In this document, we are going to describe how to build the cluster with our pre-configured dockers.

Notice: This manual may not cover all the details. It is necessary for you to search for more materials. You can also refer to other documents to set up a Hadoop cluster as long as there are at least 2 slaves nodes and one master node. There are some references:

1. <https://clubhouse.io/developer-how-to/how-to-set-up-a-hadoop-cluster-in-docker/>
2. <https://github.com/kiwenlau/hadoop-cluster-docker>

We provide a pre-defined docker, you can pull it from [**https://hub.docker.com/**](https://hub.docker.com/). (use sudo)

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. Type “sudo docker images” to see the downloaded images. (it is **nusbigdatacs4225/ubuntu-with-hadoop-spark** in my machine.)

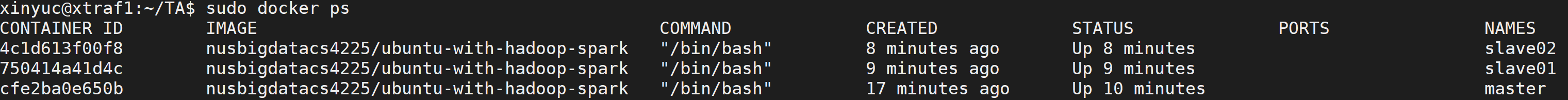
2. Create three containers by typing the followings (the creating order matters):

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

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

**You may use “Ctrl + P + Q” to exit the running containers and create the other containers.**

**Type “sudo docker ps” to see the running containers. (something like the below.)**



3. Type “**sudo docker attach [container ID]**” to open the interactive terminal with a docker and then run **“vim /etc/hosts”** to edit IP address.

172.17.0.2 master

172.17.0.3 slave01

172.17.0.4 slave02

add these lines to the opened file of all three containers.

(In master node, type “ssh slave01” then “yes” and type “ssh slave02” then “yes” to pass the verifications.

The same thing needs to be done on other two nodes to make sure they can connect to each other.)

4**.** In master node, “cd /usr/local/hadoop/etc/hadoop”, run “**vim slaves**”**,** delete **localhost** andadd **slave01 slave02** into this file.

5.In master node, initialize the hdfs and run

**“cd /usr/local/hadoop”**

**“****bin/hdfs namenode -format”**

**“****sbin/start-all.sh”**

Now, you should have successfully built a cluster with a master node and two slave nodes.

You may find you cannot directly run hadoop commands. You can add hadoop to your PATH for the shell: (1) add “export PATH=$PATH:/usr/local/hadoop/bin/” in “~/.bashrc” (2) run command “source ~/.bashrc”.

6. Package your code and submit your applications to the cluster.

You can follow **Testing your Hadoop cluster** section of this link: <https://clubhouse.io/developer-how-to/how-to-set-up-a-hadoop-cluster-in-docker/> to have a simple try.