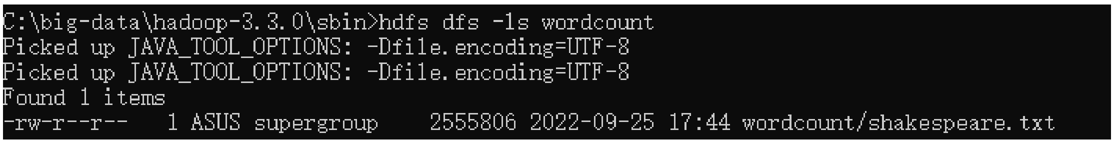
Course: MIE 1628 Cloud-based Data Analytics
Name: Wenxin Li
Student ID: 1007508724
Due Date: Sept 27, 2022

Assignment 1 Assignment on MapReduce

Q1. Line count of Shakespeare.txt

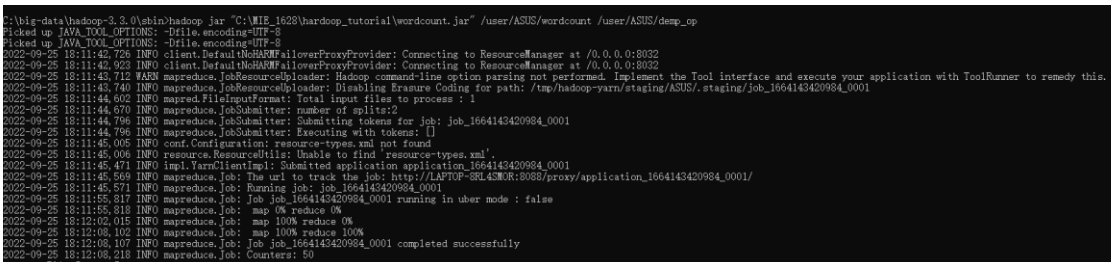
1. Upload Shakespeare.txt to hdfs



```
C:\big-data\hadoop-3.3.0\sbin>hdfs dfs -ls wordcount
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Found 1 items
-rw-r--r--   1 ASUS supergroup    2555806 2022-09-25 17:44 wordcount/shakespeare.txt
```

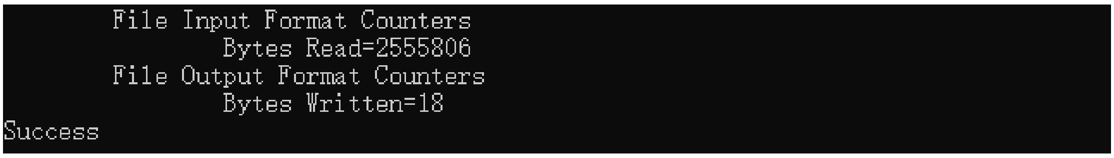
2. Input: hadoop jar <Path of jar> <Path of input> <Path of output>

hadoop jar "C:\MIE_1628\linecount\LineCount.jar" /user/ASUS/linecount /user/ASUS/result_line



```
C:\big-data\hadoop-3.3.0\sbin>hadoop jar "C:\MIE_1628\hardoop_tutorial\wordcount.jar" /user/ASUS/wordcount /user/ASUS/demp_op
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
2022-09-25 18:11:42,726 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-25 18:11:42,923 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-25 18:11:43,712 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-09-25 18:11:43,740 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ASUS/.staging/job_1664143420984_0001
2022-09-25 18:11:44,602 INFO mapred.FileInputFormat: Total input files to process : 1
2022-09-25 18:11:44,670 INFO mapreduce.JobSubmitter: number of splits:2
2022-09-25 18:11:44,796 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664143420984_0001
2022-09-25 18:11:44,796 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-09-25 18:11:45,005 INFO conf.Configuration: resource-types.xml not found
2022-09-25 18:11:45,006 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-09-25 18:11:45,471 INFO impl.YarnClientImpl: Submitted application application_1664143420984_0001
2022-09-25 18:11:45,569 INFO mapreduce.Job: The url to track the job: http://LAPTOP-8RL4SMOR:8088/proxy/application_1664143420984_0001/
2022-09-25 18:11:45,571 INFO mapreduce.Job: Running job: job_1664143420984_0001
2022-09-25 18:11:55,817 INFO mapreduce.Job: Job job_1664143420984_0001 running in uber mode : false
2022-09-25 18:11:55,818 INFO mapreduce.Job:  map 0% reduce 0%
2022-09-25 18:12:02,015 INFO mapreduce.Job:  map 100% reduce 0%
2022-09-25 18:12:08,102 INFO mapreduce.Job:  map 100% reduce 100%
2022-09-25 18:12:08,107 INFO mapreduce.Job: Job job_1664143420984_0001 completed successfully
2022-09-25 18:12:08,218 INFO mapreduce.Job: Counters: 50
```

