In this document, we are going to describe how to simulate 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 and finish the configuration for the image.

2.run the commands to create three containers.

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

**“docker run -it -h slave02 --name slave02 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 verify the configuration)

4**.** In master node, “cd /usr/local/Hadoop/etc/Hadoop”, run “**vim slaves**”**,** add **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”.

For spark environment you need to do it by yourselves.