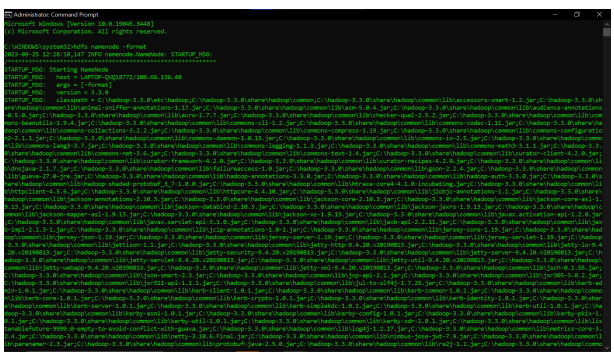
Assignment 1- Khaled Gaber - #1004144302

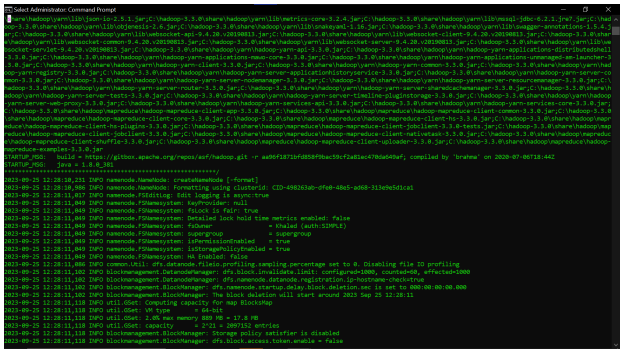
Questions:

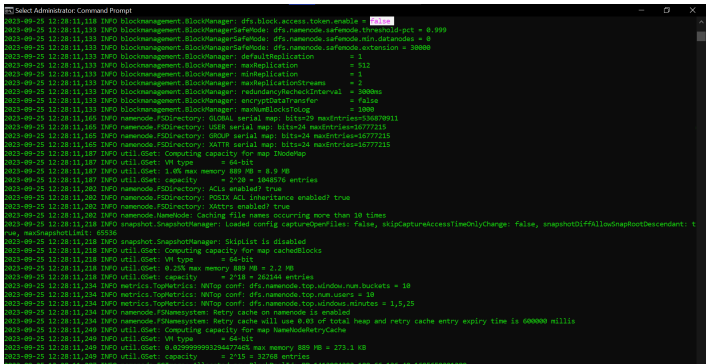
1) *[Marks: 15] Implement a Map Reduce program for counting the number of lines in a document. Use the 'shakespeare.txt' file and download it from Quercus. Please submit input/output files with code*

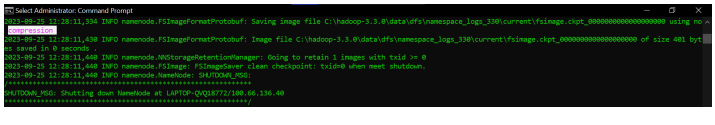
The number of lines (total including blank lines) is 58483

