Assignment 1

Task 1 Report

Start with an empty directory.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 1 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
bigdata@bigdata-VirtualBox:~$
```

Then, Add a new directory called “A1T1” for Input data and upload “testData.txt” to “A1T1”.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir A1T1
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -put testData.txt A1T1
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T1
Found 1 items
-rw-r--r--   1 bigdata supergroup         33 2024-05-02 20:20 A1T1/testData.txt
bigdata@bigdata-VirtualBox:~$
```

Compile solution1.java file using javac and create a solution1.jar file which includes all the related classes inside.

```
bigdata@bigdata-VirtualBox:~$ export HADOOP_CLASSPATH=$($HADOOP_HOME/bin/hadoop
classpath)
bigdata@bigdata-VirtualBox:~$ echo $HADOOP_CLASSPATH
/usr/share/hadoop/etc/hadoop:/usr/share/hadoop-2.7.3/share/hadoop/common/lib/*:/
usr/share/hadoop-2.7.3/share/hadoop/common/*:/usr/share/hadoop-2.7.3/share/hadoo
p/hdfs:/usr/share/hadoop-2.7.3/share/hadoop/hdfs/lib/*:/usr/share/hadoop-2.7.3/s
hare/hadoop/hdfs/*:/usr/share/hadoop-2.7.3/share/hadoop/yarn/lib/*:/usr/share/ha
doop-2.7.3/share/hadoop/yarn/*:/usr/share/hadoop-2.7.3/share/hadoop/mapreduce/li
b/*:/usr/share/hadoop-2.7.3/share/hadoop/mapreduce/*:/usr/share/hadoop/contrib/c
apacity-scheduler/*.jar
bigdata@bigdata-VirtualBox:~$
```

```
bigdata@bigdata-VirtualBox:~$ javac -cp $HADOOP_CLASSPATH solution1.java
Note: solution1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
bigdata@bigdata-VirtualBox:~$ javac -cp $HADOOP_CLASSPATH solution1.java
Note: solution1.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
bigdata@bigdata-VirtualBox:~$ jar cvf solution1.jar solution1*.class
added manifest
adding: solution1.class(in = 1254) (out= 657)(deflated 47%)
bigdata@bigdata-VirtualBox:~$
```

Then, using a solution1.jar file to move testData.txt to another directory. Once it has been moved to another directory, the data have been deleted in original directory.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop jar solution1.jar solution1 A1T1/testData.txt A1T1OUT/testData.txt
File has been deleted
bigdata@bigdata-VirtualBox:~$
```

Then, as shown below , there is no data in “A1T1” and it has been moved to “A1T1OUT”.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 3 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 20:29 A1T1
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 20:29 A1T1OUT
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T1/*
ls: `A1T1/*': No such file or directory
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T1OUT/*
-rw-r--r--   1 bigdata supergroup         33 2024-05-02 20:29 A1T1OUT/testData.txt
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -cat A1T1OUT/testData.txt
Hello World !! I am a test data.
bigdata@bigdata-VirtualBox:~$
```

Assignment 1

Task 2 Report

Start with an empty directory.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 1 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
bigdata@bigdata-VirtualBox:~$
```

Then, add a new directory called “A1T2” for Input data and upload “rain.txt” to “A1T2”.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir A1T2
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 2 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 18:56 A1T2
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -put rain.txt A1T2
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T2/*
-rw-r--r--   1 bigdata supergroup        422 2024-05-02 18:56 A1T2/rain.txt
bigdata@bigdata-VirtualBox:~$
```

Compile solution2.java file using javac and create a solution2.jar file which includes all the related classes inside.

```
bigdata@bigdata-VirtualBox:~$ javac -cp $HADOOP_CLASSPATH solution2.java
bigdata@bigdata-VirtualBox:~$ jar cvf solution2.jar solution2*.class
added manifest
adding: solution2.class(in = 1462) (out= 787)(deflated 46%)
adding: solution2$solution2Mapper.class(in = 2053) (out= 899)(deflated 56%)
adding: solution2$solution2Reducer.class(in = 2721) (out= 1229)(deflated 54%)
bigdata@bigdata-VirtualBox:~$
```

