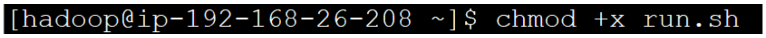
COSC2637 Big Data Processing
Lab 5 - Tutorial/Lab class (week 6)

Task - Python MapReduce program for k-means clustering

Given a set of data points in a two-dimensional space and k locations in the space as the initial cluster centroids, the task is to develop a Python MapReduce program for k-means clustering. The source code from an online source has been provided in course week 6 module on Canvas.

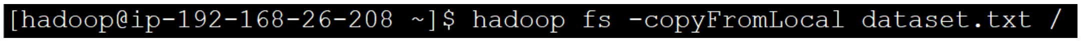
1. Upload following files to AWS EMR cluster mast node. These files include

mapper.py *# mapper*
reducer.py *# reducer*
reader.py *# check whether termination condition is satisfied*
run.sh *# control iterations and termination condition check by reader.py*
dataset.txt *# a set of data points in a 2-dimensional space*
initial_centroids_backup.txt *# contains the initial 3 cluster centroids*
hadoop-streaming-3.1.4.jar



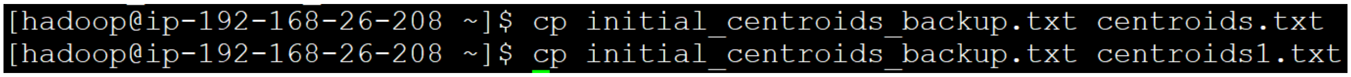
```
[hadoop@ip-192-168-26-208 ~]$ chmod +x run.sh
```

make sure run.sh is executable "chmod +x run.sh"



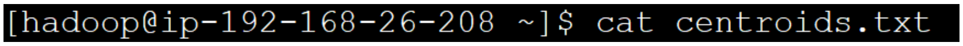
```
[hadoop@ip-192-168-26-208 ~]$ hadoop fs -copyFromLocal dataset.txt /
```

copy dataset.txt to folder "/" on HDFS by "hadoop fs -copyFromLocal dataset.txt /"



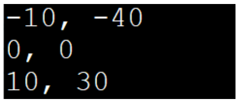
```
[hadoop@ip-192-168-26-208 ~]$ cp initial_centroids_backup.txt centroids.txt
[hadoop@ip-192-168-26-208 ~]$ cp initial_centroids_backup.txt centroids1.txt
```

centroids.txt *# initialized by initial_centroids_backup.txt, and contains results after iterations.*
centroids1.txt *# the centroids after one iteration, used for checking termination condition in run.sh*



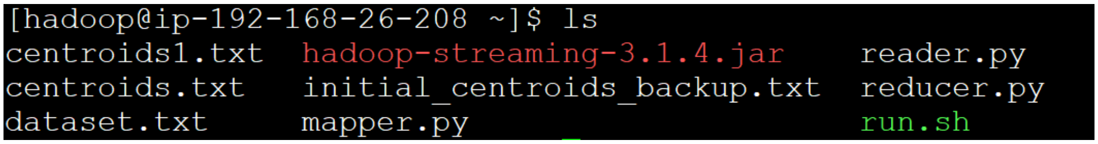
```
[hadoop@ip-192-168-26-208 ~]$ cat centroids.txt
```

The content is



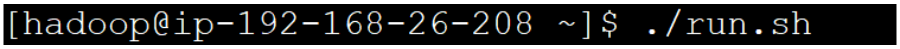
```
-10, -40
0, 0
10, 30
```

Now you should have all files ready for execution



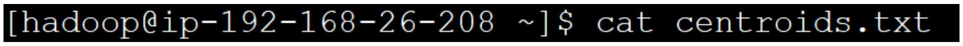
```
[hadoop@ip-192-168-26-208 ~]$ ls
centroids1.txt  hadoop-streaming-3.1.4.jar    reader.py
centroids.txt   initial_centroids_backup.txt  reducer.py
dataset.txt     mapper.py                     run.sh
```

Execute run.sh



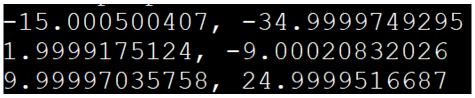
```
[hadoop@ip-192-168-26-208 ~]$ ./run.sh
```

