**Installation of Hadoop in Azure cloud using Docker**

**1. Creating a azure VM instance in cloud using free subscription or students subscription**

**https://azure.microsoft.com/en-in/free/students/**

**Ram minimum - 4GB**

**https://www.youtube.com/watch?v=fnqCSQ3WGTw**

**Note:**

**Please note down the public IP Address of the VM**

3. **Once VM is successfully created,**  Download putty in your laptop. You can connect to your Azure VM using Putty with the help of your IP Address and credentials of your vm.

https://www.putty.org/

**3. we would follow the below steps to install the docker in the Azure VM.**

4. Next step is to install the docker container. Please follow the below steps

Execute the following set of commands given below:

1 sudo apt update

It is used to update the apt package index

2 sudo apt install apt-transport-https ca-certificates curl software-properties-common

It is used to install packages to allow apt to use a repository over HTTPS

3 install the docker io

sudo apt install docker.io-

4. Once docker is successful installed. we can test the installation by below command

docker --version should give you the details about the docker.

5. Let us install the hadoop in docker by Importing the Cloudera QuickStart Image from the docker Hub

sudo docker pull cloudera/quickstart:latest

sudo docker images - should display the cloudera image.

6. Run the hadoop docker container

sudo docker run --hostname=quickstart.cloudera --privileged=true --name docker\_name -t -i -v /Users/hack:/src --publish-all=true -p 8888:8888 -p 8088:8088 -p 8042:8042 -p 19888:19888 cloudera/quickstart /usr/bin/docker-quickstart

Note: Extra information regarding the docker run command

<https://docs.docker.com/engine/reference/commandline/run/>

7. Once it is run, you can check whether a new container is created using the below command

sudo docker ps

Note down the containerid.

<https://docs.docker.com/engine/reference/commandline/ps/>

8. Open the ports so that Hadoop web ui is available from your laptop.

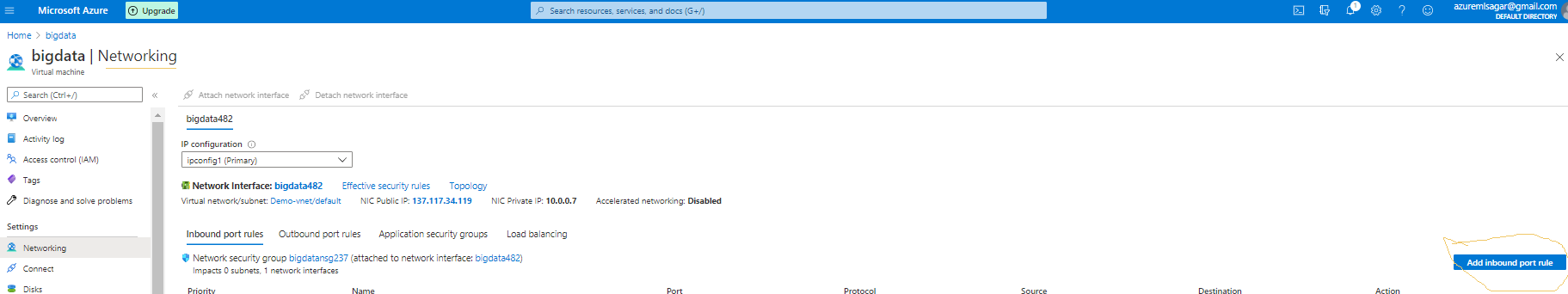
a)To check the Port Mapping from docker to Azure ports

docker inspect [CONTAINER ID]