Hdfs namenote -format



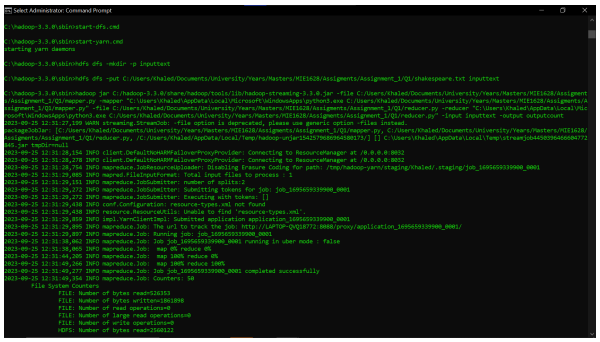


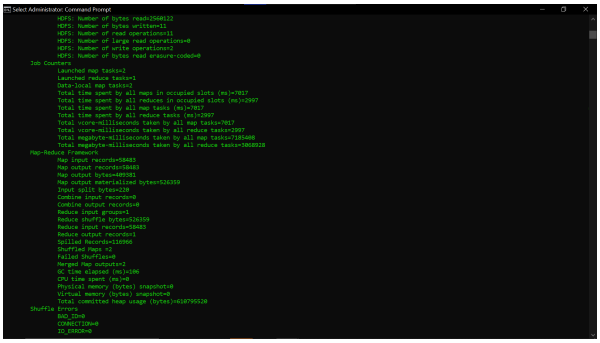


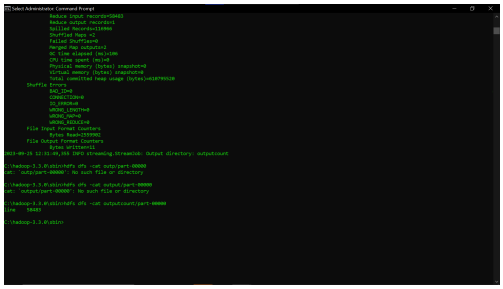


start-dfs.cmd
start-yarn.cmd

Hdfs dfs -mkdir -p inputtext





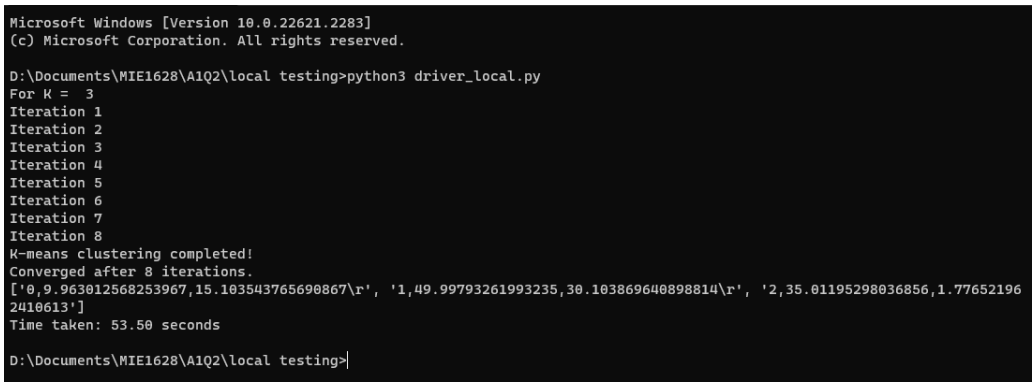


```
                Bytes Written=11
2023-09-25 12:31:49,355 INFO streaming.StreamJob: Output directory: outputcount

C:\hadoop-3.3.0\sbin>hdfs dfs -cat outp/part-00000
cat: `outp/part-00000': No such file or directory

C:\hadoop-3.3.0\sbin>hdfs dfs -cat output/part-00000
cat: `output/part-00000': No such file or directory

C:\hadoop-3.3.0\sbin>hdfs dfs -cat outputcount/part-00000
line    58483

C:\hadoop-3.3.0\sbin>hdfs dfs -ls
Found 2 items
drwxr-xr-x   - Khaled supergroup          0 2023-09-25 12:29 inputtext
drwxr-xr-x   - Khaled supergroup          0 2023-09-25 12:31 outputcount

C:\hadoop-3.3.0\sbin>hdfs dfs -ls outputcount
Found 2 items
-rw-r--r--   1 Khaled supergroup          0 2023-09-25 12:31 outputcount/_SUCCESS
-rw-r--r--   1 Khaled supergroup         11 2023-09-25 12:31 outputcount/part-00000

C:\hadoop-3.3.0\sbin>hdfs dfs -ls inputtext
Found 1 items
-rw-r--r--   1 Khaled supergroup    2555806 2023-09-25 12:29 inputtext/shakespeare.txt

C:\hadoop-3.3.0\sbin>hdfs dfs -get outputcount

C:\hadoop-3.3.0\sbin>
```

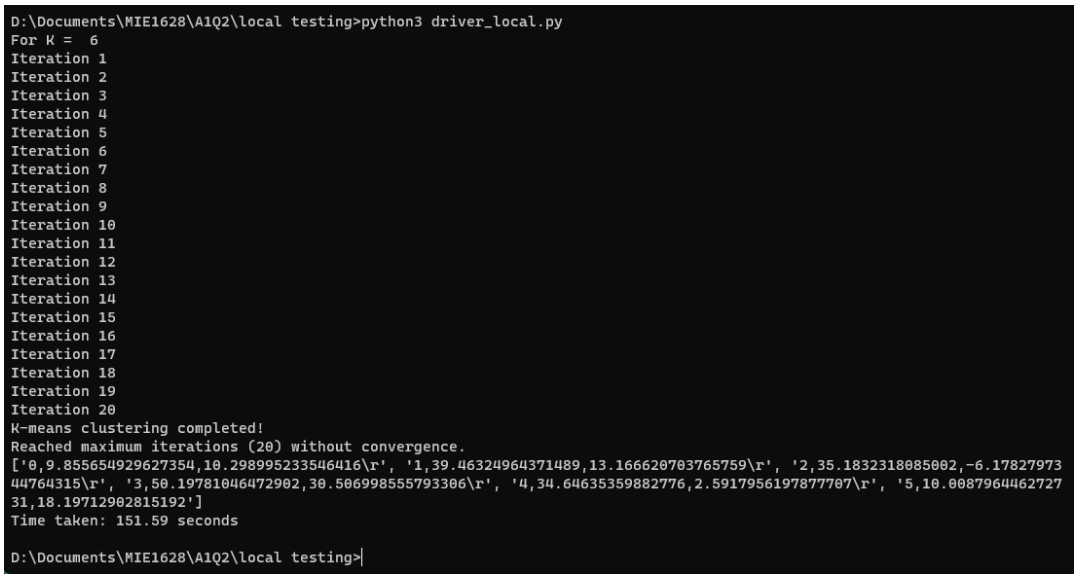
2) *[Marks: 45] Apply K-means clustering on Map Reduce using k = 3 and k = 6 clusters on the given dataset, list the cluster labels or centroids, the number of iterations for convergence or use maximum iterations = 20 and time/duration.*

Here it is working locally: type data_points.txt | python3 mapper.py | sort | python3 reducer.py

k=3



k=6



Here it is on MapReduce

k=1 (Command Testing)

