Instructions for installing virtual box VM on windows and mac for running Hadoop Lab Software

Minimum System Requirements:

- 8 GB of RAM
- 20 GB of hard-drive space
- 2 GHz dual core processor
- Windows or Mac

If you do not meet the requirements above or if you use an Apple Silicon device, please use the remote desktop solution; the instruction for that is available on LMS, in the "VM Installation" section.

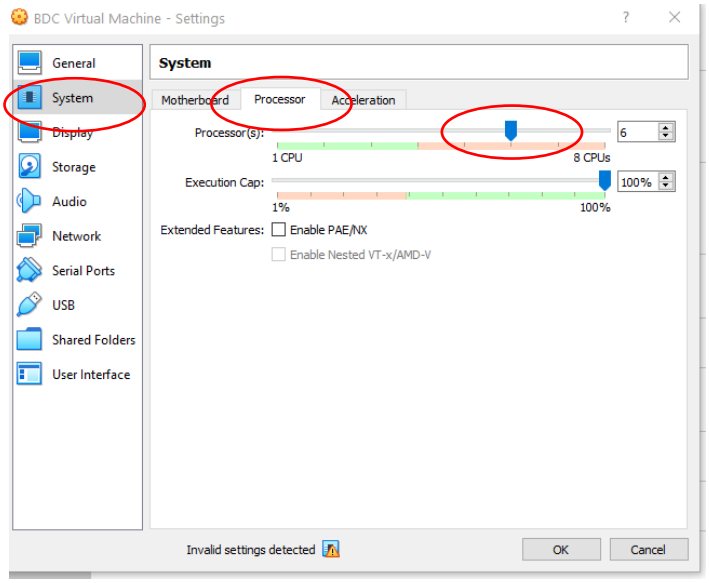
Installation Guide

Here are the steps to follow to install the required lab software for labs 2 onwards for CSE3BDC.

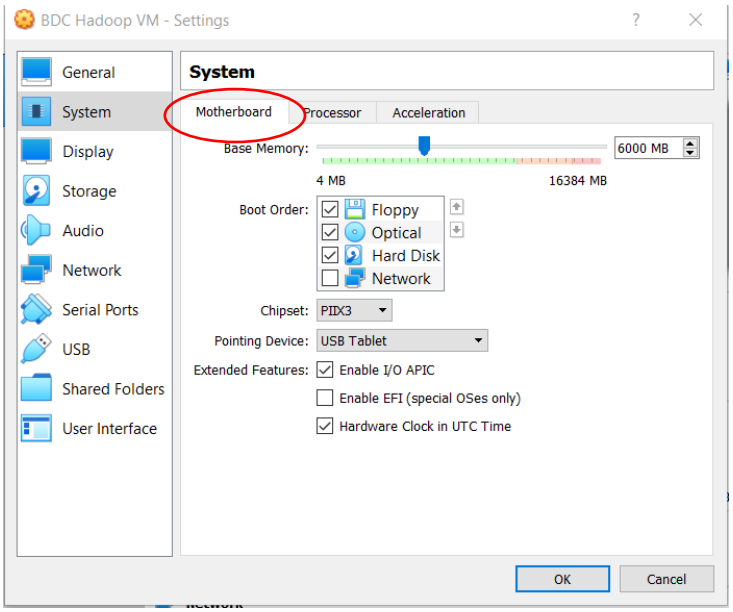
1. Download and unzip the image file from the following link:
    a. BDC VM Image

    Install VirtualBox.

