# Distributed System Project Assignment 3

Name: Vineet Kurapati
Student ID: 1002154505

Name: Sharini Jayabal
Student ID: 1002087946

→ Introduction

1. Cloning the GitHub Directory

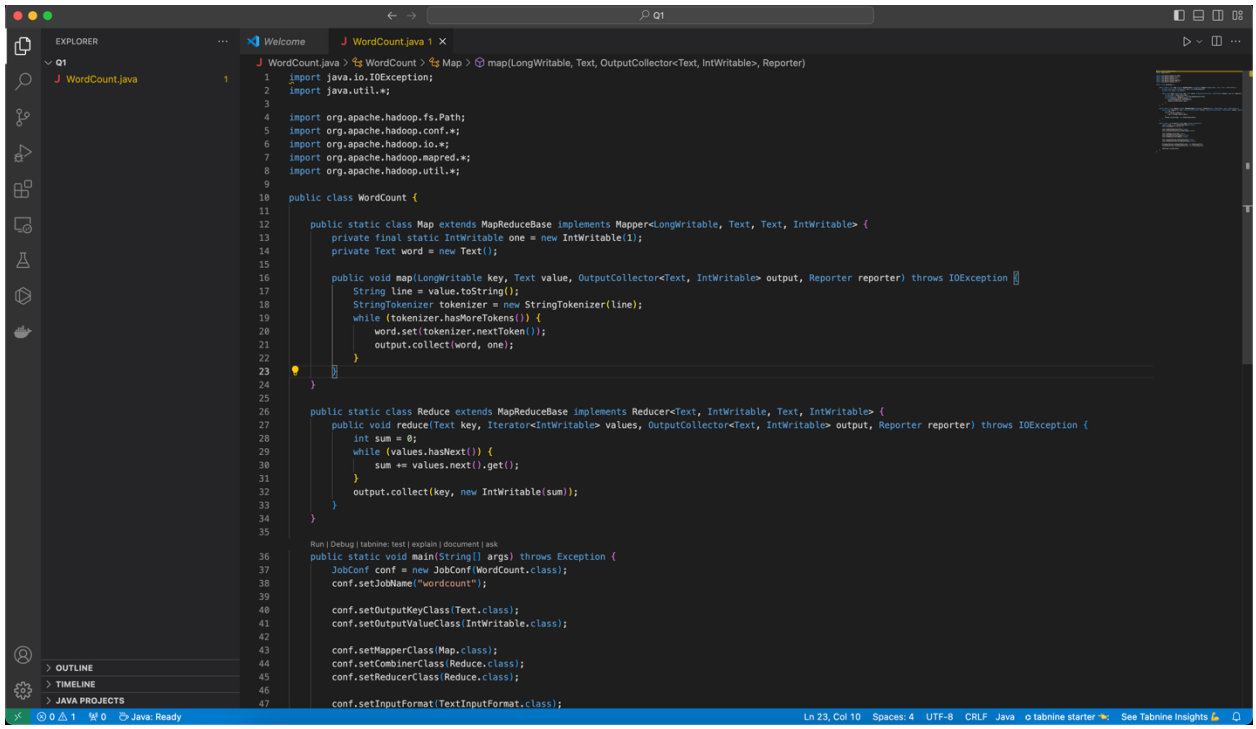
```
Desktop — -zsh — 80×24
vineetkurapati@Vineets-MacBook-Air Desktop % date;whoami;
Mon Apr 15 22:52:35 CDT 2024
vineetkurapati
vineetkurapati@Vineets-MacBook-Air Desktop % git clone https://github.com/big-da
ta-europe/docker-hadoop
Cloning into 'docker-hadoop'...
remote: Enumerating objects: 539, done.
remote: Total 539 (delta 0), reused 0 (delta 0), pack-reused 539
Receiving objects: 100% (539/539), 108.43 KiB | 902.00 KiB/s, done.
Resolving deltas: 100% (249/249), done.
```

2. Using docker compose up to start the docker environment

```
docker-hadoop — -zsh — 80×24
Mon Apr 15 22:54:17 CDT 2024
vineetkurapati
vineetkurapati@Vineets-MacBook-Air docker-hadoop % docker-compose up -d
[+] Running 1/0
 ⠿ Network docker-hadoop_default
                                                                        Creat
ed0.1s
 ⠿ Container resourcemanager
                                                                        Creat
ing0.1s
 ⠿ Container datanode
                                                                        Creat
ing0.1s
 ⠿ Container nodemanager
                                                                        Creat
ing0.1s
[+] Running 5/7menode
 ⠿ Network docker-hadoop_default

  Created0.2s istoryserver
 ⠿ Container resourcemanager

  Starting0.2s  requested image's platform (linux/amd64) does not match the dete
 ⠿ Container datanode
```

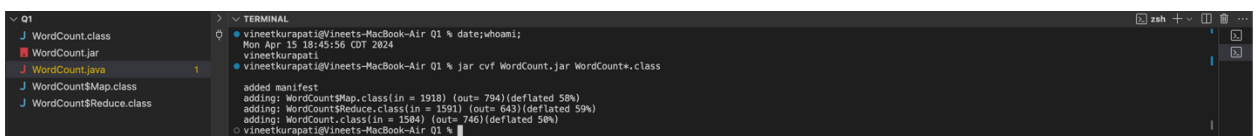
Q1 Word Count

1. Creating the Java file “WordCount.java”



```java
import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

    public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
            }
        }
    }

    public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
        public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
            int sum = 0;
            while (values.hasNext()) {
                sum += values.next().get();
            }
            output.collect(key, new IntWritable(sum));
        }
    }

    public static void main(String[] args) throws Exception {
        JobConf conf = new JobConf(WordCount.class);
        conf.setJobName("wordcount");

        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);

        conf.setMapperClass(Map.class);
        conf.setCombinerClass(Reduce.class);
        conf.setReducerClass(Reduce.class);

        conf.setInputFormat(TextInputFormat.class);
```