hadoop jar C:\hadoop-3.3.0\share\hadoop\tools\lib\hadoop-streaming-3.3.0.jar -mapper "C:\Users\khale\AppData\Local\Microsoft\WindowsApps\python3.exe C:\MIE1628\A1Q2\mapper.py" -reducer "C:\Users\khale\AppData\Local\Microsoft\WindowsApps\python3.exe C:\MIE1628\A1Q2\reducer.py" -input /input/kmeans/* -output /output/kmeans/results

```
C:\MIE1628\A1Q2>hadoop jar C:\hadoop-3.3.0\share\hadoop\tools\lib\hadoop-streaming-3.3.0.jar -mapper "C:\Users\khale\App
Data\Local\Microsoft\WindowsApps\python3.exe C:\MIE1628\A1Q2\mapper.py" -reducer "C:\Users\khale\AppData\Local\Microsoft
\WindowsApps\python3.exe C:\MIE1628\A1Q2\reducer.py" -input /input/kmeans/* -output /output/kmeans/results
2023-09-28 22:05:00,242 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-09-28 22:05:00,290 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-09-28 22:05:00,291 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-09-28 22:05:00,326 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-09-28 22:05:00,625 INFO mapred.FileInputFormat: Total input files to process : 2
2023-09-28 22:05:00,667 INFO mapreduce.JobSubmitter: number of splits:2
2023-09-28 22:05:00,717 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1923691344_0001
2023-09-28 22:05:00,717 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-28 22:05:00,793 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-09-28 22:05:00,794 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-09-28 22:05:00,794 INFO mapreduce.Job: Running job: job_local1923691344_0001
2023-09-28 22:05:00,794 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2023-09-28 22:05:00,801 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:05:00,802 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:05:00,826 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-09-28 22:05:00,828 INFO mapred.LocalJobRunner: Starting task: attempt_local1923691344_0001_m_000000_0
2023-09-28 22:05:00,840 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:05:00,840 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:05:00,844 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:05:00,860 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@27f9a9e2
2023-09-28 22:05:00,870 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:19000/input/kmeans/data_points.txt:0+37938
809
2023-09-28 22:05:00,881 INFO mapred.MapTask: numReduceTasks: 1
2023-09-28 22:05:00,921 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-09-28 22:05:00,921 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-09-28 22:05:00,921 INFO mapred.MapTask: soft limit at 83886080
2023-09-28 22:05:00,921 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-09-28 22:05:00,921 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-09-28 22:05:00,923 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuff
er
2023-09-28 22:05:00,924 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\mapper.py]
2023-09-28 22:05:00,925 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.tas
k.output.dir
2023-09-28 22:05:00,926 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.
start
2023-09-28 22:05:00,926 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.is
map
2023-09-28 22:05:00,926 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attemp
t.id
2023-09-28 22:05:00,926 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2023-09-28 22:05:00,926 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.l
ocal.dir
2023-09-28 22:05:00,927 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.f
ile
2023-09-28 22:05:00,927 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprec
ords
2023-09-28 22:05:00,927 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input
.length
2023-09-28 22:05:00,927 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2023-09-28 22:05:00,927 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2023-09-28 22:05:00,927 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task
.partition
2023-09-28 22:05:01,081 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,081 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,082 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,084 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,092 INFO streaming.PipeMapRed: Records R/W=3548/1
2023-09-28 22:05:01,118 INFO streaming.PipeMapRed: R/W/S=10000/6752/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,422 INFO streaming.PipeMapRed: R/W/S=100000/98865/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,742 INFO streaming.PipeMapRed: R/W/S=200000/198149/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:01,802 INFO mapreduce.Job: Job job_local1923691344_0001 running in uber mode : false
2023-09-28 22:05:01,802 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-28 22:05:02,064 INFO streaming.PipeMapRed: R/W/S=300000/297646/0 in:300000=300000/1 [rec/s] out:297646=297646/1
[rec/s]
2023-09-28 22:05:02,384 INFO streaming.PipeMapRed: R/W/S=400000/396925/0 in:400000=400000/1 [rec/s] out:396925=396925/1
```

```
rec/s]
2023-09-28 22:05:02,384 INFO streaming.PipeMapRed: R/W/S=400000/396925/0 in:400000=400000/1 [rec/s] out:396925=396925/1
[rec/s]
2023-09-28 22:05:02,705 INFO streaming.PipeMapRed: R/W/S=500000/496418/0 in:500000=500000/1 [rec/s] out:496418=496418/1
[rec/s]
2023-09-28 22:05:03,036 INFO streaming.PipeMapRed: R/W/S=600000/599104/0 in:300000=600000/2 [rec/s] out:299592=599184/2
[rec/s]
2023-09-28 22:05:03,360 INFO streaming.PipeMapRed: R/W/S=700000/698457/0 in:350000=700000/2 [rec/s] out:349228=698457/2
[rec/s]
2023-09-28 22:05:03,680 INFO streaming.PipeMapRed: R/W/S=800000/797731/0 in:400000=800000/2 [rec/s] out:398865=797731/2
[rec/s]
2023-09-28 22:05:04,004 INFO streaming.PipeMapRed: R/W/S=900000/897232/0 in:300000=900000/3 [rec/s] out:299077=897232/3
[rec/s]
2023-09-28 22:05:04,328 INFO streaming.PipeMapRed: R/W/S=1000000/996722/0 in:333333=1000000/3 [rec/s] out:332240=996722/
3 [rec/s]
2023-09-28 22:05:04,343 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:05:04,344 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:05:04,346 INFO mapred.LocalJobRunner:
2023-09-28 22:05:04,346 INFO mapred.MapTask: Starting flush of map output
2023-09-28 22:05:04,347 INFO mapred.MapTask: Spilling map output
2023-09-28 22:05:04,347 INFO mapred.MapTask: bufstart = 0; bufend = 38938011; bufvoid = 104857600
2023-09-28 22:05:04,347 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 22214400(88857600); length = 3999997
/6553600
2023-09-28 22:05:04,596 INFO mapred.MapTask: Finished spill 0
2023-09-28 22:05:04,602 INFO mapred.Task: Task:attempt_local1923691344_0001_m_000000_0 is done. And is in the process of
 committing
2023-09-28 22:05:04,604 INFO mapred.LocalJobRunner: Records R/W=3548/1
2023-09-28 22:05:04,605 INFO mapred.Task: Task 'attempt_local1923691344_0001_m_000000_0' done.
2023-09-28 22:05:04,609 INFO mapred.Task: Final Counters for attempt_local1923691344_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=141897
                FILE: Number of bytes written=41700544
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938009
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1000000
                Map output records=1000000
                Map output bytes=38938011
                Map output materialized bytes=40938017
                Input split bytes=101
                Combine input records=0
                Spilled Records=1000000
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=383778816
        File Input Format Counters
                Bytes Read=37938009
2023-09-28 22:05:04,609 INFO mapred.LocalJobRunner: Finishing task: attempt_local1923691344_0001_m_000000_0
2023-09-28 22:05:04,610 INFO mapred.LocalJobRunner: Starting task: attempt_local1923691344_0001_m_000001_0
2023-09-28 22:05:04,611 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:05:04,611 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:05:04,611 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:05:04,627 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@62ccfbd3
2023-09-28 22:05:04,628 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:19000/input/kmeans/centroids.txt:0+117
2023-09-28 22:05:04,630 INFO mapred.MapTask: numReduceTasks: 1
2023-09-28 22:05:04,638 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-09-28 22:05:04,638 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-09-28 22:05:04,638 INFO mapred.MapTask: soft limit at 83886080
2023-09-28 22:05:04,638 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-09-28 22:05:04,639 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-09-28 22:05:04,639 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuff