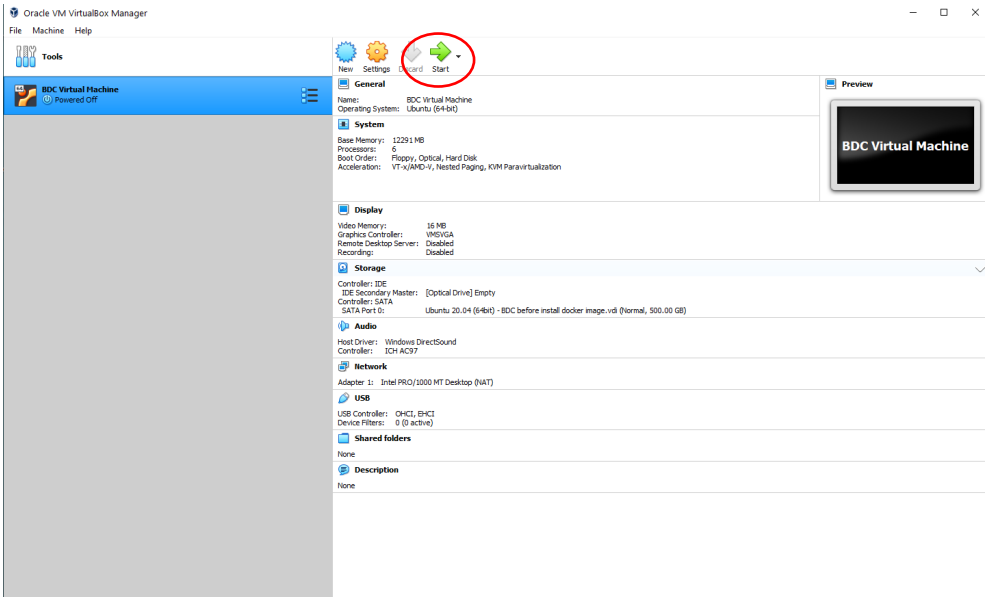
    Please download the latest version of virtual box from here:
    - https://www.virtualbox.org/wiki/Downloads
2. Go to the unzipped folder *BDC Hadoop VM* folder.
3. Double click on the *BDC Hadoop VM.vbox* file.
4. Go to System settings, and under "Processor" tab, change the number of processors to the following. If you have max of 4 CPUs, then choose 2 processors. If you have a max of 8 CPUs, change to 4. If you have max of 6 CPUs then change to 3. If you find these settings make the rest of your system run really slow then you can reduce the number of CPUs assigned to the VM to be within the green region.



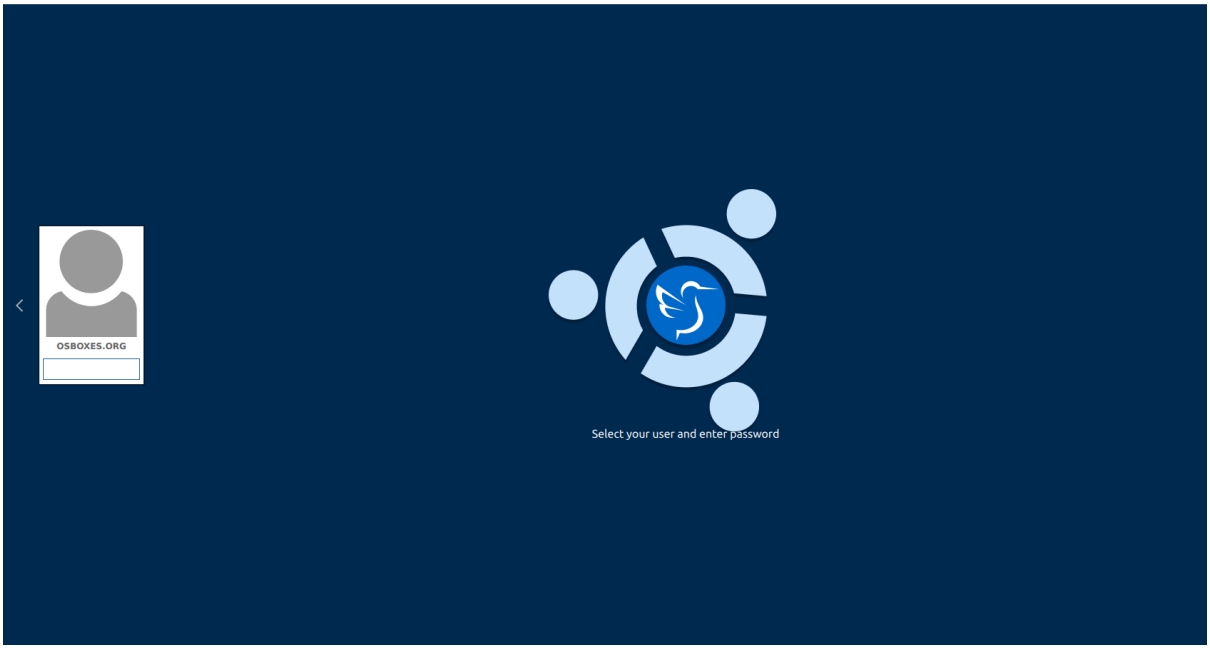
5. Next in the motherboard tab select the amount of RAM you want the VM to use. If you have 8 GB of RAM, choose 5 GB for your memory size; if you have 16 GB, choose 8 GB. If you're not sure what your memory size is, just follow the rule of thumb of dragging the slider to be near the left edge of the green part (like below).



