Assignment: 1

TEAM7: SNIGDHA JOSHI & VIPRA SHAH

This assignment is divided into two parts:

1) Data Ingestion
2) Data Wrangling

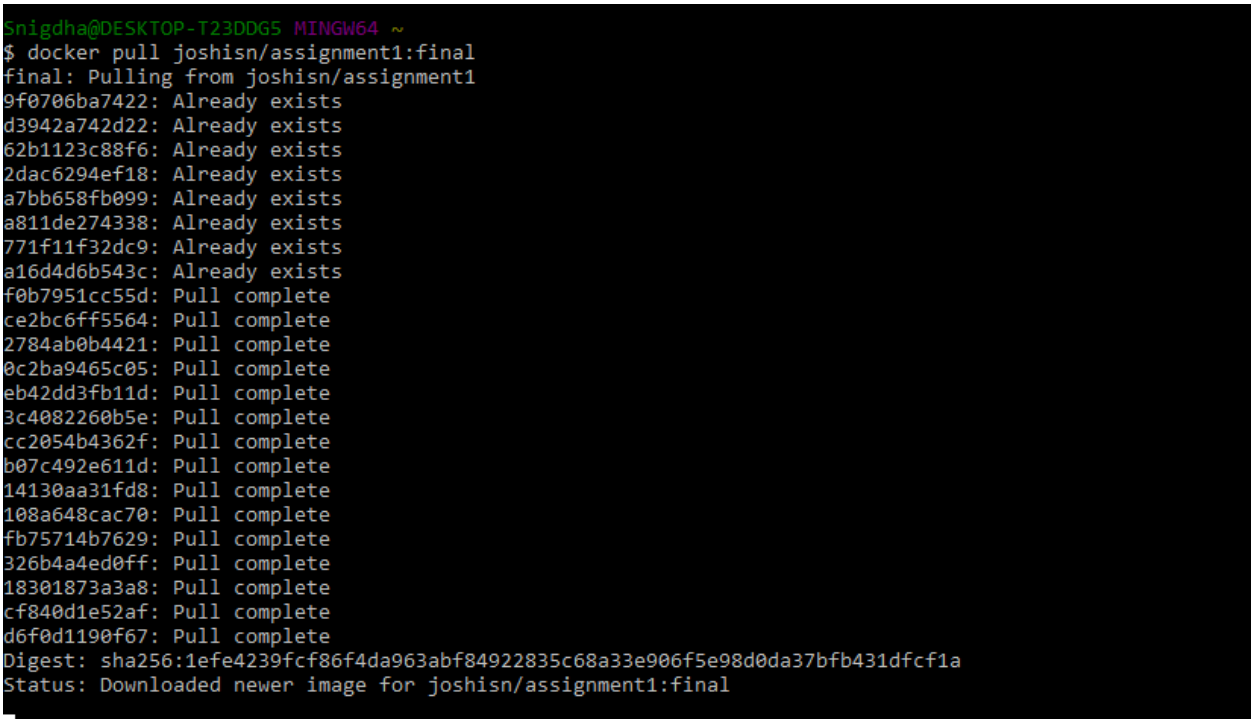
For both docker images perform following steps:

1) Pull the image from the dockerhub
2) Copy the config.json and configwrangle.json file to your local machine
3) Add your AWS access and secrete keys, create a bucket called "team7pa_assigment1" then run the following codes.
4) Commit changes
5) Execute the code

For detailed description of above steps please refer below:

Data Ingestion

1. Pull image from docker hub



2. Create the container

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Raw_Data (master)
$ docker create --name="rawdata13" joshisn/assignment1:final
1a0eeb55800aae359b25741e286f4c2f39f75727289fe941c86dff88dfc0296f
```

3. Copy config file to your local machine and edit as mentioned below:

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Raw_Data (master)
$ docker cp config.json rawdata13:/Assignment1/
```

