You can choose “VirtualBox” or docker to build a Hadoop/Spark clusters, and you should know that docker is more lightweight to build a cluster than using virtual machines. **You are not requested to use Docker**, but I think it is a good solution. This manual cannot cover all the details and you may find some problems with them, it is necessary for you to search for more materials.

We provide a pre-defined docker, you can pull it from **https://hub.docker.com/**.

The docker’s name is **nusbigdatacs4225/ubuntu-with-hadoop-spark.**

It is a ubuntu image included

1. jdk1.8.0_191(/usr/java)
2. Hadoop 2.8.5(/usr/local/hadoop)
3. Spark 2.2.0(usr/local/spark)

You can use this image to build your own clusters. (you need to check and change the configurations for your own environment)

Here is a simple example for setting a three nodes Hadoop cluster, you can write a script to do all this automatically.

1.download the docker and finish the configuration for the image.

“docker pull nusbigdatacs4225/ubuntu-with-hadoop-spark”

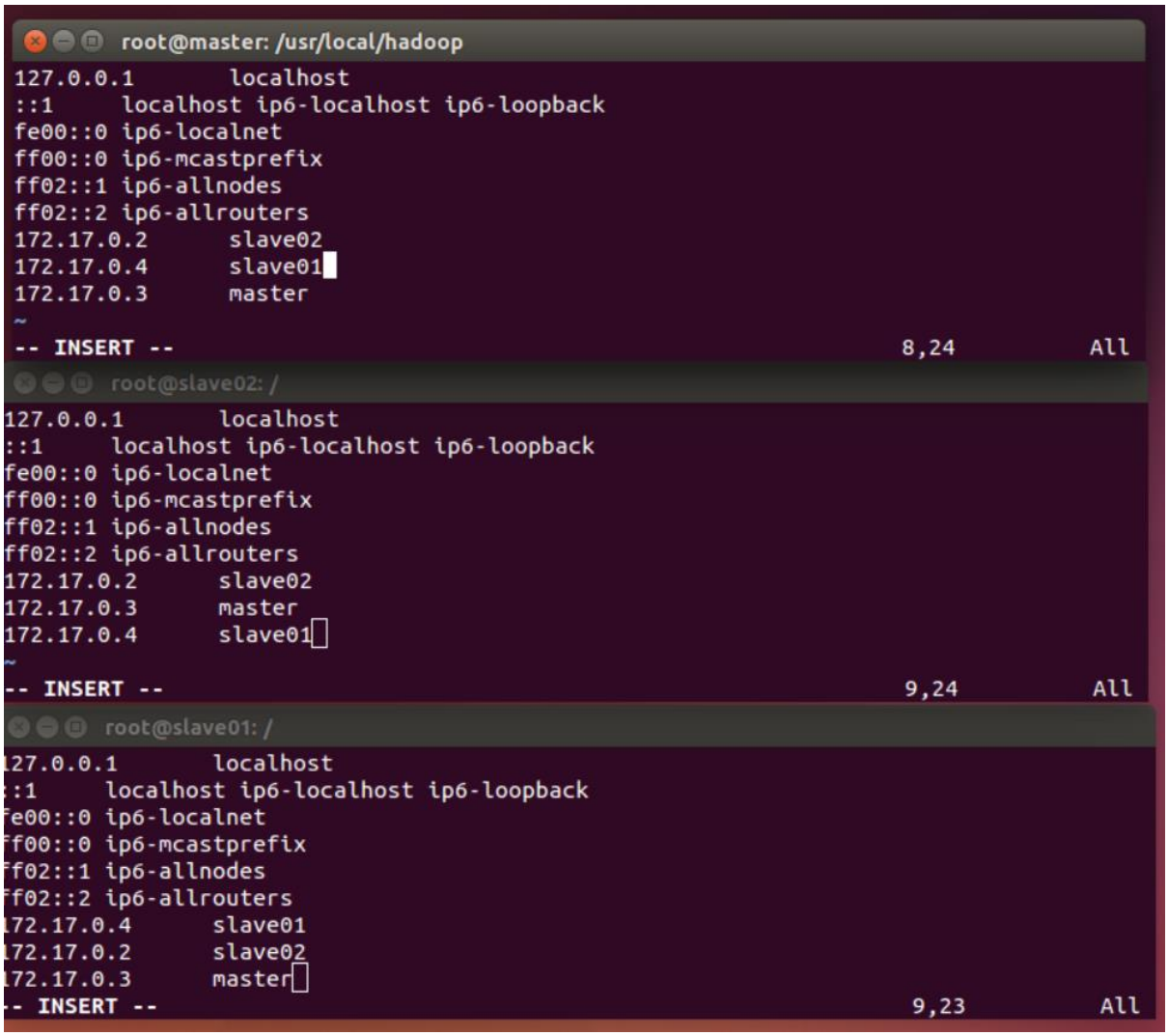
2.run the commands to create three containers.