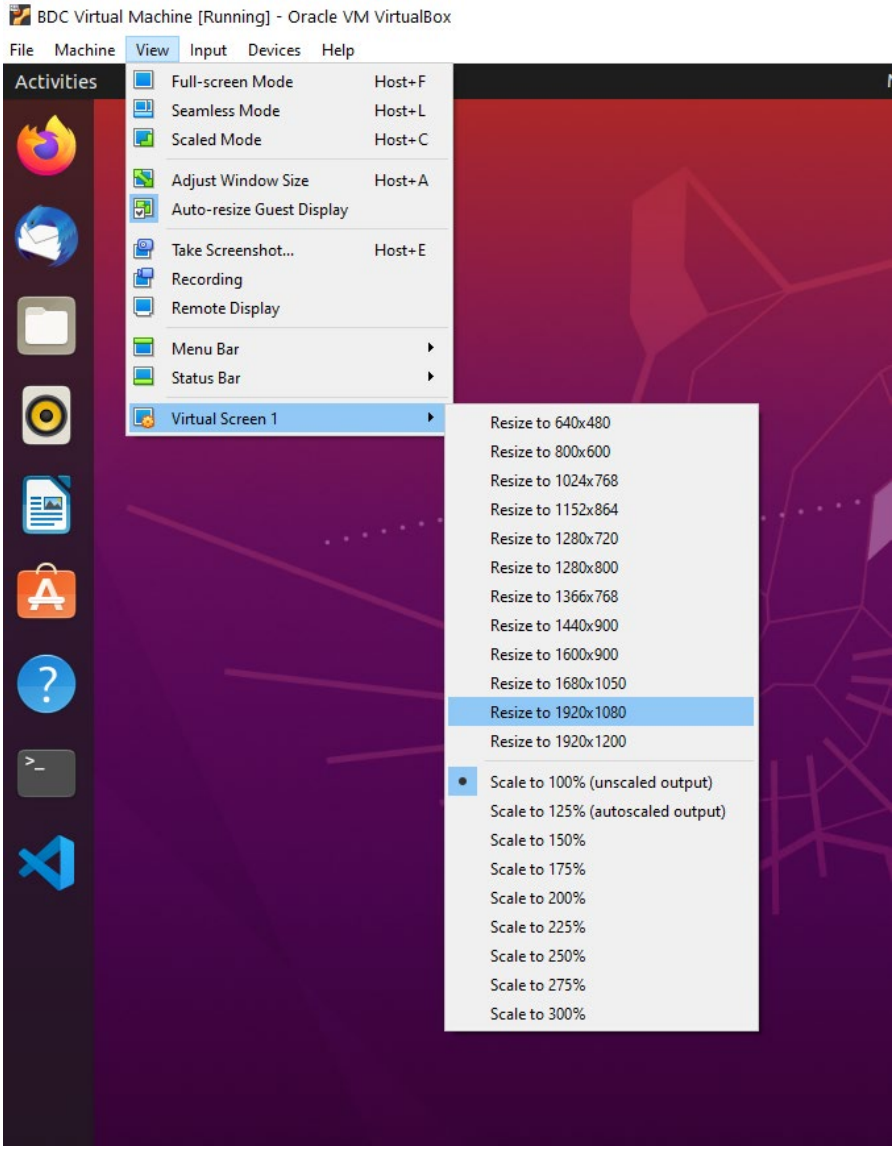
6. Next press OK and then start the VM.



7. If it asks you which operating system to boot into, please choose Ubuntu. For most people this question will not be asked but if it is asked then choose Ubuntu.
8. Login to the osboxes.org account with the password of “osboxes.org”.



