Name : **Kartavya Mandani**
Roll No.: **20BCE120**
Course: **2CS702 Big Data Analytics**
Practical No.: **7 (K-means clustering and MapReduce)**

Code

1.Mapper.py

```
import sys
import numpy as np

centroids = [[2.0, 10.0], [5.0, 8.0], [1.5, 2.0]]

for line in sys.stdin:
    line = line.strip()
    data_point = list(map(float, line.split(',')))
    data_point = np.array(data_point)
    distances = [np.linalg.norm(data_point - c) for c in
centroids]
    closest_centroid_index = np.argmin(distances)
    print(f"{closest_centroid_index}\t{','.join(map(str,
data_point))}")
```

2.Reducer.py

```
#! /usr/bin/python3
import sys
import numpy as np
def read_mapper_output(input_data):
    for line in input_data:
        yield line.strip()
def reducer():
    current_centroid = None
    sum_points = np.zeros(2)
    count = 0
    for line in read_mapper_output(sys.stdin):
```

```
        if line:
            centroid_index, point = line.split('\t', 1)
            point = np.array(list(map(float, point.split(','))))
            if current_centroid is None:
                current_centroid = int(centroid_index)
            if current_centroid != int(centroid_index):
                print(f"{current_centroid}\t{','.join(map(str,
sum_points/count))}")
                current_centroid = int(centroid_index)
                sum_points = np.zeros(2)
                count = 0
            sum_points += point
            count += 1
        else:
            break
    if current_centroid is not None:
        print(f"{current_centroid}\t{','.join(map(str,
sum_points/count))}")
reducer()
```