er
2023-09-28 22:05:04,639 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\mapper.py]
```

```
2023-09-28 22:05:04,735 INFO streaming.PipeMapRed: Records R/W=3/1
2023-09-28 22:05:04,739 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:05:04,740 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:05:04,740 INFO mapred.LocalJobRunner:
2023-09-28 22:05:04,740 INFO mapred.MapTask: Starting flush of map output
2023-09-28 22:05:04,741 INFO mapred.MapTask: Spilling map output
2023-09-28 22:05:04,741 INFO mapred.MapTask: bufstart = 0; bufend = 72; bufvoid = 104857600
2023-09-28 22:05:04,741 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214388(104857552); length = 9/6553
600
2023-09-28 22:05:04,743 INFO mapred.MapTask: Finished spill 0
2023-09-28 22:05:04,745 INFO mapred.Task: Task:attempt_local1923691344_0001_m_000001_0 is done. And is in the process of
 committing
2023-09-28 22:05:04,748 INFO mapred.LocalJobRunner: Records R/W=3/1
2023-09-28 22:05:04,748 INFO mapred.Task: Task 'attempt_local1923691344_0001_m_000001_0' done.
2023-09-28 22:05:04,749 INFO mapred.Task: Final Counters for attempt_local1923691344_0001_m_000001_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=142116
                FILE: Number of bytes written=4170660
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938126
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Map output bytes=72
                Map output materialized bytes=84
                Input split bytes=99
                Combine input records=0
                Spilled Records=3
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=383778816
        File Input Format Counters
                Bytes Read=117
2023-09-28 22:05:04,749 INFO mapred.LocalJobRunner: Finishing task: attempt_local1923691344_0001_m_000001_0
2023-09-28 22:05:04,749 INFO mapred.LocalJobRunner: map task executor complete.
2023-09-28 22:05:04,752 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2023-09-28 22:05:04,752 INFO mapred.LocalJobRunner: Starting task: attempt_local1923691344_0001_r_000000_0
2023-09-28 22:05:04,755 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:05:04,755 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:05:04,755 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:05:04,770 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@1c0cc08f
2023-09-28 22:05:04,772 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shu
ffle@409d49e4
2023-09-28 22:05:04,773 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-09-28 22:05:04,784 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=835846
16, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2023-09-28 22:05:04,785 INFO reduce.EventFetcher: attempt_local1923691344_0001_r_000000_0 Thread started: EventFetcher f
or fetching Map Completion Events
2023-09-28 22:05:04,802 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1923691344_
0001_m_000000_0 decomp: 40938013 len: 40938017 to MEMORY
2023-09-28 22:05:04,827 INFO reduce.InMemoryMapOutput: Read 40938013 bytes from map-output for attempt_local1923691344_0
001_m_000000_0
2023-09-28 22:05:04,827 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 40938013, inMemoryMapOutp
uts.size() -> 1, commitMemory -> 0, usedMemory ->40938013
2023-09-28 22:05:04,829 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1923691344_
0001_m_000001_0 decomp: 80 len: 84 to MEMORY
2023-09-28 22:05:04,833 INFO reduce.InMemoryMapOutput: Read 80 bytes from map-output for attempt_local1923691344_0001_m_
000001_0
2023-09-28 22:05:04,833 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 80, inMemoryMapOutputs.si
ze() -> 2, commitMemory -> 40938013, usedMemory ->40938093
2023-09-28 22:05:04,833 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2023-09-28 22:05:04,834 INFO mapred.LocalJobRunner: 2 / 2 copied.
2023-09-28 22:05:04,834 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and 0 on-disk map-o
utputs
```

```
2023-09-28 22:05:04,840 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 40938085 by
tes
2023-09-28 22:05:04,843 INFO mapreduce.Job:  map 100% reduce 0%
2023-09-28 22:05:05,013 INFO reduce.MergeManagerImpl: Merged 2 segments, 40938093 bytes to disk to satisfy reduce memory
limit
2023-09-28 22:05:05,014 INFO reduce.MergeManagerImpl: Merging 1 files, 40938095 bytes from disk
2023-09-28 22:05:05,014 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2023-09-28 22:05:05,014 INFO mapred.Merger: Merging 1 sorted segments
2023-09-28 22:05:05,019 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 40938087 by
tes
2023-09-28 22:05:05,019 INFO mapred.LocalJobRunner: 2 / 2 copied.
2023-09-28 22:05:05,020 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\reducer.py]
2023-09-28 22:05:05,021 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2023-09-28 22:05:05,021 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2023-09-28 22:05:05,050 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,050 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,051 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,053 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,134 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,221 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,328 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,425 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,521 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,617 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,715 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,808 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,903 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:05,997 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:05:06,094 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:1000000=1000000/1 [rec/s] out:0=0/1 [rec/s]
2023-09-28 22:05:06,098 INFO streaming.PipeMapRed: Records R/W=1000003/1
2023-09-28 22:05:06,101 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:05:06,101 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:05:06,135 INFO mapred.Task: Task:attempt_local1923691344_0001_r_000000_0 is done. And is in the process of
 committing
2023-09-28 22:05:06,137 INFO mapred.LocalJobRunner: 2 / 2 copied.
2023-09-28 22:05:06,137 INFO mapred.Task: Task attempt_local1923691344_0001_r_000000_0 is allowed to commit now
2023-09-28 22:05:06,141 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1923691344_0001_r_000000_0'
to hdfs://0.0.0.0:19000/output/kmeans/results
2023-09-28 22:05:06,143 INFO mapred.LocalJobRunner: Records R/W=1000003/1 > reduce
2023-09-28 22:05:06,143 INFO mapred.Task: Task 'attempt_local1923691344_0001_r_000000_0' done.
2023-09-28 22:05:06,145 INFO mapred.Task: Final Counters for attempt_local1923691344_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=82018376
                FILE: Number of bytes written=82638755
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938126
                HDFS: Number of bytes written=119
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=40938101
                Reduce input records=1000003
                Reduce output records=3
                Spilled Records=1000003
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1
                Total committed heap usage (bytes)=384383104
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
```

```
2023-09-28 22:05:06,865 INFO streaming.StreamJob: Output directory: /output/kmeans/results

C:\MIE1628\A1Q2>hdfs dfs -ls \
Found 4 items
drwxr-xr-x   - khale supergroup          0 2023-09-28 21:59 /input
drwxr-xr-x   - khale supergroup          0 2023-09-28 17:55 /output
drwx------   - khale supergroup          0 2023-09-28 17:42 /tmp
drwxr-xr-x   - khale supergroup          0 2023-09-28 17:26 /user

C:\MIE1628\A1Q2>hdfs dfs -get \output\* C:\MIE1628\A1Q2\Results

C:\MIE1628\A1Q2>hdfs dfs -get \input\* C:\MIE1628\A1Q2\Results

C:\MIE1628\A1Q2>hdfs dfs -cat \output\kmeans\part-00000
cat: `outputkmeanspart-00000': No such file or directory

C:\MIE1628\A1Q2>hdfs dfs -ls \output\kmeans
Found 1 items
drwxr-xr-x   - khale supergroup          0 2023-09-28 22:05 /output/kmeans/results

C:\MIE1628\A1Q2>hdfs dfs -ls \output\kmeans\results
Found 2 items
-rw-r--r--   1 khale supergroup          0 2023-09-28 22:05 /output/kmeans/results/_SUCCESS
-rw-r--r--   1 khale supergroup        119 2023-09-28 22:05 /output/kmeans/results/part-00000

C:\MIE1628\A1Q2>hdfs dfs -cat \output\kmeans\results\part-00000
0,9.921463822213996,14.994793104788538
1,43.62540411580502,19.74851471321919
2,37.01713640846856,-2.726666152642699
```