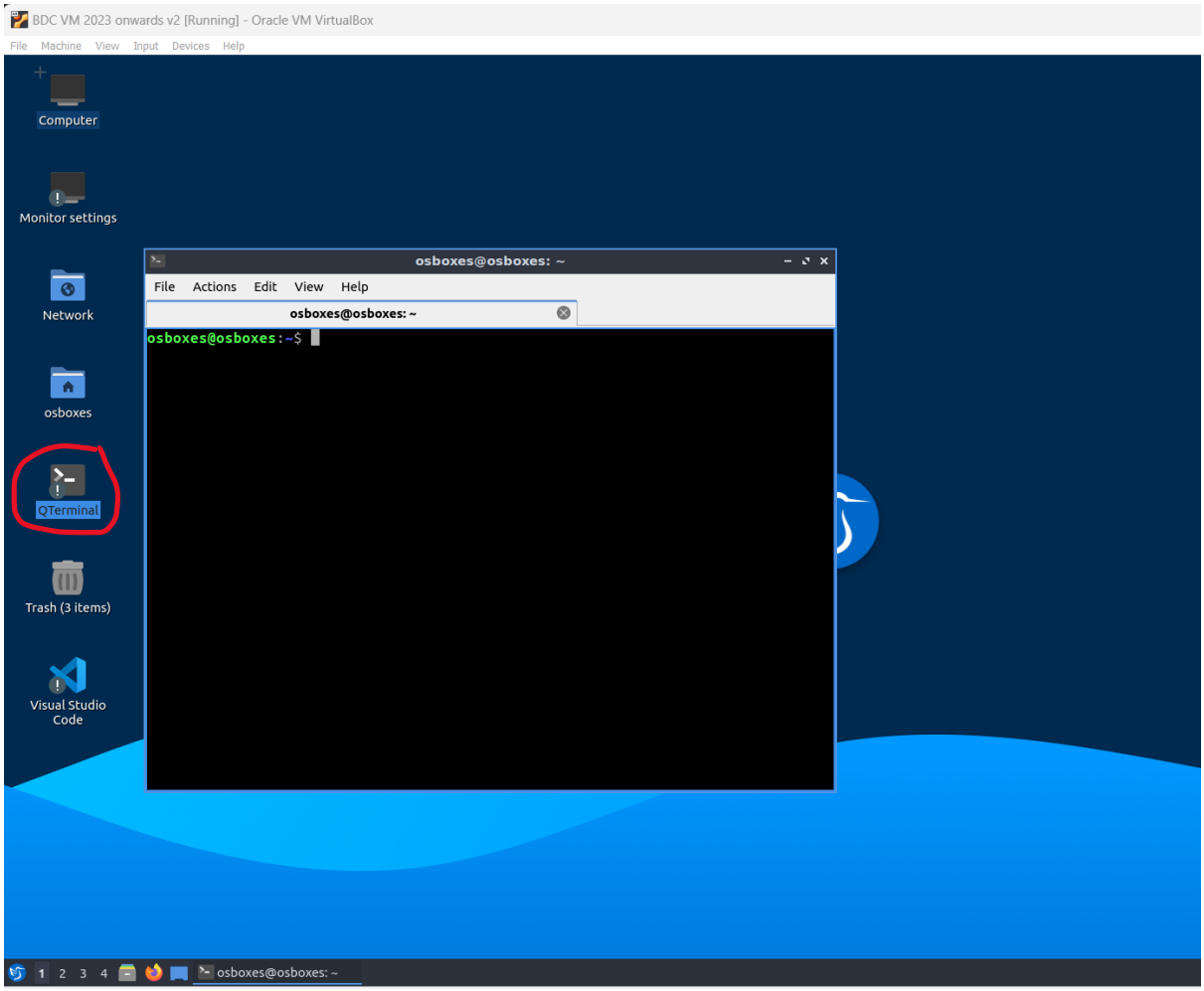
9. Sometimes when you resize the VM window the mouse controls go a bit funny, if that happens go to *view -> virtual screen 1 -> select* a resize that matches your screen (e.g. 1920 x1080). This will fix any mouse control unaligned issues.