If running successfully, check the content of centorids.txt



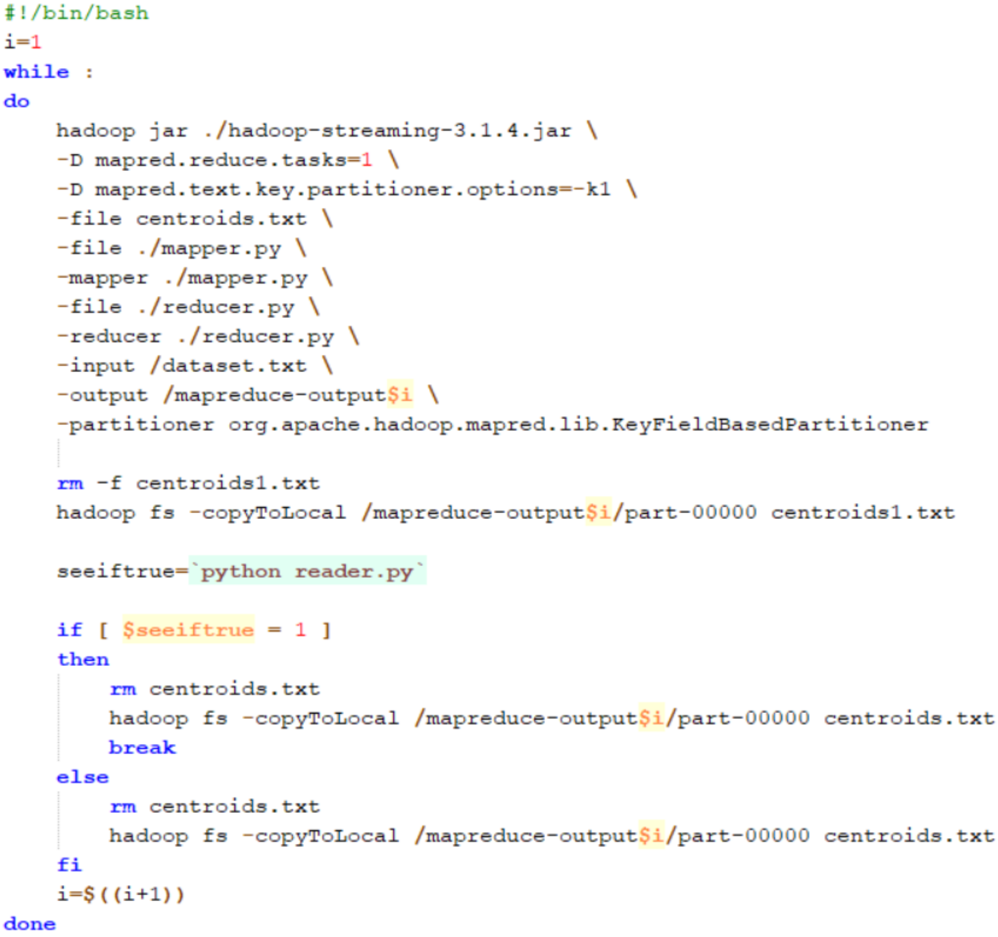
```
[hadoop@ip-192-168-26-208 ~]$ cat centroids.txt
```

You will see something like below



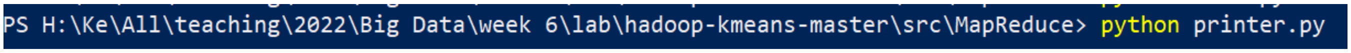
```
-15.000500407, -34.9999749295
1.9999175124, -9.00020832026
9.99997035758, 24.9999516687
```

Read through run.sh (below) to understand the flow the k-means algorithm. The detailed explanation of the code is in report.pdf