3. Input.txt <x,y>

1.0,2.0

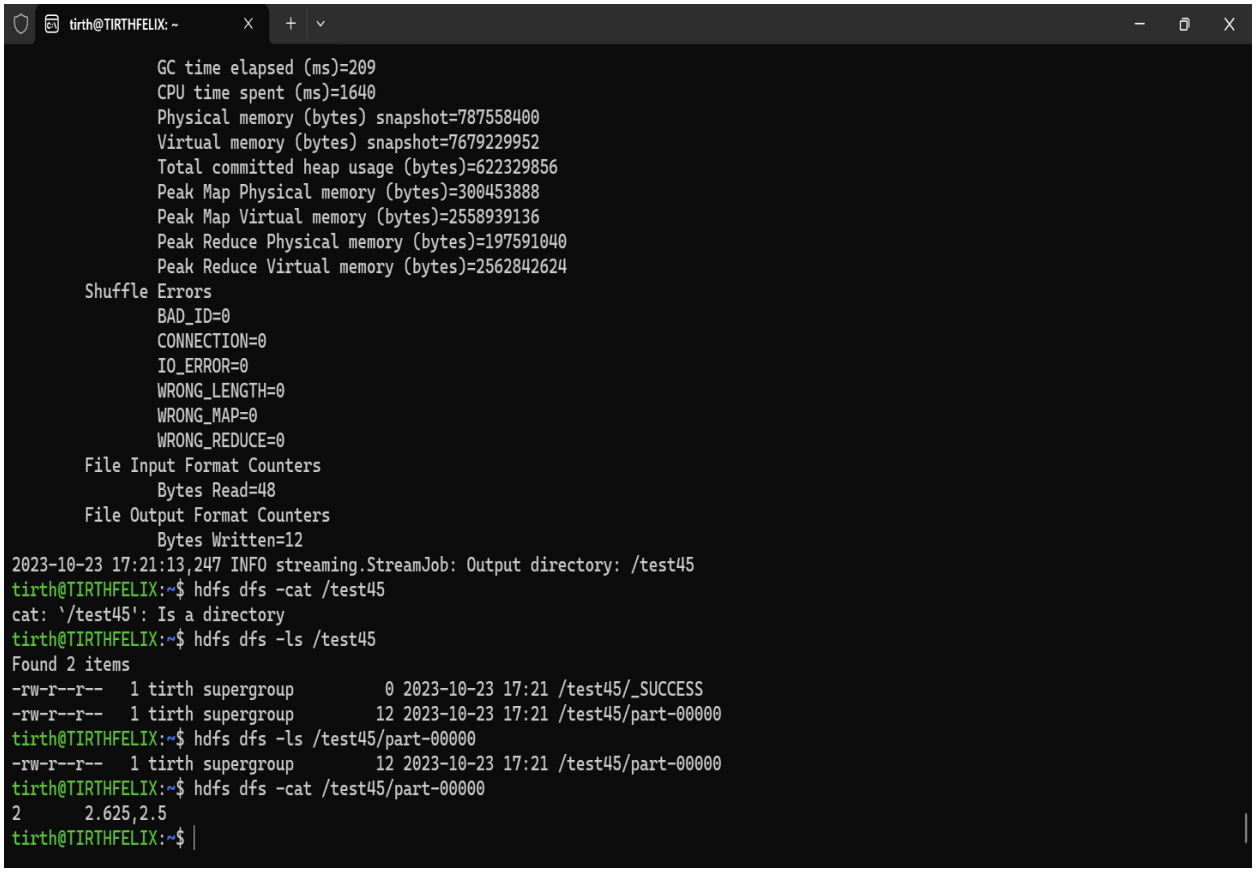
2.0,3.0

3.5,1.5

4.0,3.5

4.Input / Output:

```
tirth@TIRTHFELIX:~$ hadoop jar ~/hadoop/hadoop-3.3.2/share/hadoop/tools/lib/hadoop-streaming-3.3.2.jar -file /home/tirth/hadoop/mapper.py -mapper "python3 mapper.py" -file /home/tirth/hadoop/reducer.py -reducer "python3 reducer.py" -input /Number4.txt -output /test46
2023-10-23 17:24:04,605 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/tirth/hadoop/mapper.py, /home/tirth/hadoop/reducer.py, /tmp/hadoop-unjar7747853329624617316/] [] /tmp/streamjob7603173224858611248.jar tmpDir=null
2023-10-23 17:24:05,306 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-10-23 17:24:05,461 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-10-23 17:24:05,656 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/tirth/.staging/job_1698058215134_0019
2023-10-23 17:24:06,781 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-23 17:24:07,268 INFO mapreduce.JobSubmitter: number of splits:2
2023-10-23 17:24:07,386 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698058215134_0019
2023-10-23 17:24:07,386 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-23 17:24:07,570 INFO conf.Configuration: resource-types.xml not found
2023-10-23 17:24:07,571 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-23 17:24:07,627 INFO impl.YarnClientImpl: Submitted application application_1698058215134_0019
2023-10-23 17:24:07,664 INFO mapreduce.Job: The url to track the job: http://TIRTHFELIX.localdomain:8088/proxy/application_1698058215134_0019/
2023-10-23 17:24:07,667 INFO mapreduce.Job: Running job: job_1698058215134_0019
2023-10-23 17:24:13,806 INFO mapreduce.Job: Job job_1698058215134_0019 running in uber mode : false
2023-10-23 17:24:13,807 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-23 17:24:18,887 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-23 17:24:23,931 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-23 17:24:24,962 INFO mapreduce.Job: Job job_1698058215134_0019 completed successfully
2023-10-23 17:24:25,048 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=54
                FILE: Number of bytes written=836711
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```


tirth@TIRTHFELIX: ~
GC time elapsed (ms)=209
CPU time spent (ms)=1640
Physical memory (bytes) snapshot=787558400
Virtual memory (bytes) snapshot=7679229952
Total committed heap usage (bytes)=622329856
Peak Map Physical memory (bytes)=300453888
Peak Map Virtual memory (bytes)=2558939136
Peak Reduce Physical memory (bytes)=197591040
Peak Reduce Virtual memory (bytes)=2562842624
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=48
File Output Format Counters
Bytes Written=12
2023-10-23 17:21:13,247 INFO streaming.StreamJob: Output directory: /test45
tirth@TIRTHFELIX:~$ hdfs dfs -cat /test45
cat: `/test45': Is a directory
tirth@TIRTHFELIX:~$ hdfs dfs -ls /test45
Found 2 items
-rw-r--r-- 1 tirth supergroup 0 2023-10-23 17:21 /test45/_SUCCESS
-rw-r--r-- 1 tirth supergroup 12 2023-10-23 17:21 /test45/part-00000
tirth@TIRTHFELIX:~$ hdfs dfs -ls /test45/part-00000
-rw-r--r-- 1 tirth supergroup 12 2023-10-23 17:21 /test45/part-00000
tirth@TIRTHFELIX:~$ hdfs dfs -cat /test45/part-00000
2 2.625,2.5
tirth@TIRTHFELIX:~$


tirth@TIRTHFELIX:~$ hdfs dfs -cat /test46/part-00000
2 2.625,2.5

THE END

:)