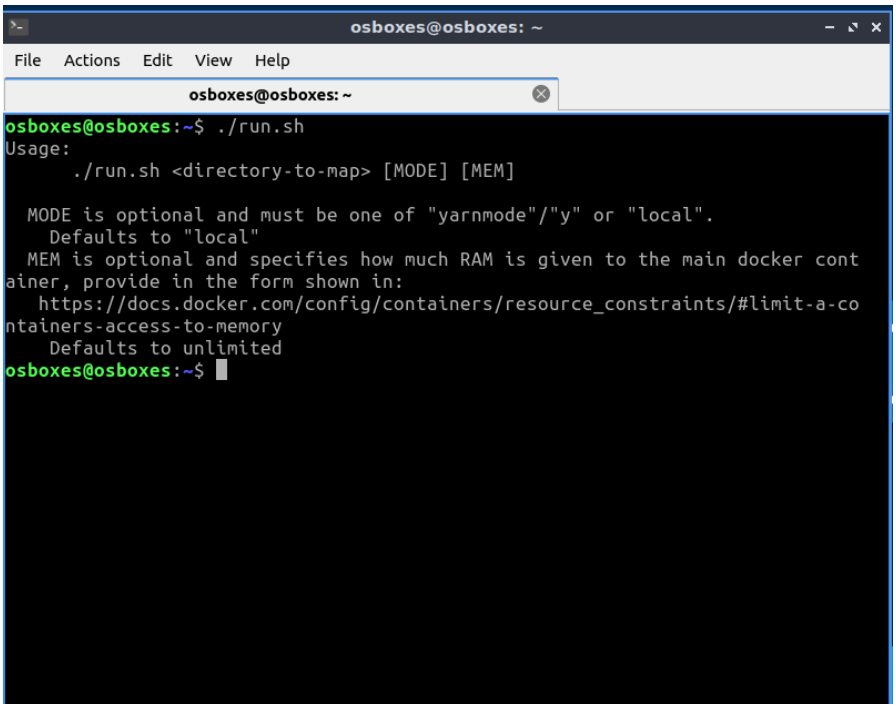
10. Sometimes the virtual machine will ask if you want to update or upgrade the operating. PLEASE DO NOT UPDATE or UPGRADE any software packages. In the past, I have updated some software and was then not able to log into the image again!
    If you really want to update, please first make a backup of the vdi file or a snapshot of the VM and then update. In case you are not able to login again you can at least revert back to your previous version.

11. Next open a QTerminal and move to the docker-hive-spark directory to fire up docker.

    To open a QTerminal, click on the terminal button shown below:



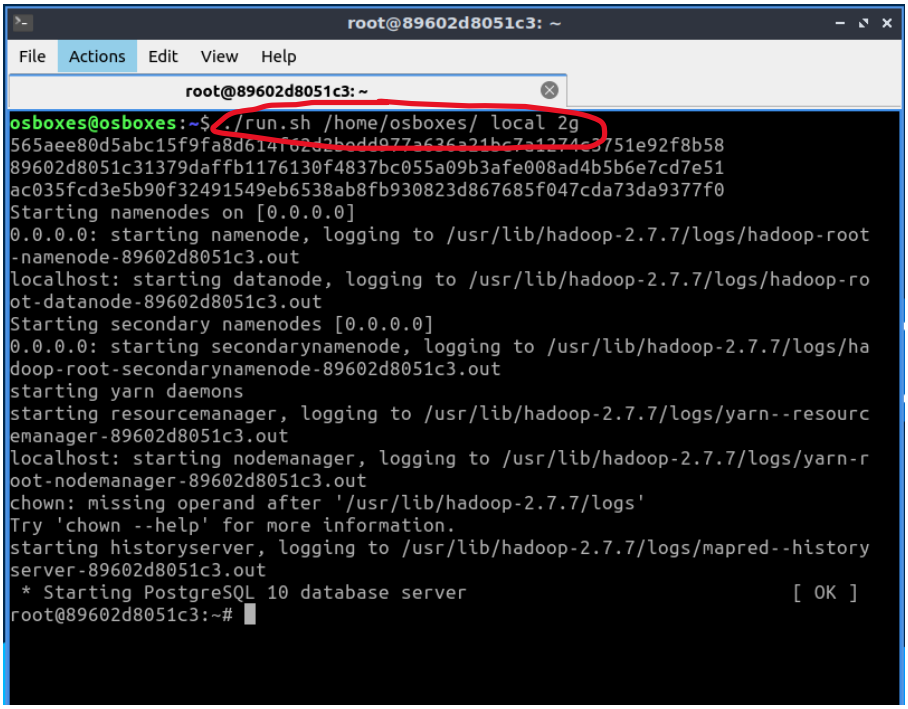
12. Then type the following to see what options you can use when running docker:
    ./run.sh