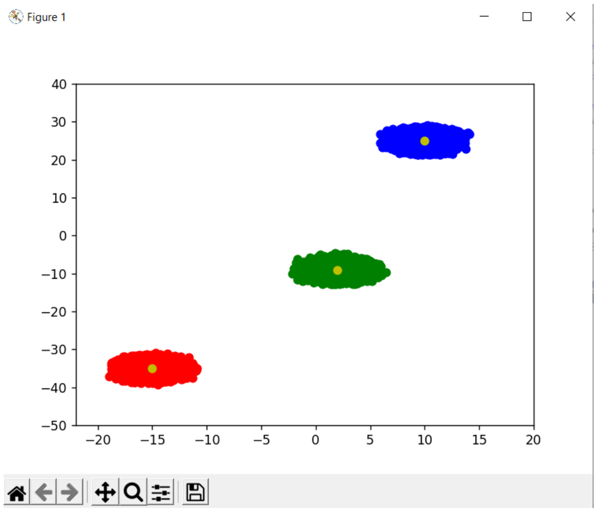
```
#!/bin/bash
i=1
while :
do
    hadoop jar ./hadoop-streaming-3.1.4.jar \
    -D mapred.reduce.tasks=1 \
    -D mapred.text.key.partitioner.options=-k1 \
    -file centroids.txt \
    -file ./mapper.py \
    -mapper ./mapper.py \
    -file ./reducer.py \
    -reducer ./reducer.py \
    -input /dataset.txt \
    -output /mapreduce-output$i \
    -partitioner org.apache.hadoop.mapred.lib.KeyFieldBasedPartitioner

    rm -f centroids1.txt
    hadoop fs -copyToLocal /mapreduce-output$i/part-00000 centroids1.txt

    seeiftrue=`python reader.py`

    if [ $seeiftrue = 1 ]
    then
        rm centroids.txt
        hadoop fs -copyToLocal /mapreduce-output$i/part-00000 centroids.txt
        break
    else
        rm centroids.txt
        hadoop fs -copyToLocal /mapreduce-output$i/part-00000 centroids.txt
    fi
    i=$((i+1))
done
```

You can visualize the data points dataset.txt and the cluster centroids results.txt with printer.py (you need make a copy of "results.txt" and name the copy as "centroids.txt"). Put centroids.txt and dataset.txt on local machine (not AWS EMR cluster and jumphost) in the same folder as printer.py. On my windows, I use "powershell". Run printer.py as:



```
PS H:\Ke\All\teaching\2022\Big Data\week 6\lab\hadoop-kmeans-master\src\MapReduce> python printer.py
```

The output will be:



Exercise: You can create a set of data points in a 2-dimensional space and perform k-means clustering. You may find the location of initial clustering centroids are essential. Why?

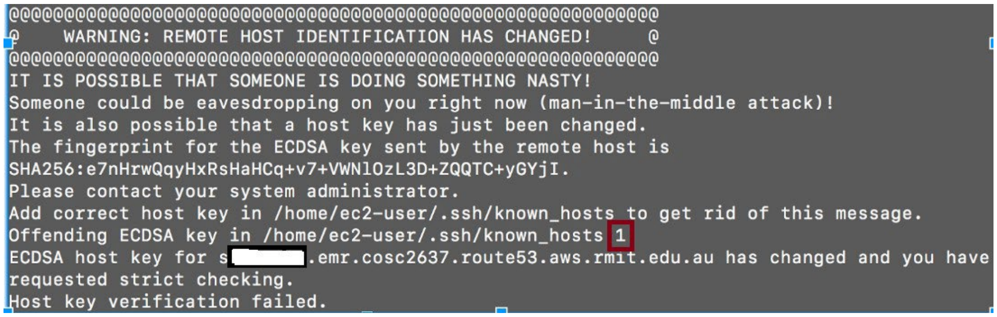
Don't forget to terminate your cluster at end of the lab class!

Also, if trying out any these exercises outside of the lab class, make sure you terminate your cluster and DO NOT leave clusters running idle for days.

******Useful Information******

- **A common issue of cluster login**

You may see the following error when ssh from your jumphost to the hadoop master node,