3. Output:



```
        File Input Format Counters
                Bytes Read=2555806
        File Output Format Counters
                Bytes Written=18
Success
```

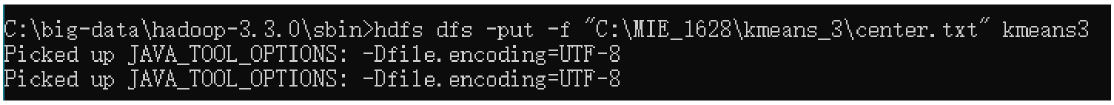
The output files is located in its relative folder.

"Assignment1_Li_Wenxin_1007508724\linecount\output"

Q2. KMeans of k=3 & k=6

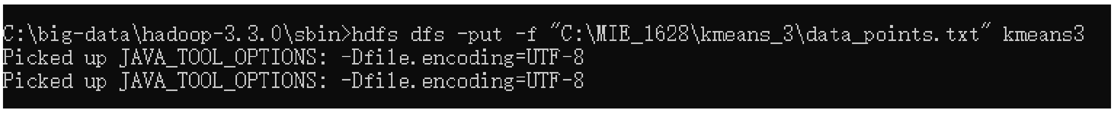
- K=3

1. Upload initial center.txt to hdfs



```
C:\big-data\hadoop-3.3.0\sbin>hdfs dfs -put -f "C:\MIE_1628\kmeans_3\center.txt" kmeans3
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
```

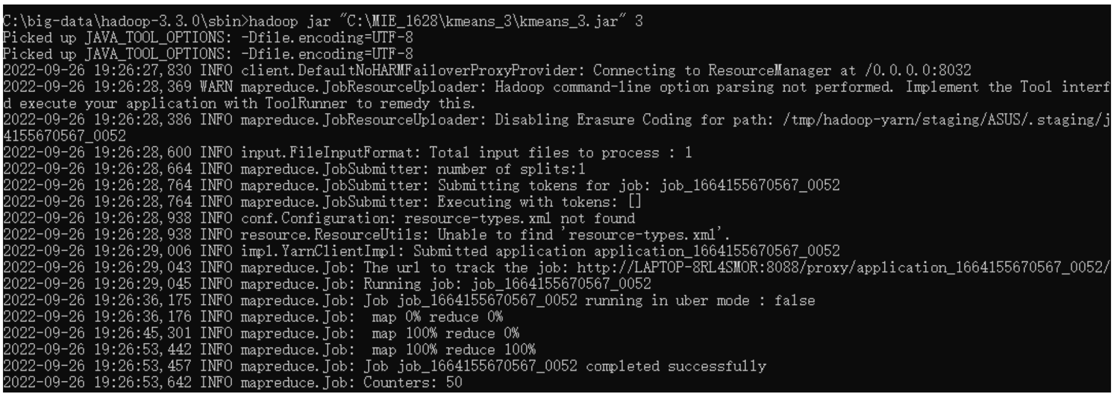
2. Upload data_points.txt to hdfs



```
C:\big-data\hadoop-3.3.0\sbin>hdfs dfs -put -f "C:\MIE_1628\kmeans_3\data_points.txt" kmeans3
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
```