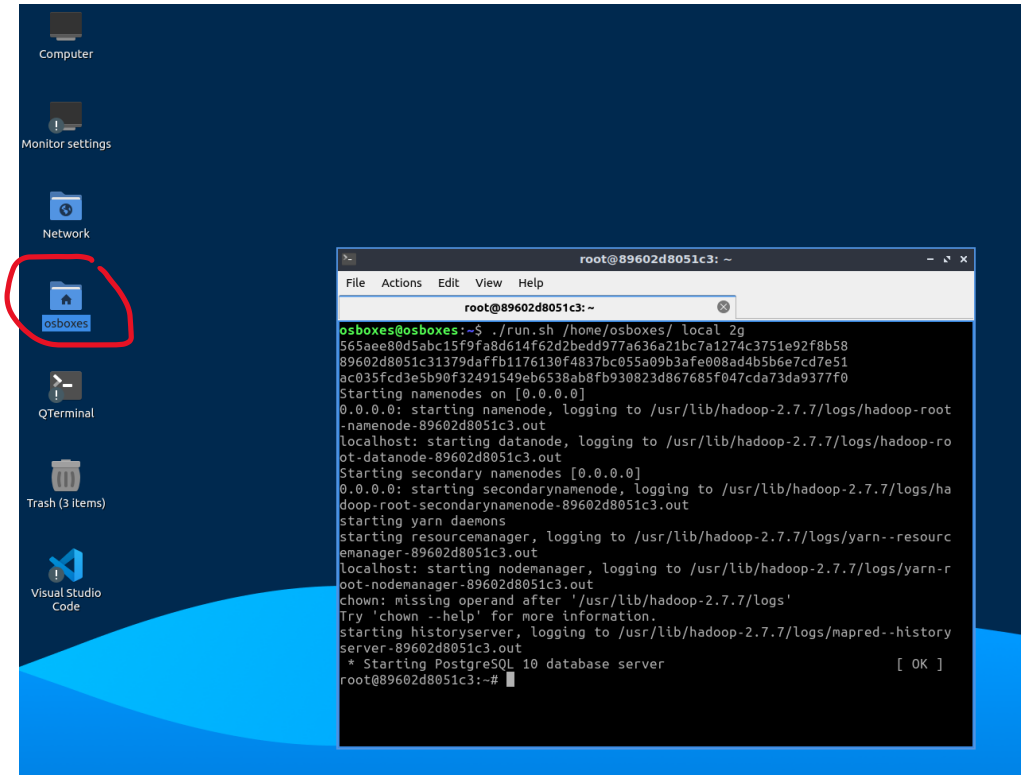
Use "./run.sh /home/osboxes/ local" if you want to run local mode. This will run a lot faster than yarnmode. I suggest you use this mode for everything, except for one of the early exercises in lab 2 where you need to look at the Map Reduce jobs where you should use the yarnmode instead.

```
root@cfae77e9b696: ~
File  Actions  Edit  View  Help
root@cfae77e9b696: ~
   https://docs.docker.com/config/containers/resource_constraints/#limit-a-co
ntainers-access-to-memory
    Defaults to unlimited
osboxes@osboxes:~$ ./run.sh /home/osboxes/ local
5b1fcc3c2921f1934d57630557eced433b54a8a9712170bcaaa95065c0d53708
cfae77e9b696dfacd706fbf14b19e9f66e150ca4c6d1f2ef4567d986e74c03e0
Error response from daemon: network with name huenetwork already exists
Starting namenodes on [0.0.0.0]
0.0.0.0: starting namenode, logging to /usr/lib/hadoop-2.7.7/logs/hadoop-root
-namenode-cfae77e9b696.out
localhost: starting datanode, logging to /usr/lib/hadoop-2.7.7/logs/hadoop-ro
ot-datanode-cfae77e9b696.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/lib/hadoop-2.7.7/logs/ha
doop-root-secondarynamenode-cfae77e9b696.out
starting yarn daemons
starting resourcemanager, logging to /usr/lib/hadoop-2.7.7/logs/yarn--resourc
emanager-cfae77e9b696.out
localhost: starting nodemanager, logging to /usr/lib/hadoop-2.7.7/logs/yarn-r
oot-nodemanager-cfae77e9b696.out
chown: missing operand after '/usr/lib/hadoop-2.7.7/logs'
Try 'chown --help' for more information.
starting historyserver, logging to /usr/lib/hadoop-2.7.7/logs/mapred--history
server-cfae77e9b696.out
 * Starting PostgreSQL 10 database server                              [ OK ]
root@cfae77e9b696:~#
```