2. Compiling the code into a “.jar” file



```
vineetkurapati@Vineets-MacBook-Air Q1 % date;whoami;
Mon Apr 15 18:45:56 CDT 2024
vineetkurapati
vineetkurapati@Vineets-MacBook-Air Q1 % jar cvf WordCount.jar WordCount*.class

added manifest
adding: WordCount$Map.class(in = 1918) (out= 794)(deflated 58%)
adding: WordCount$Reduce.class(in = 1591) (out= 643)(deflated 59%)
adding: WordCount.class(in = 1504) (out= 746)(deflated 50%)
vineetkurapati@Vineets-MacBook-Air Q1 %
```

3. Copying the .jar file and the inputfile to the docker container using “docker cp <filename> namenode:/”

```
vineetkurapati@Vineets-MacBook-Air Q1 % date;whoami;
Mon Apr 15 18:49:20 CDT 2024
vineetkurapati
vineetkurapati@Vineets-MacBook-Air Q1 % docker cp WordCount.jar namenode:/
Successfully copied 4.61kB to namenode:/
vineetkurapati@Vineets-MacBook-Air Q1 % docker cp Q1_Input1.txt namenode:/
Successfully copied 2.05kB to namenode:/
vineetkurapati@Vineets-MacBook-Air Q1 %
```

4. Once the input file has been copied into the docker container we can upload it into the Hadoop File System(hdfs) using “hdfs dfs -put <filename> <inputpath>”

```
root@11fe547620d9:/# ls
CharCount.jar  KMeansClustering.jar  Q2_Input1.txt  WordCount.jar  boot  entrypoint.sh  hadoop       home  lib64  mnt  proc  run     sbin  sys  usr
KEYS           Q1_Input1.txt         Q4_Input1.txt  bin            dev   etc            hadoop-data  lib   media  opt  root  run.sh  srv   tmp  var
root@11fe547620d9:/# date;whoami
Mon Apr 15 23:50:23 UTC 2024
root
root@11fe547620d9:/# hdfs dfs -put Q1_Input1.txt /user/root/input
2024-04-15 23:50:47,737 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@11fe547620d9:/#
```

5. Running the map reduce on the docker env

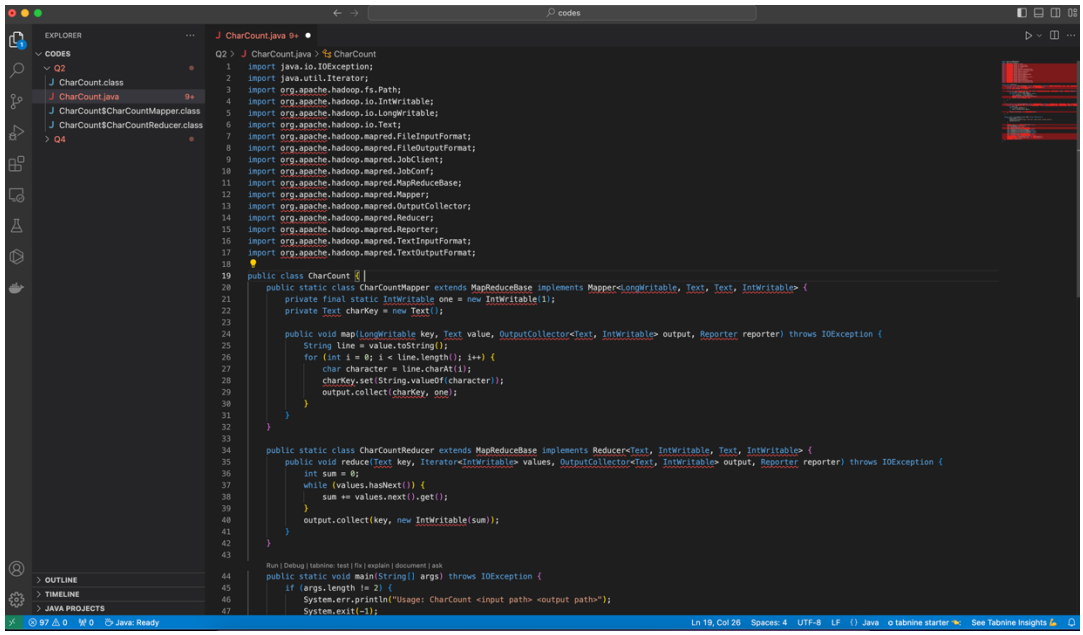
```
TERMINAL                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                
root@11fe547620d9:/# hadoop jar WordCount.jar WordCount /user/root/input/Q1_Input1.txt /user/root/outputQ1
2024-04-15 23:52:08,267 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.6:8032
2024-04-15 23:52:08,710 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.2:10200
2024-04-15 23:52:08,760 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.6:8032
2024-04-15 23:52:08,760 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.2:10200
2024-04-15 23:52:09,082 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRu
nner to remedy this.
2024-04-15 23:52:09,151 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1713222675227_0004
2024-04-15 23:52:09,396 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:52:09,784 INFO mapred.FileInputFormat: Total input files to process : 1
2024-04-15 23:52:09,854 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:52:10,337 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:52:10,794 INFO mapreduce.JobSubmitter: number of splits:2
2024-04-15 23:52:11,102 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:52:11,215 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713222675227_0004
2024-04-15 23:52:11,215 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-15 23:52:11,520 INFO conf.Configuration: resource-types.xml not found
2024-04-15 23:52:11,520 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-15 23:52:12,324 INFO impl.YarnClientImpl: Submitted application application_1713222675227_0004
2024-04-15 23:52:12,430 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1713222675227_0004/
2024-04-15 23:52:12,433 INFO mapreduce.Job: Running job: job_1713222675227_0004
2024-04-15 23:52:21,858 INFO mapreduce.Job: Job job_1713222675227_0004 running in uber mode : false
2024-04-15 23:52:21,869 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-15 23:52:30,302 INFO mapreduce.Job:  map 50% reduce 0%
2024-04-15 23:52:31,485 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-15 23:52:38,718 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-15 23:52:41,881 INFO mapreduce.Job: Job job_1713222675227_0004 completed successfully
```

