622 – Homework 1

Question:

Read this article: https://www.dataquest.io/blog/docker-data-science/. Write a short paragraph (~100 words) on why we should use Docker for data science.

Answer:

Introduction:

This document explains about docker usage in data science and the environment setup which is created with docker.

Background:

Little background on my environment setup. I have been using cloud virtual desktop (Which is windows) and a physical laptop (Again windows) for all my work. Also maintain a ubuntu server on AWS cloud for heavy load tasks. So I maintain software's separately in all the pc's. I always faced the trouble of maintaining different versions and errors while executing the code.

Until last week, I did not know much about docker. But this week assignment stressed me to learn about docker. I was wondering and asking this question to me "Why I did not know about docker until now? This would have saved lot of effort in my setup".

Usefulness of Docker:

Docker eliminated lot of setup work on any individual pc and it is far more efficient than virtual server.

1. All or most of the images in docker hub is pre-build with popular software's and packages.
2. Downloading a docker images is very quick.
3. We can also customize the image and build more custom containers on top of it.
4. Can connect docker container to host hard drives. So the files are stored locally. If the container is deleted or corrupted, we can still retain the data.
5. Can run multiple containers in different ports without affecting each other.
6. Docker also allows to perform networking and do client-server architecture across different containers.

Current Setup:

1. **Docker on Droplet cloud**
2. **Anaconda on AWS cloud**

Docker on Droplet cloud:

As part of this assignment, I have taken an instance in digital ocean. Below are details about it.

Link: https://digitalocean.com

Droplet instance: 4 GB Memory / 60 GB Disk / NYC3

OS: Ubuntu 16.04.3 x64

I followed below steps in this assignment.

Step 1:

Create an instance and apply all the updates to that PC.

Step 2:

Downloaded some custom images and tried to run that image.

```
*** System restart required ***
Last login: Tue Sep  5 23:57:10 2017 from 209.51.162.173
root@digitaltest:~# docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED
STATUS              PORTS               NAMES
root@digitaltest:~# docker images
REPOSITORY          TAG                 IMAGE ID            CREATED
SIZE
mysite              latest              aae16f9a08ac        17 hours ago
281MB
mywebsite           latest              aae16f9a08ac        17 hours ago
281MB
getmyip             latest              0b4b3c171be8        42 hours ago
166MB
<none>              <none>              f62003d3686c        42 hours ago
166MB
myubuntuimage       latest              f1517902048e        43 hours ago
166MB
ubuntu              latest              ccc7a11d65b1        3 weeks ago
120MB
redis               latest              d4f259423416        6 weeks ago
106MB
jpetazzo/web        latest              c7635dd24e9a        2 years ago
276MB
```

Step 3: Deleted all the previously build containers. And download the image *dataquestio/python3-starter*

```
REPOSITORY          TAG                 IMAGE ID            CREATED
SIZE
mysite              latest              aae16f9a08ac        17 hours ago
281MB
ubuntu              latest              ccc7a11d65b1        3 weeks ago
120MB
root@digitaltest:~# docker rmi -f aae16f9a08ac
Untagged: mysite:latest
Deleted: sha256:aae16f9a08ac84aa81fe4d9c41c44b0bf03e224d5fa3d5fa8a4d28a58d41510
Deleted: sha256:571cbd878d95f4b7bf17a750faf4339395a46edcd07033d3e62dfb569c862181
Deleted: sha256:78b83c0b3a7832f37cd43b0f2ada760a2347b4297d6ec9f3ccbe09434e2ff8ec
Deleted: sha256:54bdb6252c241dff7114f8d33475bf66b56ed78a9d254369dccdf1c4ab963b49
Deleted: sha256:9e714e6c59746dfa55cb60834d37c21494b24d668dc385294abd1dcf5c1acb28
Deleted: sha256:08b59d78d7fc217c1563bfbbbd17d198acb7583c2f85b4a0832c17e3750757b4
Deleted: sha256:11229829d07a2eae465702f58dbcb4824ffc8f8b425559fbadaa8f577e38108e
Deleted: sha256:2c0a538baea3150cc2442d5843dd5055c6ad7895ee66d44016c4f0de93bbc325
Deleted: sha256:b07a2fc84dd05f4b06003fb83736fd03bb7c69498a61a1421a0028386a9be886
Deleted: sha256:ae1e80cb775db650f899adf902babfe90be822ecc2e6b56fb259ef986c4d7bc7
Deleted: sha256:a05c1e60a633d3e7067d85fc5c628d355d76c423ee13016e7aca2255f14daf94
Deleted: sha256:aac77021895445c2d963c11a0b78cb11364ef4539d4552ba5c21ac7f70d5d422
Deleted: sha256:cfca24d591d11f9751dea35f5a20e046cf85b10bbfdd0a11521e1cba8b9e258e
Deleted: sha256:a7313a88d3b3b6ae4d3efb3bc19ced5479f750abac76e34652ea61844c4f91a1
Deleted: sha256:cf1f9415008c5abc66b04813dcebd5abba5c9913016b4a2728c8ab04869f351e
root@digitaltest:~# docker images
REPOSITORY          TAG                 IMAGE ID            CREATED
SIZE
ubuntu              latest              ccc7a11d65b1        3 weeks ago
120MB
```