k=3

This took 9 iterations in 120.76 seconds, and the finals centroids were:
The Final Centroids are: [
('0', (9.962931682425607, 15.10369315620118)),
('1', (49.99793261993191, 30.103869640898743)),
('2', (35.01195298036899, 1.7765219624106197))]

Note due to issues with running this, I followed the solution on Piazza, question @49, changing the property value from yarn to local.

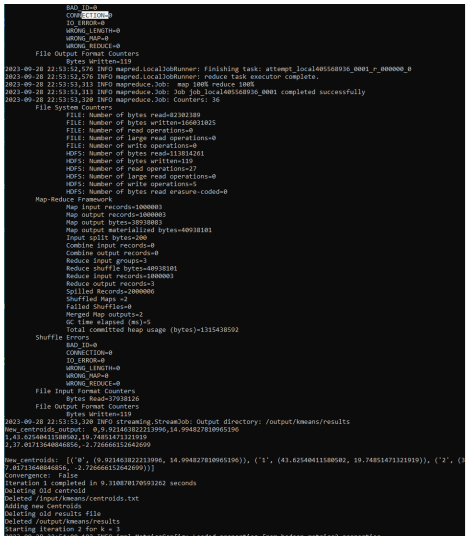
The initial centroids were taken as the first 3 points from “data_points.txt”