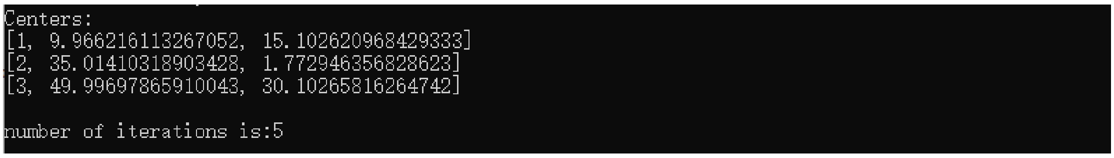
3. Input: hadoop jar <Path of jar> <k value>

hadoop jar "C:\MIE_1628\kmeans_3\kmeans_3.jar" 3



```
C:\big-data\hadoop-3.3.0\sbin>hadoop jar ″C:\MIE_1628\kmeans_3\kmeans_3.jar″ 3
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
2022-09-26 19:26:27,830 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-26 19:26:28,369 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interf
d execute your application with ToolRunner to remedy this.
2022-09-26 19:26:28,386 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ASUS/.staging/j
4155670567_0052
2022-09-26 19:26:28,600 INFO input.FileInputFormat: Total input files to process : 1
2022-09-26 19:26:28,664 INFO mapreduce.JobSubmitter: number of splits:1
2022-09-26 19:26:28,764 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664155670567_0052
2022-09-26 19:26:28,764 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-09-26 19:26:28,938 INFO conf.Configuration: resource-types.xml not found
2022-09-26 19:26:28,938 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-09-26 19:26:29,006 INFO impl.YarnClientImpl: Submitted application application_1664155670567_0052
2022-09-26 19:26:29,043 INFO mapreduce.Job: The url to track the job: http://LAPTOP-8RL4SMOR:8088/proxy/application_1664155670567_0052/
2022-09-26 19:26:29,045 INFO mapreduce.Job: Running job: job_1664155670567_0052
2022-09-26 19:26:36,175 INFO mapreduce.Job: Job job_1664155670567_0052 running in uber mode : false
2022-09-26 19:26:36,176 INFO mapreduce.Job:  map 0% reduce 0%
2022-09-26 19:26:45,301 INFO mapreduce.Job:  map 100% reduce 0%
2022-09-26 19:26:53,442 INFO mapreduce.Job:  map 100% reduce 100%
2022-09-26 19:26:53,457 INFO mapreduce.Job: Job job_1664155670567_0052 completed successfully
2022-09-26 19:26:53,642 INFO mapreduce.Job: Counters: 50
```

4. Output: Centroids, and number of iterations

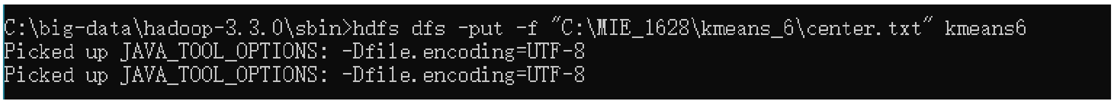


```
Centers:
[1, 9.966216113267052, 15.102620968429333]
[2, 35.01410318903428, 1.772946356828623]
[3, 49.99697865910043, 30.10265816264742]

number of iterations is:5
```

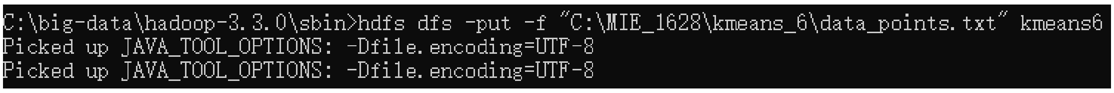
- K=6

1. Upload initial center.txt to hdfs



```
C:\big-data\hadoop-3.3.0\sbin>hdfs dfs -put -f ″C:\MIE_1628\kmeans_6\center.txt″ kmeans6
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
```