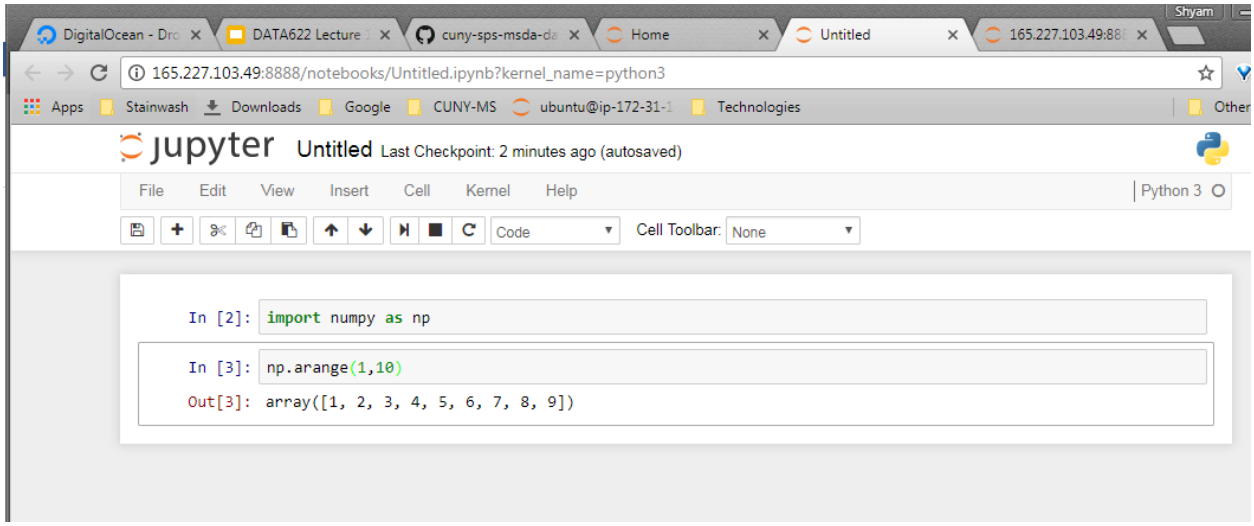
Step 4: Create a container and run it on port 8888. Some additional options performed on container.

1. Changed container port from default to port 8888.
2. Mounted the host volume to the container volume.

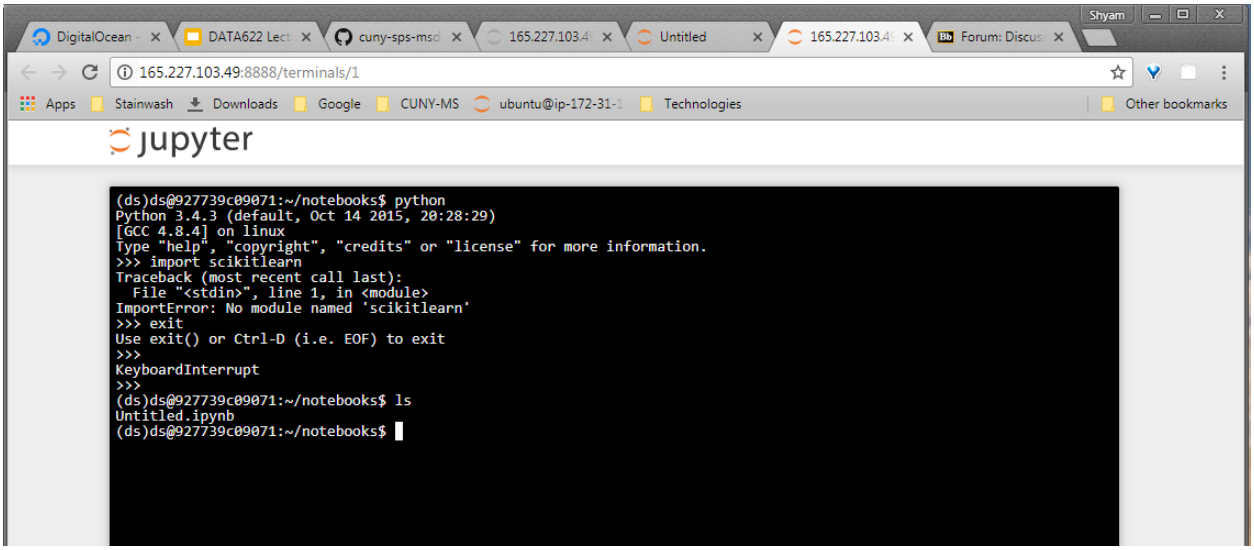
```
root@digitaltest:~# docker ps
CONTAINER ID        IMAGE               COMMAND                  CREATED
     STATUS              PORTS                    NAMES
927739c09071        3cccab0c85e2        "/home/ds/run_ipyt..."   5 seconds ago
     Up 4 seconds        0.0.0.0:8888->8888/tcp   ecstatic_ritchie
root@digitaltest:~#
```