The output for this is very long, and so I only show screenshots of the input, one iteration, and the final below:

```
C:\MIE1628\A1Q2>python3 driver.py
Starting Main()
uploading the data_points.txt file
put: `/input/kmeans/data_points.txt': File exists
initializing initial centroid values
Deleting old results file
rm: `/output/kmeans/results': No such file or directory
uploading the centroids.txt file
put: `/input/kmeans/centroids.txt': File exists
Starting iteration 1 for k = 3
2023-09-28 22:53:46,722 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-09-28 22:53:46,767 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-09-28 22:53:46,767 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-09-28 22:53:46,790 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-09-28 22:53:47,089 INFO mapred.FileInputFormat: Total input files to process : 2
2023-09-28 22:53:47,123 INFO mapreduce.JobSubmitter: number of splits:2
2023-09-28 22:53:47,168 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local405568936_0001
2023-09-28 22:53:47,168 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-28 22:53:47,243 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-09-28 22:53:47,244 INFO mapreduce.Job: Running job: job_local405568936_0001
2023-09-28 22:53:47,245 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-09-28 22:53:47,245 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2023-09-28 22:53:47,248 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:53:47,248 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:53:47,269 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-09-28 22:53:47,271 INFO mapred.LocalJobRunner: Starting task: attempt_local405568936_0001_m_000000_0
2023-09-28 22:53:47,283 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:53:47,283 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:53:47,287 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:53:47,303 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@5ba4b32a
2023-09-28 22:53:47,308 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:19000/input/kmeans/data_points.txt:0+37938
000
2023-09-28 22:53:47,319 INFO mapred.MapTask: numReduceTasks: 1
2023-09-28 22:53:47,350 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-09-28 22:53:47,350 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-09-28 22:53:47,350 INFO mapred.MapTask: soft limit at 83886080
2023-09-28 22:53:47,351 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-09-28 22:53:47,351 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-09-28 22:53:47,352 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuff
er
2023-09-28 22:53:47,354 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\mapper.py]
2023-09-28 22:53:47,356 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.tas
k.output.dir
2023-09-28 22:53:47,356 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.
start
2023-09-28 22:53:47,356 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.is
map
2023-09-28 22:53:47,356 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attemp
t.id
2023-09-28 22:53:47,357 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
2023-09-28 22:53:47,357 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.l
ocal.dir
2023-09-28 22:53:47,358 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.f
ile
2023-09-28 22:53:47,358 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprec
ords
2023-09-28 22:53:47,358 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input
.length
2023-09-28 22:53:47,358 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
2023-09-28 22:53:47,358 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
2023-09-28 22:53:47,358 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task
.partition
2023-09-28 22:53:47,484 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:47,484 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:47,485 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:47,486 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:47,490 INFO streaming.PipeMapRed: Records R/W=3548/1
2023-09-28 22:53:47,517 INFO streaming.PipeMapRed: R/W/S=10000/6752/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:47,822 INFO streaming.PipeMapRed: R/W/S=100000/58865/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:48,259 INFO mapreduce.Job: Job job_local405568936_0001 running in uber mode : false
2023-09-28 22:53:48,259 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-28 22:53:48,469 INFO streaming.PipeMapRed: R/W/S=300000/297646/0 in:300000=300000/1 [rec/s] out:297646=297646/1
[rec/s]
2023-09-28 22:53:48,791 INFO streaming.PipeMapRed: R/W/S=400000/396925/0 in:400000=400000/1 [rec/s] out:396925=396925/1
[rec/s]
2023-09-28 22:53:49,116 INFO streaming.PipeMapRed: R/W/S=500000/496418/0 in:500000=500000/1 [rec/s] out:496418=496418/1
[rec/s]
2023-09-28 22:53:49,452 INFO streaming.PipeMapRed: R/W/S=600000/599184/0 in:300000=600000/2 [rec/s] out:299592=599184/2
[rec/s]
2023-09-28 22:53:49,775 INFO streaming.PipeMapRed: R/W/S=700000/698457/0 in:350000=700000/2 [rec/s] out:349228=698457/2
[rec/s]
2023-09-28 22:53:50,098 INFO streaming.PipeMapRed: R/W/S=800000/797936/0 in:400000=800000/2 [rec/s] out:398968=797936/2
[rec/s]
2023-09-28 22:53:50,422 INFO streaming.PipeMapRed: R/W/S=900000/897232/0 in:300000=900000/3 [rec/s] out:299077=897232/3
[rec/s]
2023-09-28 22:53:50,746 INFO streaming.PipeMapRed: R/W/S=1000000/996722/0 in:333333=1000000/3 [rec/s] out:332240=996722/
3 [rec/s]
2023-09-28 22:53:50,761 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:53:50,762 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:53:50,764 INFO mapred.LocalJobRunner:
2023-09-28 22:53:50,764 INFO mapred.MapTask: Starting flush of map output
2023-09-28 22:53:50,764 INFO mapred.MapTask: Spilling map output
2023-09-28 22:53:50,766 INFO mapred.MapTask: bufstart = 0; bufend = 38938011; bufvoid = 104857600
2023-09-28 22:53:50,767 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 22214400(88857600); length = 3999997
/6553600
2023-09-28 22:53:51,007 INFO mapred.MapTask: Finished spill 0
2023-09-28 22:53:51,014 INFO mapred.Task: Task:attempt_local405568936_0001_m_000000_0 is done. And is in the process of
committing
2023-09-28 22:53:51,015 INFO mapred.LocalJobRunner: Records R/W=3548/1
2023-09-28 22:53:51,015 INFO mapred.Task: Task 'attempt_local405568936_0001_m_000000_0' done.
2023-09-28 22:53:51,019 INFO mapred.Task: Final Counters for attempt_local405568936_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=141897
                FILE: Number of bytes written=41697566
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938009
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=1000000
                Map output records=1000000
                Map output bytes=38938011
                Map output materialized bytes=40938017
                Input split bytes=101
                Combine input records=0
                Spilled Records=1000000
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=368050176
        File Input Format Counters
                Bytes Read=37938009
2023-09-28 22:53:51,019 INFO mapred.LocalJobRunner: Finishing task: attempt_local405568936_0001_m_000000_0
2023-09-28 22:53:51,020 INFO mapred.LocalJobRunner: Starting task: attempt_local405568936_0001_m_000001_0
2023-09-28 22:53:51,021 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:53:51,022 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:53:51,022 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:53:51,040 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@753f10a8
2023-09-28 22:53:51,041 INFO mapred.MapTask: Processing split: hdfs://0.0.0.0:19000/input/kmeans/centroids.txt:0+117
2023-09-28 22:53:51,043 INFO mapred.MapTask: numReduceTasks: 1
2023-09-28 22:53:51,075 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2023-09-28 22:53:51,075 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2023-09-28 22:53:51,075 INFO mapred.MapTask: soft limit at 83886080
2023-09-28 22:53:51,075 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2023-09-28 22:53:51,076 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2023-09-28 22:53:51,077 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuff
```

```
2023-09-28 22:53:51,078 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\mapper.py]