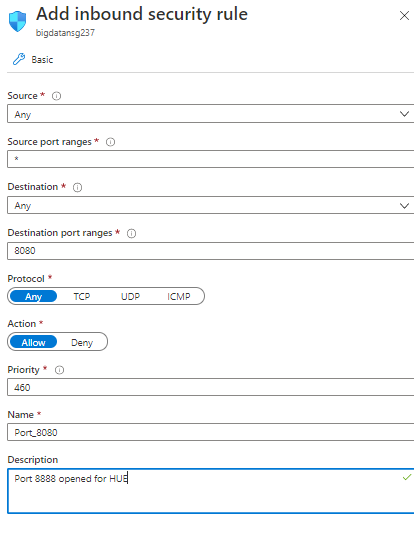


**b) Open the ports 8088,8888,8042,19888 in azure portal**

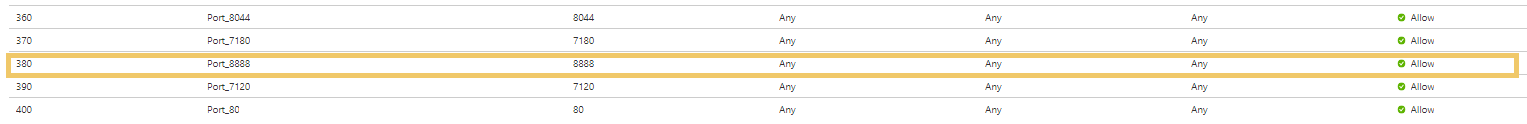
Click the networking option in azure portal in your subscription



Add inbound rules for port 8888 for hue



c)You can verify it on the dashboard screen once added successfully



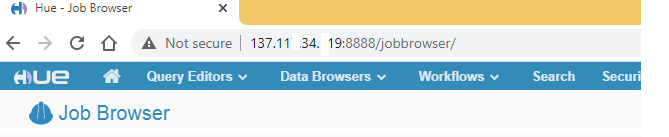
d) Similarly add the inbound rules for port 8088,19888 and 8042.

**9.Accessing Hue from your windows machine**

To check the successful execution of the Job you can use the Hue interface with the application id. IP address is the public address of your Azure VM.

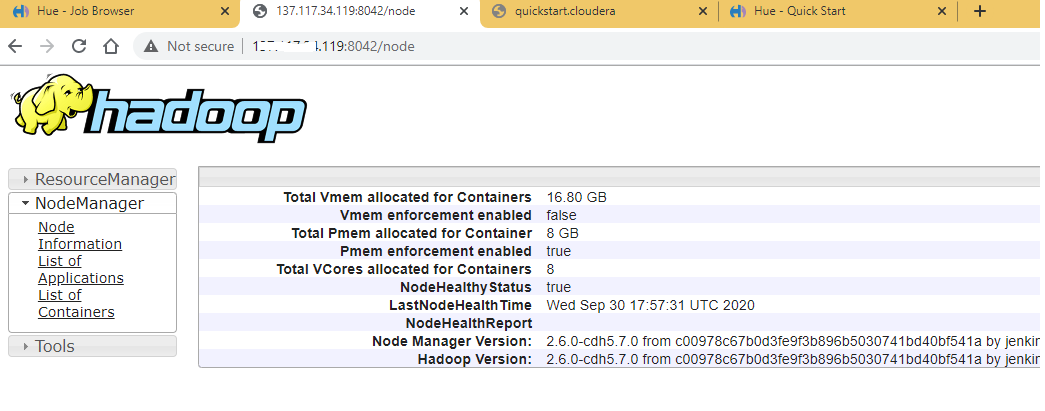
Use that ip and form the below url

[http://<ipaddress>:8888/](http://192.168.56.102:8888/about/)



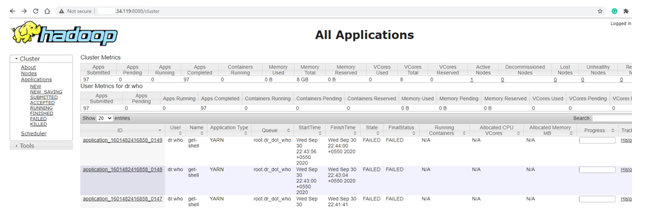
Resource Manager

[http://<ipaddress>:8088/](http://192.168.56.102:8888/about/)



Job History Server

[http://<ipaddress>:8042/](http://192.168.56.102:8888/about/)



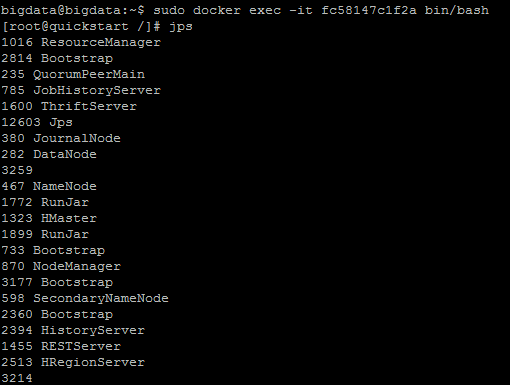
10. Get the docker command line and run the hadoop command

sudo docker exec -it <containerid> bin/bash

hadoop fs -ls , it will list all the files in the users home directory

11. To check whether all the hadoop services are running, Run the jps command in your docker container

> jps



Please check whether the below services are running

1. Resource Manager

2. Namenode

3. NodeManager

4. DataNode

5. Job History Server

6. Journal Node

If any of the services are not running.

docker restart containerid

**Appendix**

**1. Useful docker commands**

a) sudo docker stop [CONTAINER ID]

It is used to stop one or more running containers.