Step 5: Below is the screenshot of the Jupyter notebooks.

URL: http://165.227.103.49:8888/tree



Terminal window in Jupyter:



Challenges faced:

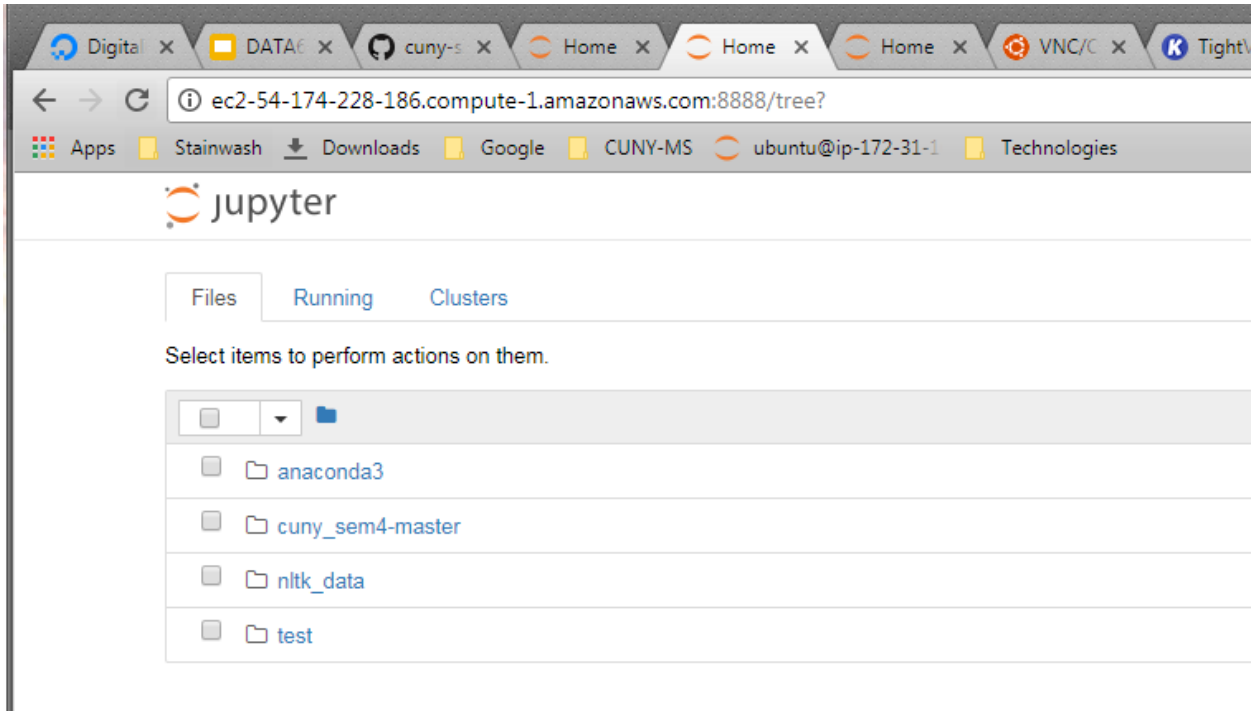
Below are some challenges faced on the docker setup.

1. Issues while mapping the volume and ports to the container.
2. Some errors while setting up on Windows pc.

Anaconda on AWS cloud:

In slack we discussed about anaconda and docker. I have installed anaconda complete instance in AWS instance.

Note: This setup is not using docker.



Challenges faced:

Below are some challenges faced on the anaconda setup.

1. It took lot of time to install anaconda on Ubuntu server.
2. Lot of time spent to setup the jupyter notebook on anaconda.
3. Password reset and initial setup was little tough.