2023-09-28 22:53:51,092 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,169 INFO streaming.PipeMapRed: Records R/W=3/1
2023-09-28 22:53:51,173 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:53:51,174 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:53:51,174 INFO mapred.LocalJobRunner:
2023-09-28 22:53:51,174 INFO mapred.MapTask: Starting flush of map output
2023-09-28 22:53:51,176 INFO mapred.MapTask: Spilling map output
2023-09-28 22:53:51,176 INFO mapred.MapTask: bufstart = 0; bufend = 72; bufvoid = 104857600
2023-09-28 22:53:51,177 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214388(104857552); length = 9/6553
600
2023-09-28 22:53:51,180 INFO mapred.MapTask: Finished spill 0
2023-09-28 22:53:51,182 INFO mapred.Task: Task:attempt_local405568936_0001_m_000001_0 is done. And is in the process of
committing
2023-09-28 22:53:51,183 INFO mapred.LocalJobRunner: Records R/W=3/1
2023-09-28 22:53:51,184 INFO mapred.Task: Task 'attempt_local405568936_0001_m_000001_0' done.
2023-09-28 22:53:51,185 INFO mapred.Task: Final Counters for attempt_local405568936_0001_m_000001_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=142116
                FILE: Number of bytes written=41697682
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938126
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Map output bytes=72
                Map output materialized bytes=84
                Input split bytes=99
                Combine input records=0
                Spilled Records=3
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=473432064
        File Input Format Counters
                Bytes Read=117
2023-09-28 22:53:51,185 INFO mapred.LocalJobRunner: Finishing task: attempt_local405568936_0001_m_000001_0
2023-09-28 22:53:51,185 INFO mapred.LocalJobRunner: map task executor complete.
2023-09-28 22:53:51,189 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2023-09-28 22:53:51,191 INFO mapred.LocalJobRunner: Starting task: attempt_local405568936_0001_r_000000_0
2023-09-28 22:53:51,195 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-09-28 22:53:51,195 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outpu
t directory:false, ignore cleanup failures: false
2023-09-28 22:53:51,195 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2023-09-28 22:53:51,212 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBase
dProcessTree@37bb9fb0
2023-09-28 22:53:51,214 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shu
ffle@4b95d714
2023-09-28 22:53:51,215 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2023-09-28 22:53:51,224 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=835846
16, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2023-09-28 22:53:51,225 INFO reduce.EventFetcher: attempt_local405568936_0001_r_000000_0 Thread started: EventFetcher fo
r fetching Map Completion Events
2023-09-28 22:53:51,245 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local405568936_0
001_m_000000_0 decomp: 40938013 len: 40938017 to MEMORY
2023-09-28 22:53:51,276 INFO reduce.InMemoryMapOutput: Read 40938013 bytes from map-output for attempt_local405568936_00
01_m_000000_0
2023-09-28 22:53:51,277 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 40938013, inMemoryMapOutp
uts.size() -> 1, commitMemory -> 0, usedMemory ->40938013
2023-09-28 22:53:51,279 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local405568936_0
001_m_000001_0 decomp: 80 len: 84 to MEMORY
2023-09-28 22:53:51,280 INFO reduce.InMemoryMapOutput: Read 80 bytes from map-output for attempt_local405568936_0001_m_0
00001_0
2023-09-28 22:53:51,280 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 80, inMemoryMapOutputs.si
ze() -> 2, commitMemory -> 40938013, usedMemory ->40938093
2023-09-28 22:53:51,281 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
```

```
2023-09-28 22:53:51,282 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and 0 on-disk map-o
utputs
2023-09-28 22:53:51,286 INFO mapred.Merger: Merging 2 sorted segments
2023-09-28 22:53:51,286 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size: 40938085 by
tes
2023-09-28 22:53:51,294 INFO mapreduce.Job:  map 100% reduce 0%
2023-09-28 22:53:51,451 INFO reduce.MergeManagerImpl: Merged 2 segments, 40938093 bytes to disk to satisfy reduce memory
limit
2023-09-28 22:53:51,452 INFO reduce.MergeManagerImpl: Merging 1 files, 40938095 bytes from disk
2023-09-28 22:53:51,452 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2023-09-28 22:53:51,452 INFO mapred.Merger: Merging 1 sorted segments
2023-09-28 22:53:51,457 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 40938087 by
tes
2023-09-28 22:53:51,458 INFO mapred.LocalJobRunner: 2 / 2 copied.
2023-09-28 22:53:51,458 INFO streaming.PipeMapRed: PipeMapRed exec [C:\Users\khale\AppData\Local\Microsoft\WindowsApps\p
ython3.exe, C:\MIE1628\A1Q2\reducer.py]
2023-09-28 22:53:51,459 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2023-09-28 22:53:51,459 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2023-09-28 22:53:51,487 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,487 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,488 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,490 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,567 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,658 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,754 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,853 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:51,951 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,049 INFO streaming.PipeMapRed: R/W/S=500000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,144 INFO streaming.PipeMapRed: R/W/S=600000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,240 INFO streaming.PipeMapRed: R/W/S=700000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,335 INFO streaming.PipeMapRed: R/W/S=800000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,429 INFO streaming.PipeMapRed: R/W/S=900000/0/0 in:NA [rec/s] out:NA [rec/s]
2023-09-28 22:53:52,525 INFO streaming.PipeMapRed: R/W/S=1000000/0/0 in:1000000=1000000/1 [rec/s] out:0=0/1 [rec/s]
2023-09-28 22:53:52,529 INFO streaming.PipeMapRed: Records R/W=1000003/1
2023-09-28 22:53:52,533 INFO streaming.PipeMapRed: MRErrorThread done
2023-09-28 22:53:52,534 INFO streaming.PipeMapRed: mapRedFinished
2023-09-28 22:53:52,565 INFO mapred.Task: Task:attempt_local405568936_0001_r_000000_0 is done. And is in the process of
committing
2023-09-28 22:53:52,566 INFO mapred.LocalJobRunner: 2 / 2 copied.
2023-09-28 22:53:52,566 INFO mapred.Task: Task attempt_local405568936_0001_r_000000_0 is allowed to commit now
2023-09-28 22:53:52,573 INFO output.FileOutputCommitter: Saved output of task 'attempt_local405568936_0001_r_000000_0' t
o hdfs://0.0.0.0:19000/output/kmeans/results
2023-09-28 22:53:52,573 INFO mapred.LocalJobRunner: Records R/W=1000003/1 > reduce
2023-09-28 22:53:52,574 INFO mapred.Task: Task 'attempt_local405568936_0001_r_000000_0' done.
2023-09-28 22:53:52,575 INFO mapred.Task: Final Counters for attempt_local405568936_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=82018376
                FILE: Number of bytes written=82635777
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=37938126
                HDFS: Number of bytes written=119
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=40938101
                Reduce input records=1000003
                Reduce output records=3
                Spilled Records=1000003
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=2
                Total committed heap usage (bytes)=473956352
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
```



```
2023-09-28 22:55:40,754 INFO streaming.StreamJob: Output directory: /output/kmeans/results
New_centroids_output:  0,9.962931682425607,15.10369315620118
1,49.99793261993191,30.103869640898743
2,35.01195298036899,1.7765219624106197

New_centroids:  [('0', (9.962931682425607, 15.10369315620118)), ('1', (49.99793261993191, 30.103869640898743)), ('2', (3
5.01195298036899, 1.7765219624106197))]
Convergence:  True
Iteration 9 completed in 9.318159341812134 seconds
9 Iterations in: 120.76522064208984 seconds
The Final Centroids are: [('0', (9.962931682425607, 15.10369315620118)), ('1', (49.99793261993191, 30.103869640898743)),
 ('2', (35.01195298036899, 1.7765219624106197))]
```