6. Checking the output using “hds dfs -cat <output path>

```
root@11fe547620d9:/# date;whoami;
Mon Apr 15 23:53:39 UTC 2024
root
root@11fe547620d9:/# hdfs dfs -cat /user/root/outputQ1/part-00000
2024-04-15 23:54:11,935 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Birdsong        1
Dreams  1
In      3
Nature's        1
Sunrise 1
Whispers        1
anew,   1
awaken, 1
beauty  1
breeze, 1
dawn's  1
each    1
ease.   1
embrace,        1
floats  1
found,  1
gentle  2
grace.  1
hue.    1
in      1
moment, 1
morning's       1
on      1
paints  1
quiet   1
sky     1
soft    1
sound.  1
sweet,  1
symphony        1
the     3
tranquil        1
with    1
world's 1
root@11fe547620d9:/#
```

Q2 Character Count

1. Creating the Java file “CharCount.java”




2. Compiling the code into a “.jar” file

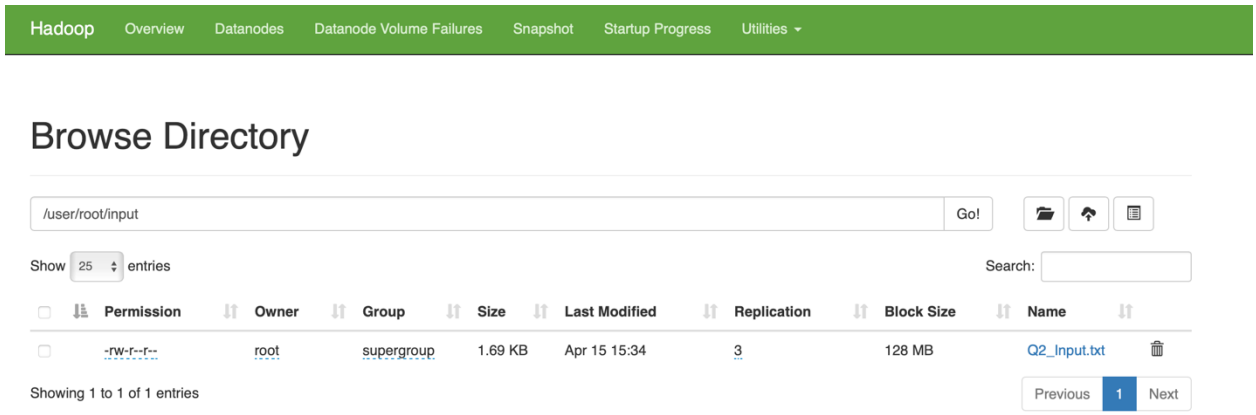
```
vineetkurapati@Vineets-MacBook-Air Q2 % javac -classpath /Users/vineetkurapati/Downloads/hadoop-3.4.0/share/hadoop/common/hadoop-common-3.4.0.jar:/Users/vineetkurapati/Downloads/hadoop-3
.4.0/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.4.0.jar CharCount.java

vineetkurapati@Vineets-MacBook-Air Q2 % jar cvf CharCount.jar CharCount*.class

added manifest
adding: CharCount$CharCountMapper.class(in = 1955) (out= 803)(deflated 58%)
adding: CharCount$CharCountReducer.class(in = 1611) (out= 642)(deflated 60%)
adding: CharCount.class(in = 1771) (out= 891)(deflated 49%)
vineetkurapati@Vineets-MacBook-Air Q2 % date;whoami;
Mon Apr 15 15:22:28 CDT 2024
vineetkurapati
```

3. Copying the .jar file and the inputfile to the docker container using “docker cp <filename> namenode:/”

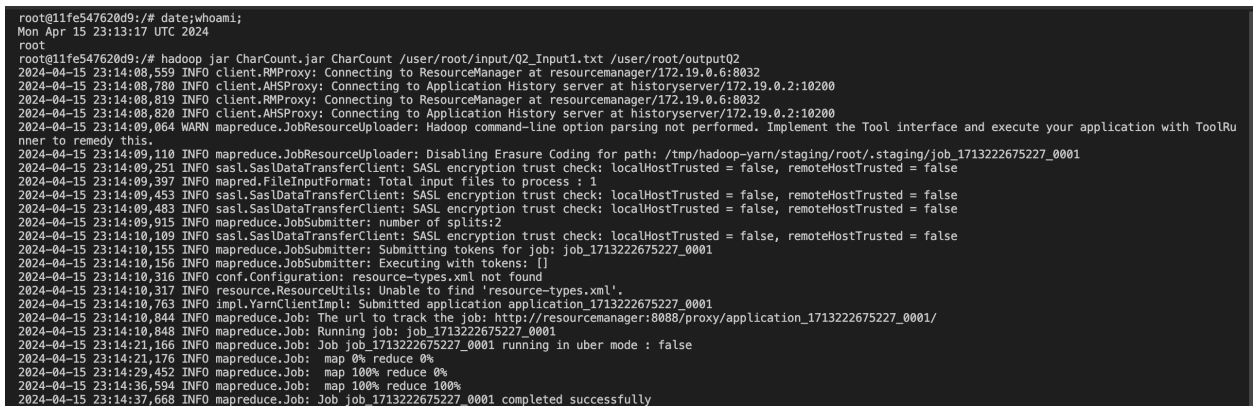
```
vineetkurapati@Vineets-MacBook-Air codes % docker exec -it namenode /bin/bash
root@1a194fa4925e:/# date;whoami;
Mon Apr 15 20:27:49 UTC 2024
root
root@1a194fa4925e:/# ls
CharCount.jar  KMeansClustering.jar  boot  entrypoint.sh  hadoop       home       input.txt  lib64  mnt  output.txt  root  run.sh  small.txt  sys  usr
KEYS           bin                   dev   etc            hadoop-data  inpu2.txt  lib        media  opt  proc        run   sbin    srv        tmp  var
root@1a194fa4925e:/#
```