4. Start the container

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Raw_Data (master)
$ docker start -i rawdata13
/Assignment1
1
2017-06-25
```

5. Commit the container to save the changes

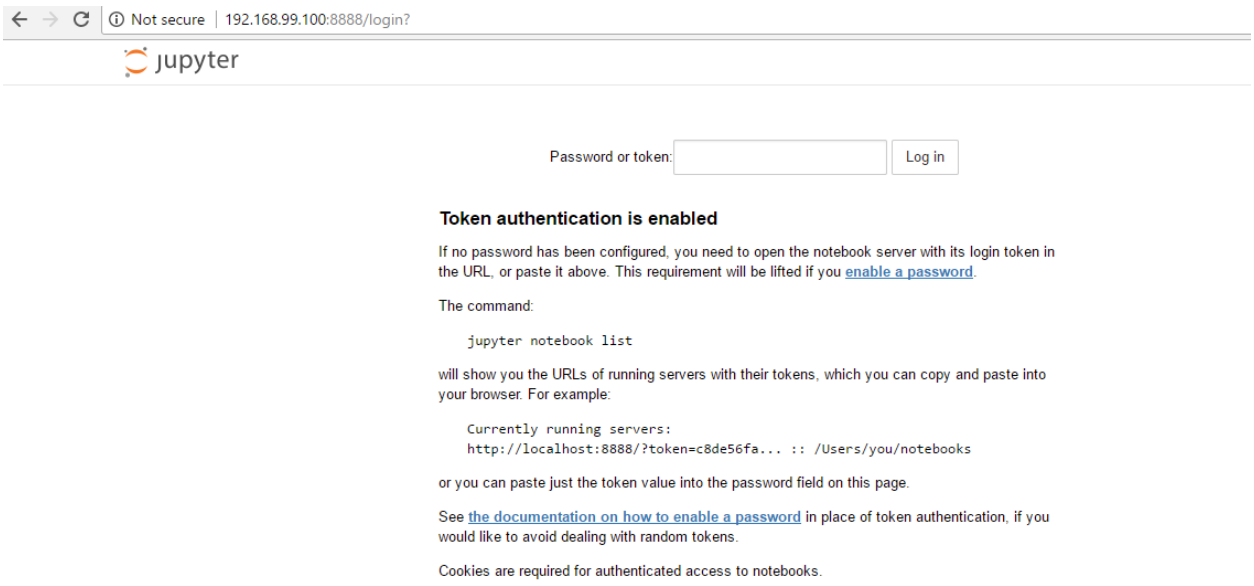
```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Raw_Data (master)
$ docker commit rawdata13 joshisn/assignment1newimage
sha256:07493232db067285be1d8614ba6c626d1481d0649e05070ebe26baea9abc450e
```

6. Run jupyter notebook

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Raw_Data (master)
$ docker run -it -d --name "rawdata_2" -p 8888:8888 joshisn/assignment1newimage /bin/bash -c 'jupyter notebook --no-bro
wser --allow-root --ip=* --NotebookApp.password="$PASSWD" "$@"'
272d75e6db86843baa4373c1a8874b33585d3e78462240531d0b24b0653b8561
```

7. Connect jupyter notebook

   http://<yourdocker ip address>:8888



Data Wrangling

1. Run following commands to pull the image and create container and run wrangle.py file

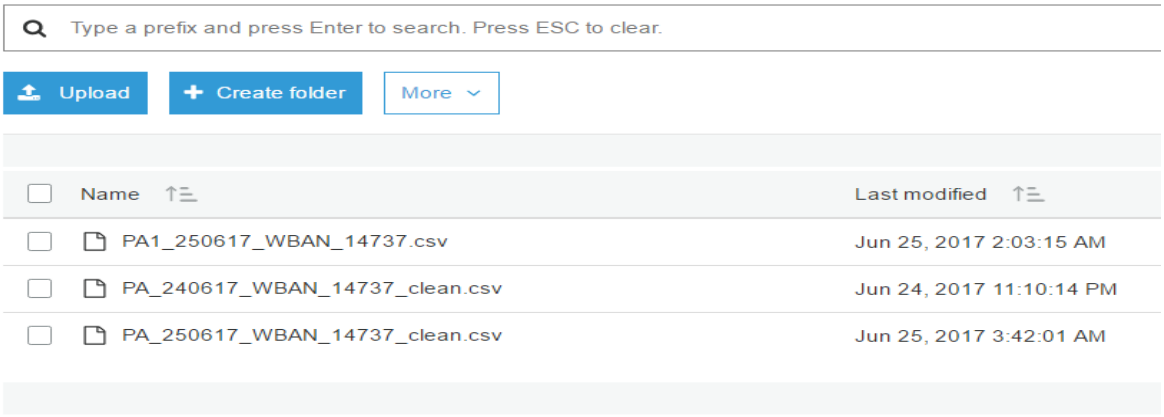
   Note: before running Create command, copy and edit the configwrangle.json file as mentioned in Data Ingestion part.

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Clean_Data (master)
$ docker pull joshisn/assignment1:final
final: Pulling from joshisn/assignment1
Digest: sha256:71464840f073694af87742782c47e0c3087332252a9c8b9df9bff91d9b7b48ea
Status: Image is up to date for joshisn/assignment1:final
```

```
Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Clean_Data (master)
$ docker create --name="cleandatacontainer_new" joshisn/assignment1:final
2ef4743b59e70f425a1e758f2e51a55c5c30c6df6883455671dd809a6aec5320

Snigdha@DESKTOP-T23DDG5 MINGW64 ~/Documents/ADS/Assignment1/Clean_Data (master)
$ docker start -i cleandatacontainer_new
```

After performing above command clean.csv file will be uploaded on S3 bucket.



To open jupyter notebook follow command mentioned in Data Ingestion part.