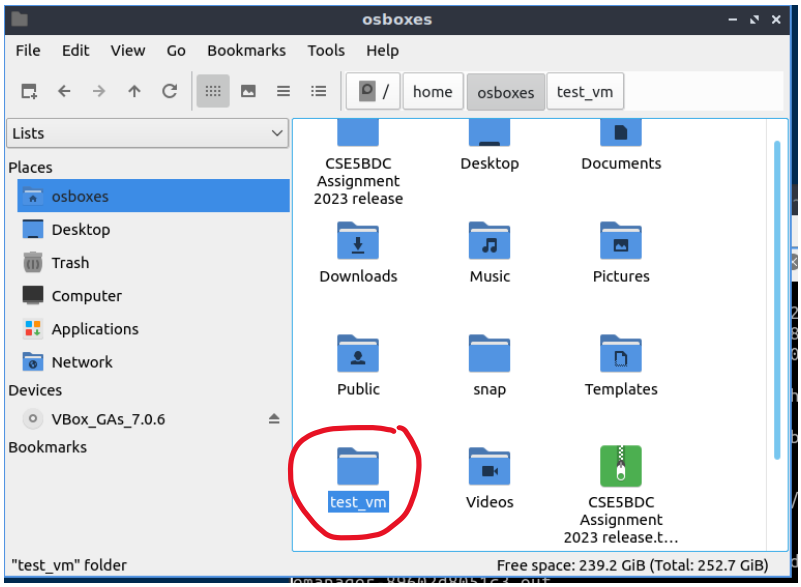
Finally, the last parameter is if you want to limit the amount of RAM you allocate to the *docker* image. I suggest you select at least 2GB. If you have allocated 4GB of RAM to the VM then I would give between 2GB to 3GB to docker. If you selected 6GB of RAM to the VM then I would give 4GB. Or you can just leave this parameter blank. In the example below I set the VM to use just 2GB of RAM. A tutor was able to do all the labs and the assignment using just 1GB of RAM assigned to docker.



13. Now we are in the docker image, we can try out a Hive script to see if everything is working as they should. Let's first open the code in the visual studio code editor to take a look at it. In First click the osboxes folder.



Look for the test_vm folder then move into the folder and then find file t1-wordcount.hql and then *right click and open with visual studios*.



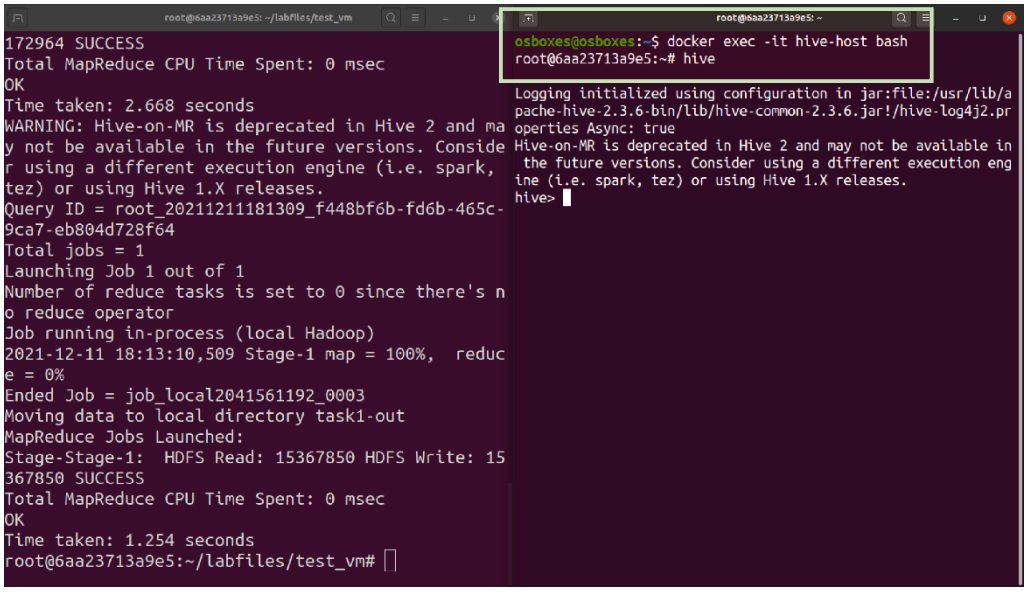
14. Go back to the terminal installing the docker image. If the install has finished you should be at a root prompt. Now change to the labfiles directory. This will take you to your home osboxes directory. Now change to test_vm directory.
    Then type the following command to run hive to test if your machine is fast enough to run the labs: hive -f t1-wordcount.hql (command also included in screenshot below)