4. Once the input file has been copied into the docker container we can upload it into the Hadoop File System(hdfs) using “hdfs dfs -put <filename> <inputpath>”

```
TERMINAL                                                                                                                                                    docker

root@1a194fa4925e:/# date;whoami;
Mon Apr 15 20:34:17 UTC 2024
root
root@1a194fa4925e:/# hdfs dfs -put Q2_Input.txt /user/root/input
2024-04-15 20:34:22,754 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@1a194fa4925e:/#
```

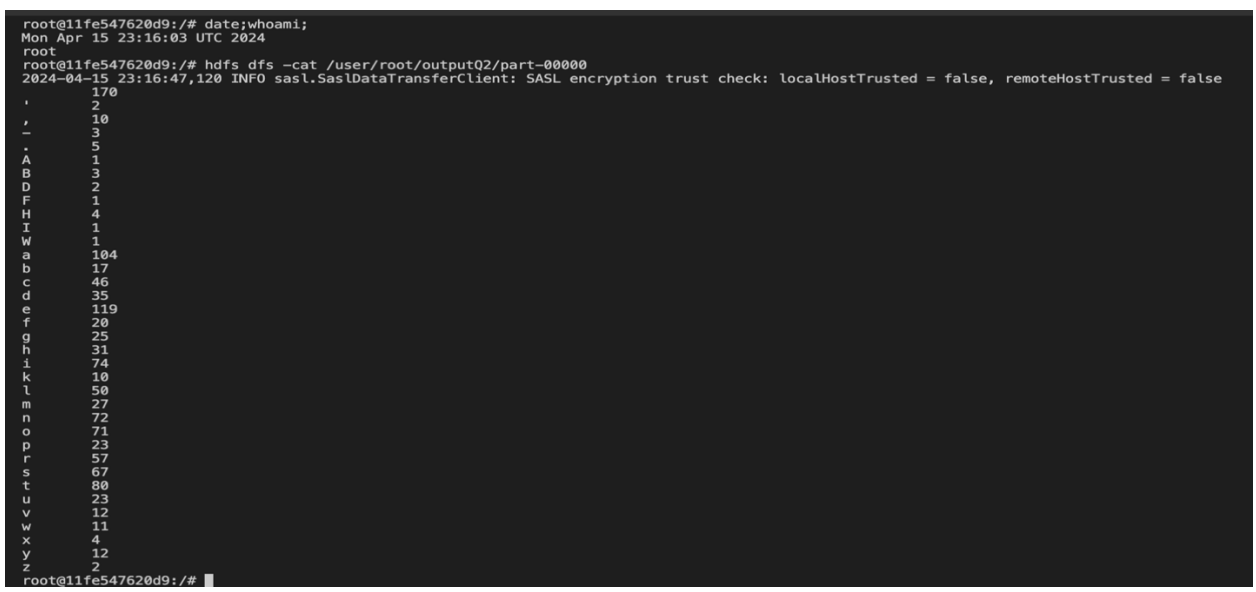



5. Running the map reduce on the docker env



```
root@11fe547620d9:/# date;whoami;
Mon Apr 15 23:13:17 UTC 2024
root
root@11fe547620d9:/# hadoop jar CharCount.jar CharCount /user/root/input/Q2_Input1.txt /user/root/outputQ2
2024-04-15 23:14:08,559 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.6:8032
2024-04-15 23:14:08,780 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.2:10200
2024-04-15 23:14:08,819 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.6:8032
2024-04-15 23:14:08,820 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.2:10200
2024-04-15 23:14:09,064 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRu
nner to remedy this.
2024-04-15 23:14:09,110 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1713222675227_0001
2024-04-15 23:14:09,251 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:14:09,397 INFO mapred.FileInputFormat: Total input files to process : 1
2024-04-15 23:14:09,453 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:14:09,483 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:14:09,915 INFO mapreduce.JobSubmitter: number of splits:2
2024-04-15 23:14:10,109 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:14:10,155 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713222675227_0001
2024-04-15 23:14:10,156 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-15 23:14:10,316 INFO conf.Configuration: resource-types.xml not found
2024-04-15 23:14:10,317 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-15 23:14:10,763 INFO impl.YarnClientImpl: Submitted application application_1713222675227_0001
2024-04-15 23:14:10,844 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1713222675227_0001/
2024-04-15 23:14:10,848 INFO mapreduce.Job: Running job: job_1713222675227_0001
2024-04-15 23:14:21,166 INFO mapreduce.Job: Job job_1713222675227_0001 running in uber mode : false
2024-04-15 23:14:21,176 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-15 23:14:29,452 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-15 23:14:36,594 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-15 23:14:37,668 INFO mapreduce.Job: Job job_1713222675227_0001 completed successfully
```

