Project2B: Harp Kmeans

Ethan Li, Rohit Patil

Indiana University Bloomington

I. Descriptions

I.1 Main Steps

The main steps are as follows:

```
Begin
generate data and initial centroids and write them to HDFS
repeat
    configure a job
    run the job
until reach maxIteration
End
```

In each mapper task:

```
Begin
setup, get initial values.
load data and centroids
do local computation and assign each data point to closet centroid
for each centroid, do local aggregation and then call allreduce
finalize computation and write the results back to HDFS.
End
```

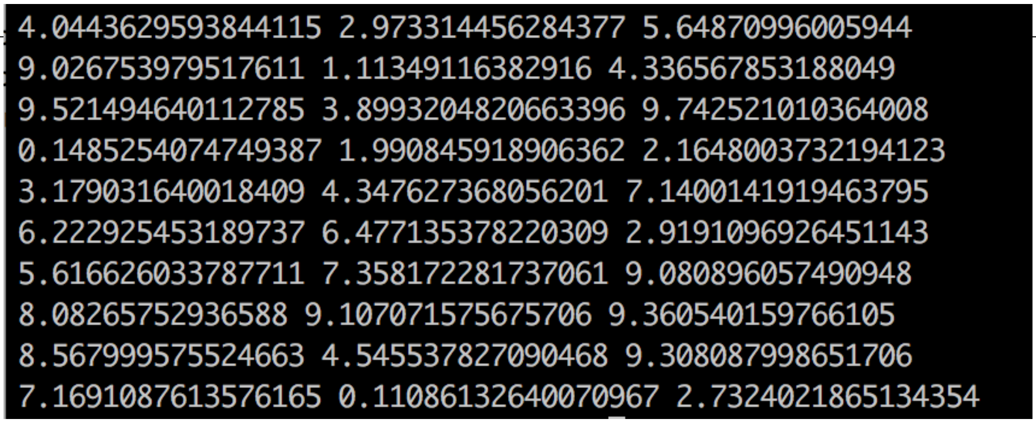
I.2 data flow

The data is being generated in memory and write back to local filesystem, and then copied to hdfs. Every iteration, the data will be loaded into HDFS.
The centroids is being generated firstly and write back to HDFS. In every iteration, they will be loaded, calculated, updated and then finally write back to HDFS. Note that in the update part, the centroids will be transferred among all mappers.

II. Test

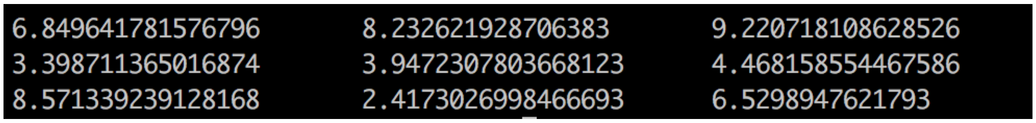
To save the space, we use a small test here to get the data and final output centroids. We use the following command to run the test and got the input data at Figure 1 and output centroids at Figure 2.

```
hadoop jar harp3-app-hadoop-2.6.0.jar \
edu.iu.km.KmeansMapCollective 10 3 3 2 5 /kmeans /tmp/kmeans
```



```
4.0443629593844115 2.973314456284377 5.64870996005944
9.026753979517611 1.11349116382916 4.336567853188049
9.521494640112785 3.8993204820663396 9.742521010364008
0.1485254074749387 1.990845918906362 2.1648003732194123
3.179031640018409 4.347627368056201 7.1400141919463795
6.222925453189737 6.477135378220309 2.9191096926451143
5.616626033787711 7.358172281737061 9.080896057490948
8.08265752936588 9.107071575675706 9.360540159766105
8.567999575524663 4.545537827090468 9.308087998651706
7.1691087613576165 0.11086132640070967 2.7324021865134354
```

Figure 1: data



```
6.849641781576796        8.232621928706383        9.220718108628526
3.398711365016874        3.9472307803668123       4.468158554467586
8.571339239128168        2.4173026998466693       6.5298947621793
```

Figure 2: centroids