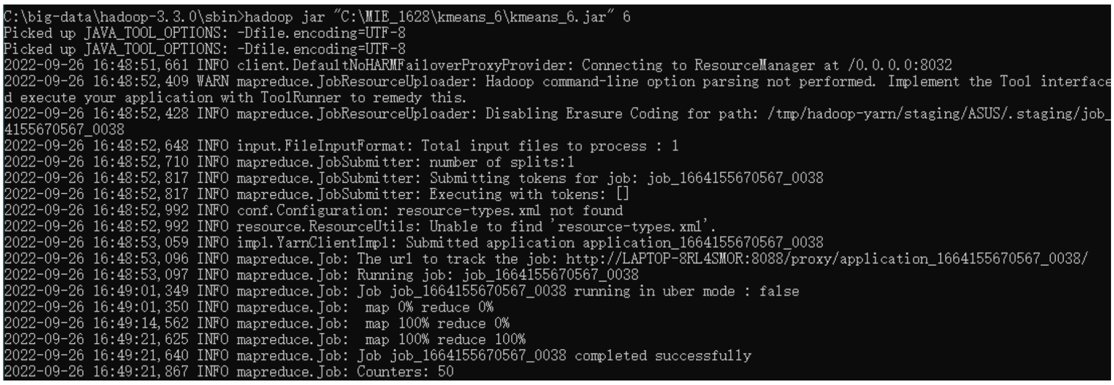
2. Upload data_points.txt to hdfs



```
C:\big-data\hadoop-3.3.0\sbin>hdfs dfs -put -f ″C:\MIE_1628\kmeans_6\data_points.txt″ kmeans6
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
```

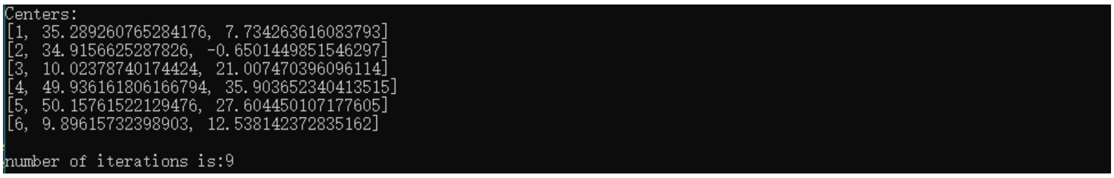
3. Input: hadoop jar <Path of jar> <k value>

hadoop jar "C:\MIE_1628\kmeans_6\kmeans_6.jar" 6



```
C:\big-data\hadoop-3.3.0\sbin>hadoop jar ″C:\MIE_1628\kmeans_6\kmeans_6.jar″ 6
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
Picked up JAVA_TOOL_OPTIONS: -Dfile.encoding=UTF-8
2022-09-26 16:48:51,661 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-09-26 16:48:52,409 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface
d execute your application with ToolRunner to remedy this.
2022-09-26 16:48:52,428 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ASUS/.staging/job_
4155670567_0038
2022-09-26 16:48:52,648 INFO input.FileInputFormat: Total input files to process : 1
2022-09-26 16:48:52,710 INFO mapreduce.JobSubmitter: number of splits:1
2022-09-26 16:48:52,817 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664155670567_0038
2022-09-26 16:48:52,817 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-09-26 16:48:52,992 INFO conf.Configuration: resource-types.xml not found
2022-09-26 16:48:52,992 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-09-26 16:48:53,059 INFO impl.YarnClientImpl: Submitted application application_1664155670567_0038
2022-09-26 16:48:53,096 INFO mapreduce.Job: The url to track the job: http://LAPTOP-8RL4SMOR:8088/proxy/application_1664155670567_0038/
2022-09-26 16:48:53,097 INFO mapreduce.Job: Running job: job_1664155670567_0038
2022-09-26 16:49:01,349 INFO mapreduce.Job: Job job_1664155670567_0038 running in uber mode : false
2022-09-26 16:49:01,350 INFO mapreduce.Job:  map 0% reduce 0%
2022-09-26 16:49:14,562 INFO mapreduce.Job:  map 100% reduce 0%
2022-09-26 16:49:21,625 INFO mapreduce.Job:  map 100% reduce 100%
2022-09-26 16:49:21,640 INFO mapreduce.Job: Job job_1664155670567_0038 completed successfully
2022-09-26 16:49:21,867 INFO mapreduce.Job: Counters: 50
```