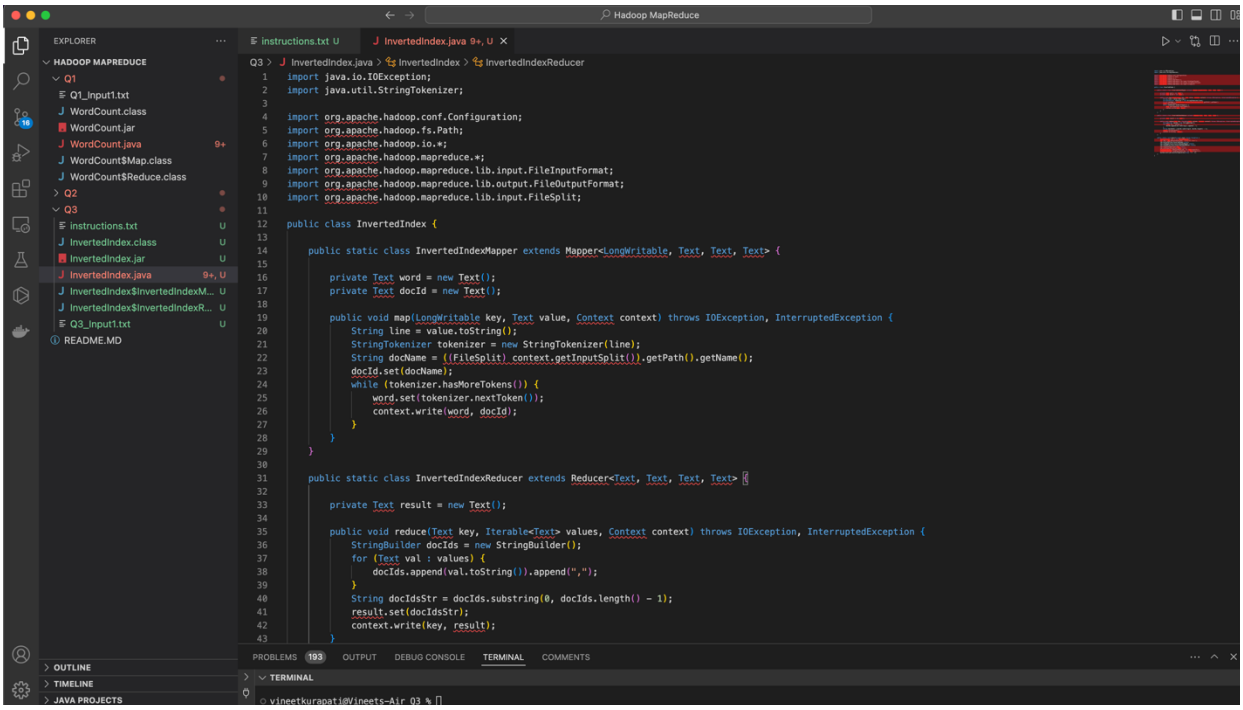
6. Checking the output using “hds dfs -cat <output path>



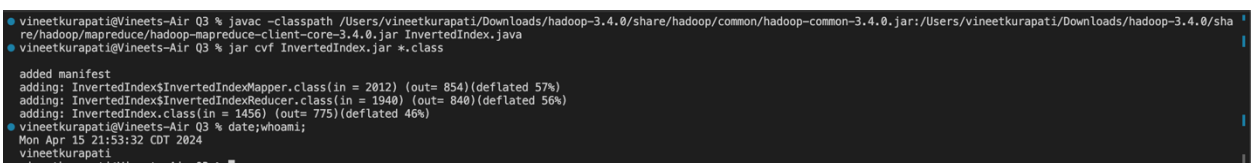
```
root@11fe547620d9:/# date;whoami;
Mon Apr 15 23:16:03 UTC 2024
root
root@11fe547620d9:/# hdfs dfs -cat /user/root/outputQ2/part-00000
2024-04-15 23:16:47,120 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
'       170
,       2
-       10
.       3
A       5
B       1
D       3
F       2
H       1
I       4
W       1
a       1
b       104
c       17
d       46
e       35
f       119
g       20
h       25
i       31
k       74
l       10
m       50
n       27
o       72
p       71
r       23
s       57
t       67
u       80
v       23
w       12
x       11
y       4
z       12
root@11fe547620d9:/#
```

Q3 Inverted Index

1. Creating the Java file “InvertedIndex.java”




2. Compiling the code into a “.jar” file




3. Copying the .jar file and the inputfile to the docker container using “docker cp <filename> namenode:/”

4. Once the input file has been copied into the docker container we can upload it into the Hadoop File System(hdfs) using “hdfs dfs -put <filename> <inputpath>”

```
root@dae5d126041a:/# date;whoami;
Tue Apr 16 03:03:07 UTC 2024
root
root@dae5d126041a:/# hdfs dfs -put input_one_q3.txt input_two_q3.txt input_three_q3.txt input_four_q3.txt input_five_q3.txt /user/root/input

2024-04-16 03:03:10,924 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:03:11,113 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:03:11,605 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:03:12,049 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:03:12,523 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@dae5d126041a:/#
root@dae5d126041a:/#
```

