046 localhost:4046
LocalForward 4047 localhost:4047
LocalForward 4048 localhost:4048
LocalForward 4049 localhost:4049
LocalForward 4050 localhost:4050
LocalForward 4051 localhost:4051
LocalForward 4052 localhost:4052
LocalForward 4053 localhost:4053
LocalForward 4054 localhost:4054
LocalForward 4055 localhost:4055
LocalForward 4056 localhost:4056
LocalForward 4057 localhost:4057
LocalForward 4058 localhost:4058
LocalForward 4059 localhost:4059
LocalForward 4060 localhost:4060
```

2. On the VM

1. update apt repo metadata

```
# update apt repo metadata
sudo apt update
```

2. install java

```
# install java
sudo apt-get install -y openjdk-8-jdk
```

3. Edit `/etc/hosts`

Replace the `192.168.x.y` and `hostname` with your VM's IP and hostname.

And send the IP and hostname to me. I will add them all to `/etc/hosts` on every node in our cluster.

```
192.168.x.y hostname

192.168.2.156 master
192.168.2.85 node1
192.168.2.94 node2
192.168.2.182 node3
```

4. Env variable so the workers know which Python to use

```
echo "export PYSPARK_PYTHON=python3" >> ~/.bashrc
source ~/.bashrc
```

5. install git

```
sudo apt-get install -y git
```

4. install the python package manager 'pip' -- it is recommended to do this directly


```
sudo apt-get install -y python3-pip
```

5. install pyspark (the matching version as the cluster), and some other useful deps

- It should be **pyspark==3.2.1**

Slack

We are no longer supporting this browser, so you'll need to switch to one of our supported browsers to keep using Slack. We know this can be a pain, and we're sorry for asking you to do it. You can learn more about why we no longer support some browsers in our [FAQ](#).

 <https://app.slack.com/client/T02T9DFGWNB/C02T9DFHZ63/thread/C02T9DFHZ63-1647271091.859459>

```
python3 -m pip install pyspark==3.2.1 --user
python3 -m pip install pandas --user
python3 -m pip install matplotlib --user
```

6. Install Jupyter Lab

```
python3 -m pip install jupyterlab
```

7. Start the notebook

follow the instructions you see -- copy the 'localhost' link into your browser.

```
jupyter lab
```

3. Test

1. get our repo

```
git clone https://github.com/hd-zhao-uu/1TD169_Project.git
```

2. run wordcount.ipynb in the `src` folder

- change the `appName`

```
In [18]: from pyspark.sql import SparkSession
from operator import add
import re
import json
import time

In [19]: # New API
spark_session = SparkSession\
    .builder\
    .master("spark://master:7077") \
    .appName("haodong_zhao_wordcount")\
    .config("spark.executor.cores",2)\
    .config("spark.dynamicAllocation.enabled", True)\
    .config("spark.dynamicAllocation.shuffleTracking.enabled", True)\
    .config("spark.shuffle.service.enabled", False)\
    .config("spark.dynamicAllocation.executorIdleTimeout", "30s")\
    .config("spark.driver.port", 9998)\
    .config("spark.blockManager.port", 10005)\
    .getOrCreate()

# Old API (RDD)
spark_context = spark_session.sparkContext

spark_context.setLogLevel("WARN")
```



It's time to start our project! I have built up the cluster and shown how to connect to the cluster. After you have successfully completed the setting, we should discuss what tasks each person should be responsible for. $\forall (\geq \nabla \leq *)o$

Reference

spark-jupyter/spark_cluster_deployment.txt at main · JSFRi/spark-jupyter

You can't perform that action at this time. You signed in with another tab or window.
You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

 https://github.com/JSFRi/spark-jupyter/blob/main/spark_cluster_deployment.txt

JSFRi/**spark-jupyter**



 1 Contributor  0 Issues  2 Stars  3 Forks