4. Output: Centroids and number of iterations

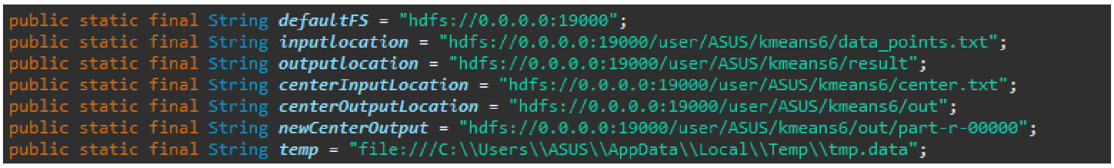


```
Centers:
[1, 35.289260765284176, 7.734263616083793]
[2, 34.9156625287826, -0.6501449851546297]
[3, 10.02378740174424, 21.007470396096114]
[4, 49.936161806166794, 35.903652340413515]
[5, 50.15761522129476, 27.604450107177605]
[6, 9.89615732398903, 12.538142372835162]

number of iterations is:9
```

- Setting:

1. The input path and output path are already written in the java file. Thus, do not need to be entered manually.



```
public static final String defaultFS = "hdfs://0.0.0.0:19000";
public static final String inputlocation = "hdfs://0.0.0.0:19000/user/ASUS/kmeans6/data_points.txt";
public static final String outputlocation = "hdfs://0.0.0.0:19000/user/ASUS/kmeans6/result";
public static final String centerInputLocation = "hdfs://0.0.0.0:19000/user/ASUS/kmeans6/center.txt";
public static final String centerOutputLocation = "hdfs://0.0.0.0:19000/user/ASUS/kmeans6/out";
public static final String newCenterOutput = "hdfs://0.0.0.0:19000/user/ASUS/kmeans6/out/part-r-00000";
public static final String temp = "file:///C:\\Users\\ASUS\\AppData\\Local\\Temp\\tmp.data";
```

2. The output files is located in its relative folder.

"Assignment1_Li_Wenxin_1007508724\kmeans_3\output"
"Assignment1_Li_Wenxin_1007508724\kmeans_6\output"

Q3. Explain advantages and disadvantages of using K-Means Clustering with MapReduce.
Answer:

- Advantage:

We can use MapReduce's distributed feature, and the process of computing center can be parallelized. When the amount of data is huge, MapReduce has advantages over unit price.

- Disadvantages:

K-means is an algorithm that needs multiple iterations, that is, the output of this calculation is used as the input of the next calculation. In MapReduce, each time the data needs to be written to HDFS, and then read, which takes some time. In the case of a small amount of data, the performance is not better than the unit price.

Q4. Can we reduce the number of distance comparison by applying the Canopy Selection? Which distance metric should we use for the canopy clustering and why?

Answer: Yes, the key idea of the canopy algorithm is that one can greatly reduce the number of distance computations required for clustering by first cheaply partitioning the data into overlapping subsets, and then only measuring distances among pairs of data points that belong to a common subset.
The distance metric like cosine-similarity which based on inverted index is best for the canopy clustering. It is very cheap, and it can also be applied to text and high dimensional real-valued data.

Q5. Is it possible to apply Canopy Selection on MapReduce? If yes, then explain in words, how would you implement it?

Answer: Yes, it is possible to apply Canopy Selection on MapReduce. In order to achieve this, the processed data must first be milled into a suitable format. Mapper performs canopy clustering on the points in its input set by an inverted index and two thresholds(T1>T2). It then outputs the centers of its canopies. Reducer clusters the canopy centers to produce the final canopy centers.

Q6. Is it possible to combine the Canopy Selection with K-Means on MapReduce? If yes, then explain in words, how would you do that?

Answer: Yes, it is possible to combine the Canopy Selection with K-Means on MapReduce. Canopy

Selection can be used as the initial step of K-Means clustering. In MapReduce, we can get the final canopy centers(details is in Q5) by Canopy Selection. Then, we use the canopy centers as the initial centroids of K-Means. After that, we can use the parallel K-Means algorism on MapReduce to get the more accuracy centroids.