5. Running the map reduce on the docker env (Running all the testcases at once to check the processing of all the inputs)

```
root@dae5d126041a:/# date;whoami;
Tue Apr 16 03:04:23 UTC 2024
root
root@dae5d126041a:/# hadoop jar InvertedIndex.jar InvertedIndex /user/root/input /user/root/outputQ3
2024-04-16 03:05:26,660 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.5:8032
2024-04-16 03:05:26,879 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.4:10200
2024-04-16 03:05:27,117 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRu
nner to remedy this.
2024-04-16 03:05:27,217 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1713235919592_0001
2024-04-16 03:05:27,347 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:05:27,488 INFO input.FileInputFormat: Total input files to process : 5
2024-04-16 03:05:27,523 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:05:27,984 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:05:28,424 INFO mapreduce.JobSubmitter: number of splits:5
2024-04-16 03:05:28,607 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-16 03:05:29,058 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713235919592_0001
2024-04-16 03:05:29,059 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-16 03:05:29,291 INFO conf.Configuration: resource-types.xml not found
2024-04-16 03:05:29,292 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-16 03:05:30,282 INFO impl.YarnClientImpl: Submitted application application_1713235919592_0001
2024-04-16 03:05:30,379 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1713235919592_0001/
2024-04-16 03:05:30,380 INFO mapreduce.Job: Running job: job_1713235919592_0001
2024-04-16 03:05:41,836 INFO mapreduce.Job: Job job_1713235919592_0001 running in uber mode : false
2024-04-16 03:05:41,848 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-16 03:05:51,214 INFO mapreduce.Job:  map 40% reduce 0%
2024-04-16 03:05:52,245 INFO mapreduce.Job:  map 60% reduce 0%
2024-04-16 03:05:57,344 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-16 03:06:02,441 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-16 03:06:05,570 INFO mapreduce.Job: Job job_1713235919592_0001 completed successfully
```