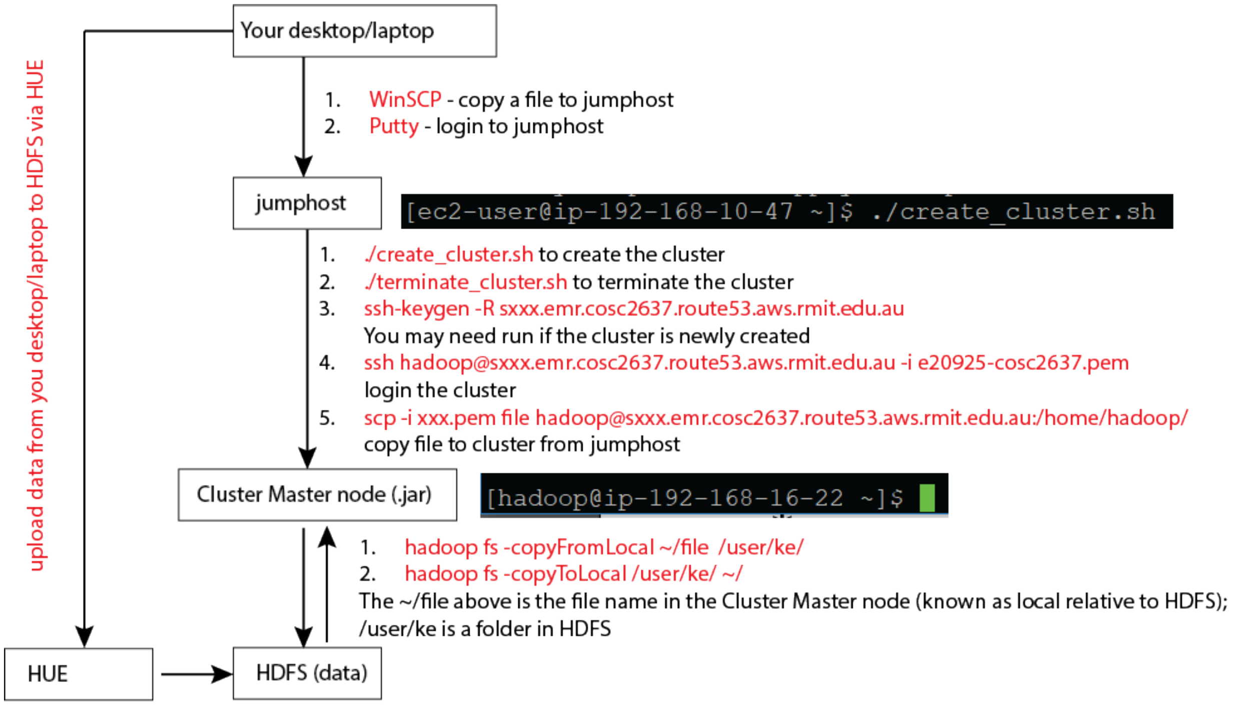


```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the ECDSA key sent by the remote host is
SHA256:e7nHrwQqyHxRsHaHCq+v7+VWNlOzL3D+ZQQTC+yGYjI.
Please contact your system administrator.
Add correct host key in /home/ec2-user/.ssh/known_hosts to get rid of this message.
Offending ECDSA key in /home/ec2-user/.ssh/known_hosts 1
ECDSA host key for s[redacted].emr.cosc2637.route53.aws.rmit.edu.au has changed and you have
requested strict checking.
Host key verification failed.
```

Every time a new cluster is created it is a new set of hosts so the ssh host key changes, and ssh gives you a warning that it has been changed. To fix the problem, please run the following:

```
$ssh-keygen -R sxxxxxxx.emr.cosc2637.route53.aws.rmit.edu.au
```

- ## AWS EMR - Cheat sheet



upload data from you desktop/laptop to HDFS via HUE

Your desktop/laptop

1. WinSCP - copy a file to jumphost
2. Putty - login to jumphost

jumphost

```
[ec2-user@ip-192-168-10-47 ~]$ ./create_cluster.sh
```

1. ./create_cluster.sh to create the cluster
2. ./terminate_cluster.sh to terminate the cluster
3. ssh-keygen -R sxxx.emr.cosc2637.route53.aws.rmit.edu.au
   You may need run if the cluster is newly created
4. ssh hadoop@sxxx.emr.cosc2637.route53.aws.rmit.edu.au -i e20925-cosc2637.pem
   login the cluster
5. scp -i xxx.pem file hadoop@sxxx.emr.cosc2637.route53.aws.rmit.edu.au:/home/hadoop/
   copy file to cluster from jumphost

Cluster Master node (.jar)

```
[hadoop@ip-192-168-16-22 ~]$
```

1. hadoop fs -copyFromLocal ~/file /user/ke/
2. hadoop fs -copyToLocal /user/ke/ ~/

The ~/file above is the file name in the Cluster Master node (known as local relative to HDFS); /user/ke is a folder in HDFS

HUE

HDFS (data)