“docker run -it -h master --name master nusbigdatacs4225/ubuntu-with-hadoop-spark”
“docker run -it -h slave01 --name slave01 nusbigdatacs4225/ubuntu-with-hadoop-spark”

“docker run -it -h slave02 --name slave02 nusbigdatacs4225/ubuntu-with-hadoop-spark”

3.run **“vim /etc/hosts”** to check IP address.
e.g.
master: 172.17.0.2
slave01:172.17.0.3
slave02:172.17.0.4

add this to all three containers. (you can use “**ssh slave01“**in master container to check this configuration)

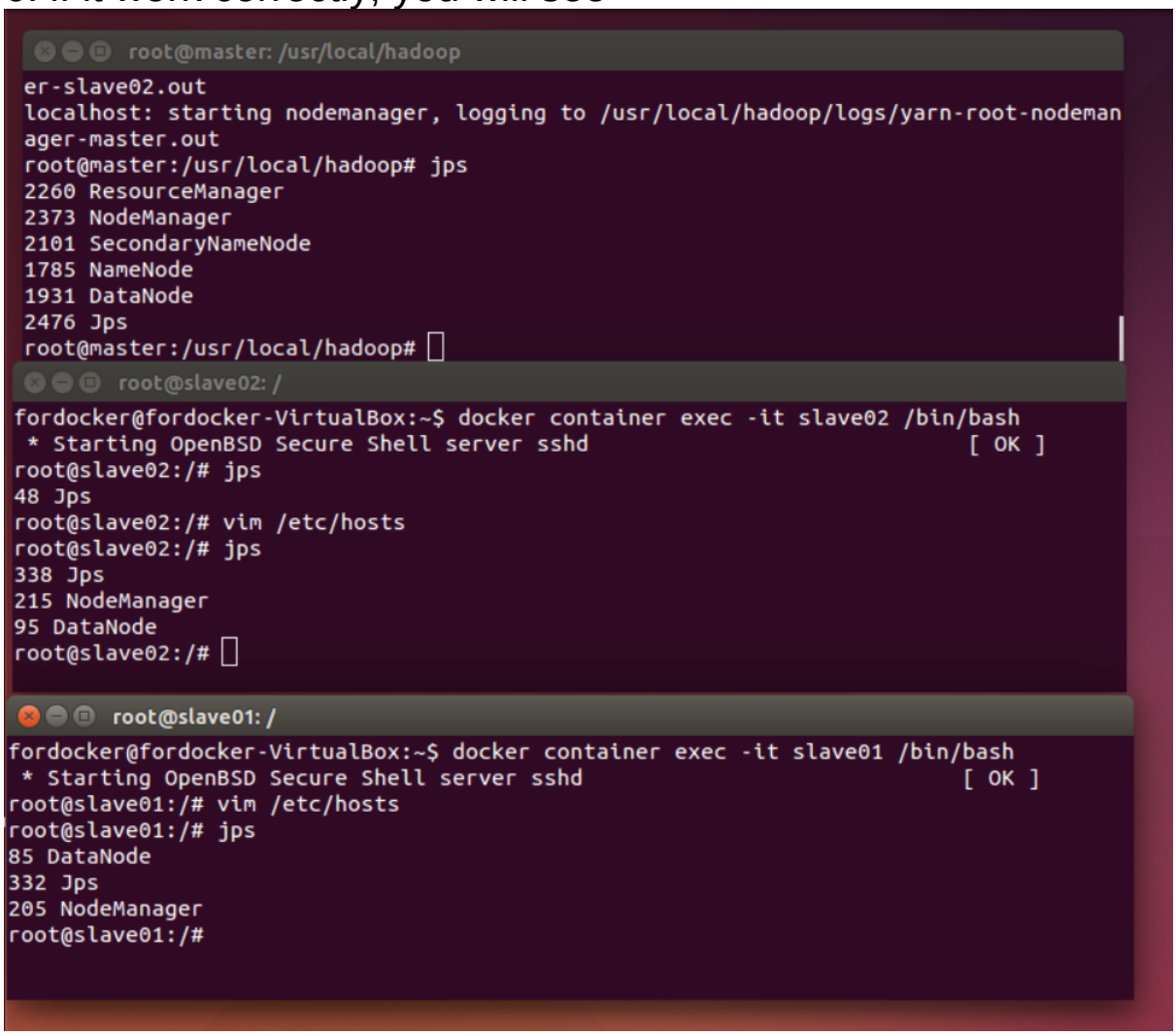


4**.**In master container, cd /usr/local/hadoop/etc/hadoop, run “**vim slaves**”**,** add **slave01 slave02** into this file.

5.initialize the hdfs and run

“cd /usr/local/Hadoop”
“bin/hdfs namenode -format”
“sbin/start-all.sh”

6. if it work correctly, you will see



For spark environment you need to do it by yourselves.