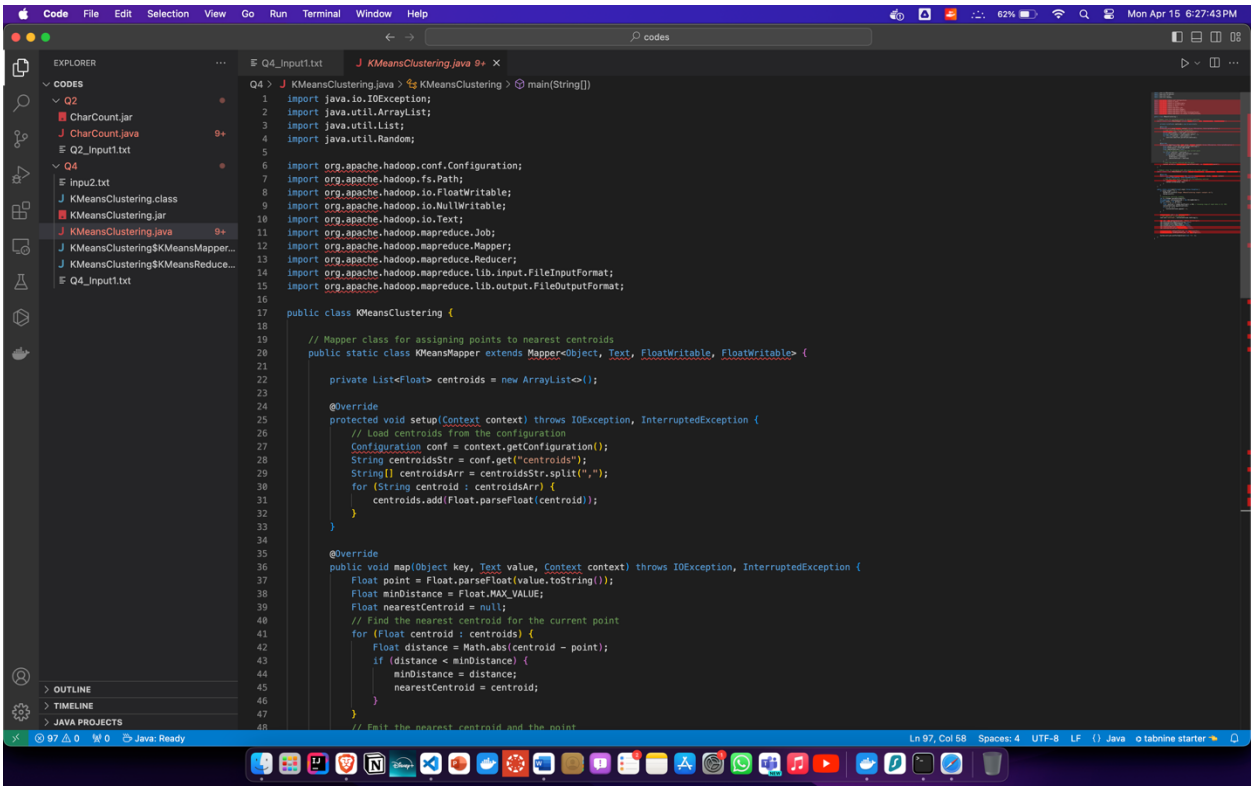
6. Checking the output using “hds dfs -cat <output path>

```
root@dae5d126041a:/# hdfs dfs -cat /user/root/outputQ3/part-r-00000
2024-04-16 03:07:50,869 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1:      input_one_q3.txt,input_two_q3.txt,input_four_q3.txt,input_five_q3.txt,input_three_q3.txt
2:      input_three_q3.txt,input_four_q3.txt,input_one_q3.txt,input_two_q3.txt,input_five_q3.txt
Adventure       input_four_q3.txt
Alice   input_two_q3.txt
Birds   input_one_q3.txt
Bob     input_two_q3.txt
Document        input_two_q3.txt,input_two_q3.txt,input_one_q3.txt,input_one_q3.txt,input_three_q3.txt,input_three_q3.txt,input_five_q3.txt,input_five_q3.txt,input_four_q3.txt,inp
ut_four_q3.txt
Innovation      input_five_q3.txt
It's    input_three_q3.txt
People  input_three_q3.txt
Technology      input_five_q3.txt
The     input_one_q3.txt,input_four_q3.txt
advancing       input_five_q3.txt
always  input_four_q3.txt
and     input_two_q3.txt,input_two_q3.txt,input_three_q3.txt
are     input_three_q3.txt,input_four_q3.txt,input_one_q3.txt
be      input_four_q3.txt
books   input_two_q3.txt
brightly        input_one_q3.txt
cats    input_three_q3.txt
changing        input_five_q3.txt
chirping        input_one_q3.txt
dogs    input_three_q3.txt
drives  input_five_q3.txt
enjoys  input_two_q3.txt
experiences.    input_four_q3.txt
explored.       input_four_q3.txt
fields, input_five_q3.txt
find    input_three_q3.txt
from    input_three_q3.txt
full    input_four_q3.txt
future. input_five_q3.txt
happily input_one_q3.txt
in      input_one_q3.txt,input_one_q3.txt,input_five_q3.txt
is      input_five_q3.txt,input_one_q3.txt,input_four_q3.txt
likes   input_two_q3.txt
listen  input_two_q3.txt
live.   input_five_q3.txt
movies. input_two_q3.txt
music.  input_two_q3.txt
new     input_four_q3.txt
of      input_four_q3.txt
our     input_five_q3.txt
outside.        input_three_q3.txt
playing input_two_q3.txt
progress        input_five_q3.txt
raining input_three_q3.txt
rapidly,        input_five_q3.txt
read    input_two_q3.txt
rushing input_three_q3.txt
seekers input_four_q3.txt
seeking input_four_q3.txt
shaping input_five_q3.txt
shelter input_three_q3.txt
shining input_one_q3.txt
sky.    input_one_q3.txt
sports  input_two_q3.txt
storm.  input_three_q3.txt
```

```
People  input_three_q3.txt
Technology      input_five_q3.txt
The     input_one_q3.txt,input_four_q3.txt
advancing       input_five_q3.txt
always  input_four_q3.txt
and     input_two_q3.txt,input_two_q3.txt,input_three_q3.txt
are     input_three_q3.txt,input_four_q3.txt,input_one_q3.txt
be      input_four_q3.txt
books   input_two_q3.txt
brightly        input_one_q3.txt
cats    input_three_q3.txt
changing        input_five_q3.txt
chirping        input_one_q3.txt
dogs    input_three_q3.txt
drives  input_five_q3.txt
enjoys  input_two_q3.txt
experiences.    input_four_q3.txt
explored.       input_four_q3.txt
fields, input_five_q3.txt
find    input_three_q3.txt
from    input_three_q3.txt
full    input_four_q3.txt
future. input_five_q3.txt
happily input_one_q3.txt
in      input_one_q3.txt,input_one_q3.txt,input_five_q3.txt
is      input_five_q3.txt,input_one_q3.txt,input_four_q3.txt
likes   input_two_q3.txt
listen  input_two_q3.txt
live.   input_five_q3.txt
movies. input_two_q3.txt
music.  input_two_q3.txt
new     input_four_q3.txt
of      input_four_q3.txt
our     input_five_q3.txt
outside.        input_three_q3.txt
playing input_two_q3.txt
progress        input_five_q3.txt
raining input_three_q3.txt
rapidly,        input_five_q3.txt
read    input_two_q3.txt
rushing input_three_q3.txt
seekers input_four_q3.txt
seeking input_four_q3.txt
shaping input_five_q3.txt
shelter input_three_q3.txt
shining input_one_q3.txt
sky.    input_one_q3.txt
sports  input_two_q3.txt
storm.  input_three_q3.txt
sun     input_one_q3.txt
the     input_one_q3.txt,input_one_q3.txt,input_three_q3.txt,input_five_q3.txt
to      input_three_q3.txt,input_two_q3.txt,input_two_q3.txt,input_four_q3.txt
trees.  input_one_q3.txt
various input_five_q3.txt
waiting input_four_q3.txt
watching        input_two_q3.txt
way     input_five_q3.txt
we      input_five_q3.txt
wonders input_four_q3.txt
world   input_four_q3.txt
root@dae5d126041a:/#
```

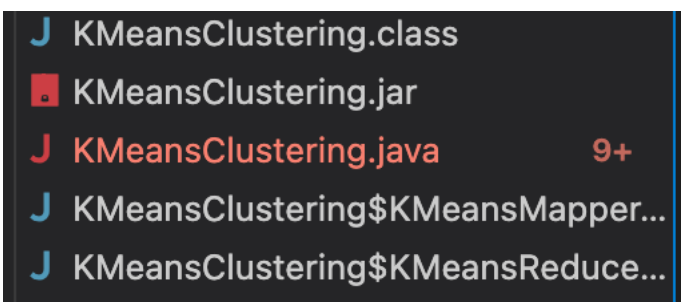
Q4 KMeans Clustering

1. Creating the Java file “KMeansClustering.java”



```java
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import java.util.Random;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class KMeansClustering {

    // Mapper class for assigning points to nearest centroids
    public static class KMeansMapper extends Mapper<Object, Text, FloatWritable, FloatWritable> {

        private List<Float> centroids = new ArrayList<>();

        @Override
        protected void setup(Context context) throws IOException, InterruptedException {
            // Load centroids from the configuration
            Configuration conf = context.getConfiguration();
            String centroidsStr = conf.get("centroids");
            String[] centroidsArr = centroidsStr.split(",");
            for (String centroid : centroidsArr) {
                centroids.add(Float.parseFloat(centroid));
            }
        }

        @Override
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            Float point = Float.parseFloat(value.toString());
            Float minDistance = Float.MAX_VALUE;
            Float nearestCentroid = null;
            // Find the nearest centroid for the current point
            for (Float centroid : centroids) {
                Float distance = Math.abs(centroid - point);
                if (distance < minDistance) {
                    minDistance = distance;
                    nearestCentroid = centroid;
                }
            }
            // Emit the nearest centroid and the point
```

