Exercise 0

Install docker and run the container

Ubuntu/Debian

You can install docker using apt: `apt-get install docker.io`.

Other Operating Systems

Go to `https://docs.docker.com/engine/installation/` and follow the instructions for your system.

Run the container

Run the container:

```
$ sudo docker run -d --name dwws --restart always \
-v /home/user/share/:/home/student/share/ acrrd/dw-workstation
```

where `/home/user/share` is the absolute path to the directory you want to use to share data with the container. The command download the image (∼1.3Gb) if not already present on the system. You have to run this command only once. You can use: `sudo docker ps -a` to see the active containers. You can use: `sudo docker start dwws` to start the container and `sudo docker stop dwws` to stop it.
The container embeds a working version of HIVE installed and it runs an ssh server. We just need to know its ip address to connect to it, the command `sudo docker inspect dwws` output information about the container, search for `"IPAddress"` to find the ip.

```
$ sudo docker inspect dwws | grep \"IPAddress\"
  "IPAddress": "172.17.0.2",
          "IPAddress": "172.17.0.2",
```

We can login on the container using ssh with user `student` and password `foobar`:
`ssh student@172.17.0.2`.
If the connection does not work try to start the container.
On your host machine, move the two files `earthquakes.csv` and `GlobalSuperstoreOrders2016.csv` in the directory you want to share with the container.
On the container, runing `ls share` you should be able to see the two files above.
If you delete the container you loose all the data except the one in `/home/user/share`.
In case you need it, root password is `toor`.
To launch the HIVE shell we just need to run the command `hive`.

Import local data into HIVE

In this simple example we will import the content of `earthquakes.csv` into HIVE. First of all, we should create a table with heading that matches the metadata of our csv file.

```
hive> CREATE TABLE Earthquakes (id string, latitude double,
      longitude double, depth double, mag double)
      row format delimited fields terminated by ',';
```

Next, we can just use the following code to load the data into the table we just created.

```
hive> LOAD DATA LOCAL INPATH '/home/student/share/earthquakes.csv'
      OVERWRITE INTO TABLE Earthquakes;
```

Now you can query your table using the HQL language. For instance, run the following command to check if your import is correct.

```
  hive> SELECT * FROM Earthquakes;
```

If everything worked properly, import the content of `GlobalSuperstoreOrders2016.csv` into a new table and check its content.
Now try to solve the following queries:

1. Select the orders with High Order Priority
2. Select the orders that were placed from november 2014 to february 2015
3. Select the five most profitable orders from the previous query