<https://docs.docker.com/engine/reference/commandline/stop/>

b) command you can use again to restart your container.

docker restart containerid

2. Development of MR Job

Refer documentation - Setting up\_Hadoop\_Project\_Eclipse document

3. Deploying Jar into hadoop and executing the Jar

**a)Transferring the MR JAR file to Azure Virtual Machine**

To transfer the file from your laptop windows machine to a Azure virtual machine using winscp.User your azure machine ip address and credentials

[**https://data-flair.training/blogs/transfer-data-from-windows-to-ubuntu/**](https://data-flair.training/blogs/transfer-data-from-windows-to-ubuntu/)

b) Transfer the file from azure VM to hadoop docker container src directory

sudo docker -cp <containerid>:/src file

c) Execute the Jar with the below command

su cloudera // so all the jobs are executed under user cloudera by default it is root

yarn jar MR\_WordCnt.jar WCDriver webin/ out/ -Dlog4j.configuration=log4j.properties -D mapred.map.tasks = 1

or

hadoop jar MR\_WordCnt.jar WCDriver webin/ out/ -Dlog4j.configuration=log4j.properties -D mapred.map.tasks = 1

jarName: MR\_WordCnt.jar

className: WCDriver

in inputfilePathInhdfs

out outputfilePathInhdfs

-Dlog4j.configuration=log4j.properties - This file should be their where you are running the jar, so it overrides other log4j.property file available in hadoop by default.

-D mapred.map.tasks = 1 - Map task is one so it is easy to debug. Otherwise there would be multiple map task if your file is greater than 128MB(default block size).

Note:

if hadoop command is not recognised means you are not in docker bash

sudo docker exec -it <containerid> bin/bash

7. The input files needs to be available in hdfs. so the local file input.txt needs to be moved to hdfs using the below command

# create the in directory in hdfs

**hdfs fs -mkdir webin**

# move the file from local file system to hdfs

**hdfs fs -put input.txt in/**

hdfs fs -ls /user/cloudera/in

9. Check the job status and the log files for errors in Hue using the job browser with the application ID created on submission. You and can get that application ID from the console itself after starting your MapReduce job.

It will be similar to as shown below.

18/07/11 14:39:23 INFO impl.YarnClientImpl: Submitted application application\_1531299441901\_0001

9. On successful execution check the output file in the hdfs out directory mentioned while running the MR job

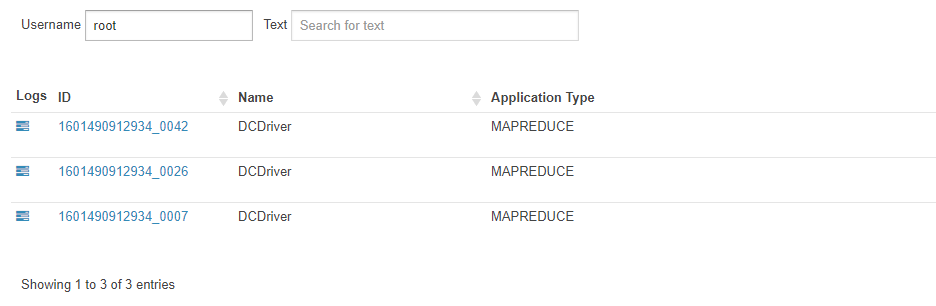
**hdfs dfs -cat /out/part-r-00000**

**Sample Output file: part-r-00000**

d) checking the job status in Hue

[http://<ipaddress>:8888/](http://192.168.56.102:8888/about/)

Click on Job browser and search for root user jobs. Then click on the job id and check for logs.



**E. Debugging your mapreduce code**

**Job History Server WebUI: <ur ipaddress of your vm >:19888**