2. Compiling the code into a “.jar” file




3. Copying the .jar file and the inputfile to the docker container using “docker cp <filename> namenode:/”

```
vineetkurapati@Vineets-MacBook-Air Q4 % date;whoami;
Mon Apr 15 18:25:43 CDT 2024
vineetkurapati
vineetkurapati@Vineets-MacBook-Air Q4 % docker cp KMeansClustering.jar namenode:/
Successfully copied 5.63kB to namenode:/
vineetkurapati@Vineets-MacBook-Air Q4 % docker cp Q4_Input1.txt namenode:/
Successfully copied 4.1kB to namenode:/
vineetkurapati@Vineets-MacBook-Air Q4 %
```

4. Once the input file has been copied into the docker container we can upload it into the Hadoop File System(hdfs) using “hdfs dfs -put <filename> <inputpath>”

```
root@11fe547620d9:/# date;whoami;
Mon Apr 15 23:32:40 UTC 2024
root
root@11fe547620d9:/# ls
CharCount.jar  KMeansClustering.jar  Q4_Input1.txt  boot  entrypoint.sh  hadoop       home  lib64  mnt  proc  run     sbin  sys  usr
KEYS           Q2_Input1.txt         bin            dev   etc            hadoop-data  lib   media  opt  root  run.sh  srv   tmp  var
root@11fe547620d9:/# hdfs dfs -put Q4_Input1.txt /user/root/input
2024-04-15 23:33:03,539 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@11fe547620d9:/#
```

5. Running the map reduce on the docker env

```
root@11fe547620d9:/# date;whoami;
Mon Apr 15 23:33:50 UTC 2024
root
root@11fe547620d9:/# hadoop jar KMeansClustering.jar KMeansClustering /user/root/input/Q4_Input1.txt /user/root/outputQ4
Usage: KMeansClustering <input> <output> <k>
root@11fe547620d9:/# hadoop jar KMeansClustering.jar KMeansClustering /user/root/input/Q4_Input1.txt /user/root/outputQ4 10
2024-04-15 23:34:54,315 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.19.0.6:8032
2024-04-15 23:34:54,563 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.19.0.2:10200
2024-04-15 23:34:54,782 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-15 23:34:54,825 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1713222675227_0003
2024-04-15 23:34:54,954 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:34:55,148 INFO input.FileInputFormat: Total input files to process : 1
2024-04-15 23:34:55,217 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:34:55,725 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:34:56,162 INFO mapreduce.JobSubmitter: number of splits:1
2024-04-15 23:34:56,407 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-04-15 23:34:56,896 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1713222675227_0003
2024-04-15 23:34:56,898 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-15 23:34:57,144 INFO conf.Configuration: resource-types.xml not found
2024-04-15 23:34:57,145 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-15 23:34:57,913 INFO impl.YarnClientImpl: Submitted application application_1713222675227_0003
2024-04-15 23:34:58,083 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1713222675227_0003/
2024-04-15 23:34:58,085 INFO mapreduce.Job: Running job: job_1713222675227_0003
2024-04-15 23:35:06,440 INFO mapreduce.Job: Job job_1713222675227_0003 running in uber mode : false
2024-04-15 23:35:06,450 INFO mapreduce.Job:  map 0% reduce 0%
2024-04-15 23:35:13,715 INFO mapreduce.Job:  map 100% reduce 0%
2024-04-15 23:35:19,812 INFO mapreduce.Job:  map 100% reduce 100%
2024-04-15 23:35:22,913 INFO mapreduce.Job: Job job_1713222675227_0003 completed successfully
```

6. Checking the output using “hds dfs -cat <output path>

```
root@11fe547620d9:/# hdfs dfs -cat /user/root/outputQ4/part-r-00000
2024-04-15 23:36:55,706 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1.234   9.44916
9.1011  9.44916
5.678   9.44916
13.1415 13.097822
21.1819 21.082932
17.1617 21.082932
25.2021 21.818167
33.2425 31.1153
29.2223 31.1153
45.3031 31.1153
41.2829 31.1153
37.2627 31.1153
65.4041 67.08318
61.3839 67.08318
57.3637 67.08318
53.3435 67.08318
49.3233 67.08318
73.4445 68.83744
69.4243 68.83744
81.4849 80.79666
77.4647 80.79666
89.5253 82.33521
85.5051 82.33521
955.8683        97.48579
951.8481        97.48579
947.8279        97.48579
943.8077        97.48579
939.7875        97.48579
935.7673        97.48579
931.7471        97.48579
927.7269        97.48579
923.7067        97.48579
919.6865        97.48579
915.6663        97.48579
911.6461        97.48579
907.6259        97.48579
903.6057        97.48579
899.5855        97.48579
895.5653        97.48579
891.5451        97.48579
887.5249        97.48579
883.5047        97.48579
879.4845        97.48579
875.4643        97.48579
871.4441        97.48579
867.4239        97.48579
863.4037        97.48579
859.3835        97.48579
855.3633        97.48579
851.3431        97.48579
847.3229        97.48579
843.3027        97.48579
839.2825        97.48579
835.2623        97.48579
831.2421        97.48579
827.2219        97.48579
823.2017        97.48579
819.1815        97.48579
815.1613        97.48579
```