Output files were saved using:

hdfs dfs -get /output/kmeans/results/* C:\MIE1628\A1Q2\Final_Results\k3

k=6

This did not converge, maxing out at 20 iterations.
20 Iterations in: 291.4041225910187 seconds
The Final Centroids are: [
('0', (9.855663875034324, 10.30023604782646)),
('1', (39.46324964371434, 13.166620703765755)),
('2', (35.18323180850043, -6.17827973447644)),
('3', (50.19781046472971, 30.506998555793103)),
('4', (34.64635359882755, 2.5917956197877845)),
('5', (10.008599091020962, 18.198393029289274))]

The initial centroids were taken as the first 6 points from “data_points.txt”

```
C:\MIE1628\A1Q2>python3 driver.py
Starting Main()
uploading the data_points.txt file
put: `/input/kmeans/data_points.txt': File exists
initializing initial centroid values
Deleting old results file
Deleted /output/kmeans/results
uploading the centroids.txt file
put: `/input/kmeans/centroids.txt': File exists
Starting iteration 1 for k = 6
```

```
New_centroids:  [('0', (9.855663875034324, 10.30023604782646)), ('1', (39.46324964371434, 13.166620703765755)), ('2', (3
5.18323180850043, -6.17827973447644)), ('3', (50.19781046472971, 30.506998555793103)), ('4', (34.64635359882755, 2.59179
56197877845)), ('5', (10.008599091020962, 18.198393029289274))]
Convergence:  False
Iteration 20 completed in 10.352383136749268 seconds
Saving Output file
20 Iterations in: 295.6279318332672 seconds
The Final Centroids are: [('0', (9.855663875034324, 10.30023604782646)), ('1', (39.46324964371434, 13.166620703765755)),
 ('2', (35.18323180850043, -6.17827973447644)), ('3', (50.19781046472971, 30.506998555793103)), ('4', (34.64635359882755
, 2.5917956197877845)), ('5', (10.008599091020962, 18.198393029289274))]
```

3) *[Marks: 10] Explain the advantages and disadvantages of using K-Means Clustering with MapReduce*.

Advantages:

- Able to handle expansive datasets that are beyond the storage capacity of a single machine.
- Work parallelization that results in significantly enhanced clustering speed.
- Robust fault tolerance mechanism: in case a node experiences a failure during computation, the framework autonomously redistributes the workload to other operational nodes, ensuring the processing advances without disruption.

Disadvantages:

- Increased complexity: Implementing K-means using MapReduce is slightly more complex than running it locally
- Overhead costs: The process entails multiple phases including data partitioning, mapping, shuffling, and reducing, which incur additional computational and communication overheads. In situations where data processing tasks are small or simplistic, the overhead associated with configuring and operating the MapReduce framework could just not be worth it
- Since we need to divide the tasks into 2 parts, the Map and Reduce functions, this might not be suitable for all types of tasks, possibly limiting its suitability for certain clustering or analytical challenges.

Please read the paper which is provided with the assignment in the Quercus and answer the following questions.

4) [Marks: 10] Can we reduce the number of distance comparisons by applying the Canopy Selection? Which distance metric should we use for the canopy clustering and why?

- Yes, Canopy Selection can significantly diminish the number of distance comparisons required. The primary concept revolves around economically partitioning the data into subsets initially, followed by computing the distances solely between points within the same subset, thereby lowering the overall number of distance computations needed.
- The distance metric we decide to use is dependant on the nature of the data, as is task-specific

5) [Marks: 10] Is it possible to apply Canopy Selection on MapReduce? If yes, then explain in words, how would you implement it.

Yes it is, we can structure it in two primary phases: Canopy creation and allocation of data points.

Canopy Creation

- Mappers process segments of data points to generate initial canopies using a specified algorithm.
    - A random point is selected to initiate a canopy.
    - The algorithm iterates through remaining points:
        - Points within distance < T1 from the initial point are added to the canopy.
        - If the distance is also < T2, the point is removed from the set.
    - This loop continues until all points are either clustered or removed, generating a set of canopies.
    - Mappers return the centroids of these initial canopies.
- Reducers receive the initial centroids:
    - Apply the canopy measure and thresholds once more to refine the canopy centroids.
    - Output a final set of canopy centroids for the next phase.

Clustering Phase:

- In this phase, Mappers are fed the finalized canopy centroids, utilizing them to classify the data points, subsequently returning pairs composed of (canopy centroid identifier, data point).
- Reducers then consolidate data based on the canopy centroids' identifiers, culminating in the output of the final canopies.

6) [Marks: 10] Is it possible to combine the Canopy Selection with K-Means on MapReduce? If yes, then explain in words, how would you do that.

Yes it is possible, and would likely result in a speedup.The main potential steps are outlined below:

1. Canopy Selection Implementation:
    - Execute the canopy selection as described in Question 5 to partition the data into rough clusters or canopies initially.

2. K-means Initialization:
    - Within each canopy, initialize k prototypes.

3. K-means Iterations:

- Mappers:

    - Each data point is associated with a list of canopies it belongs to, and consequently, a list of candidate prototypes.
    - Mappers calculate the closest prototype for each point, outputting pairs of the closest prototype ID and the data point.

- Combiners:
    - This stage can aggregate data from mappers to reduce the amount of data transferred to reducers. They can compile lists of points associated with each prototype ID.

- Reducers:
    - Post-combination, reducers receive pairs consisting of a prototype ID and a list of points associated with that prototype.
    - The new centroids are computed by calculating the mean of the points within each cluster.
    - Output the new centroids (prototype ID, mean of the points) for the iterations.