    If everything finishes within 1 or 2 mins then it should be good enough to do the labs. If this take more than 5 mins to finish then your machine maybe too slow.
    Trouble shooting: if the directory test_vm cannot be found. Please double check that the second parameter to ./run.sh /home/osboxes/ is typed in correctly.

15. Now let's start another terminal window and use that for the hive interpreter. So, we will have two windows open. For the second window we will use the following command to connect into the docker container above.

```
docker exec -it hive-host bash
```

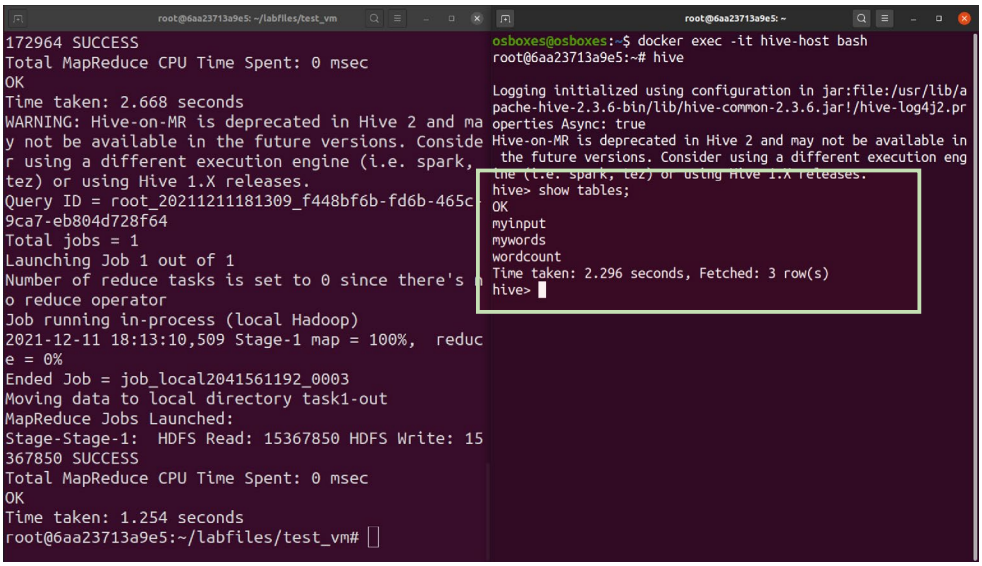
Then open the hive interpreter using the hive command. See the right window below:



16. Now on the right window you can type the following in the hive interpreter to see what tables were created.

    show tables;

    If everything works correctly you should see three tables (myinput, mywords and wordcount)



    Note the first time you use the hive interpreter it will be slow. But if you run it again it should be super fast. Try the above command again in the hive interpreter.

17. Saving via OneDrive. If drag and drop does not work for some reason. You can upload your work to your own OneDrive account for backup. Just need to log into your OneDrive account using the FireFox browser using your La Trobe student account. If you have trouble logging into OneDrive, you can upload your files to your google drive account or anywhere else on the web. Alternatively, if you have a DropBox account, you can use that too.