Then, using a solution2.jar file to make a desired output for input text data.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop jar solution2.jar solution2 A1T2/rain.txt A1T2OUT
24/05/02 19:19:53 INFO client.RMProxy: Connecting to ResourceManager at bigdata-VirtualBox/10.0.2.15:8032
24/05/02 19:19:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
24/05/02 19:19:54 INFO input.FileInputFormat: Total input paths to process : 1
24/05/02 19:19:54 INFO mapreduce.JobSubmitter: number of splits:1
24/05/02 19:19:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1714635690331_0008
24/05/02 19:19:54 INFO impl.YarnClientImpl: Submitted application application_1714635690331_0008
24/05/02 19:19:54 INFO mapreduce.Job: The url to track the job: http://bigdata-VirtualBox:8088/proxy/application_1714635690331_0008/
24/05/02 19:19:54 INFO mapreduce.Job: Running job: job_1714635690331_0008
24/05/02 19:19:59 INFO mapreduce.Job: Job job_1714635690331_0008 running in uber mode : false
24/05/02 19:19:59 INFO mapreduce.Job:  map 0% reduce 0%
24/05/02 19:20:03 INFO mapreduce.Job:  map 100% reduce 0%
24/05/02 19:20:08 INFO mapreduce.Job:  map 100% reduce 100%
24/05/02 19:20:08 INFO mapreduce.Job: Job job_1714635690331_0008 completed successfully
```

After creating an output file, then using "-cat" command to display the output.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 3 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 18:56 A1T2
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 19:20 A1T2OUT
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T2OUT/*
-rw-r--r--   1 bigdata supergroup          0 2024-05-02 19:20 A1T2OUT/_SUCCESS
-rw-r--r--   1 bigdata supergroup        264 2024-05-02 19:20 A1T2OUT/part-r-00000
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -cat A1T2OUT/part-r-00000
Mandalay                 203  90  10
NSW                      900 900 900
Queensland                75  50  25
Shan                      80  50  10
South Australia          300 300 300
Victoria                 225 125  10
Western Australia        410 200  10
Yangon                   170 100  20
bigdata@bigdata-VirtualBox:~$
```

Input Data are as follows.

```
Queensland      Gold Coast      25
Victoria        Melbourne       125
Victoria        Geelong 90
Victoria        Wodonga 10
NSW     Lismore 900
Queensland      Brisbane        50
South Australia Adelaide        300
Western Australia       Perth   200
Western Australia       Albany  200
Western Australia       Broome  10
Yangon  Insein  20
Yangon  Yankin  50
Yangon  Mingalardon     100
Shan    Kalaw   10
Shan    Lashio  20
Shan    Intaw   50
Mandalay        Kyaukse 90
Mandalay        Mahlaing        70
Mandalay        Pyawbwe 10
Mandalay        Bagan   33
```

Queensland	Gold Coast	25
Victoria	Melbourne	125
Victoria	Geelong	90
Victoria	Wodonga	10
NSW	Lismore	900
Queensland	Brisbane	50
South Australia	Adelaide	300
Western Australia	Perth	200
Western Australia	Albany	200
Western Australia	Broome	10
Yangon	Insein	20
Yangon	Yankin	50
Yangon	Mingalardon	100
Shan	Kalaw	10
Shan	Lashio	20
Shan	Intaw	50
Mandalay	Kyaukse	90
Mandalay	Mahlaing	70
Mandalay	Pyawbwe	10
Mandalay	Bagan	33

Assignment 1

Task 3 Report

Start with an empty directory.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 1 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
bigdata@bigdata-VirtualBox:~$
```

Then, add a new directory called “A1T3” for Input data and upload “grep.txt” to “A1T3”.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -mkdir A1T3
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 2 items
drwxr-xr-x   - bigdata supergroup          0 2017-07-03 01:33 .sparkStaging
drwxr-xr-x   - bigdata supergroup          0 2024-05-02 20:02 A1T3
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -put grep.txt A1T3
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T3
Found 1 items
-rw-r--r--   1 bigdata supergroup       1025 2024-05-02 20:03 A1T3/grep.txt
bigdata@bigdata-VirtualBox:~$
```

Compile solution3.java file using javac and create a solution3.jar file which includes all the related classes inside.

```
bigdata@bigdata-VirtualBox:~$ javac -cp $HADOOP_CLASSPATH solution3.java
bigdata@bigdata-VirtualBox:~$ jar cvf solution3.jar solution3*.class
added manifest
adding: solution3.class(in = 1465) (out= 792)(deflated 45%)
adding: solution3$solution3Mapper.class(in = 2590) (out= 1143)(deflated 55%)
adding: solution3$solution3Reducer.class(in = 2264) (out= 978)(deflated 56%)
bigdata@bigdata-VirtualBox:~$
```

Then, using a solution2.jar file to make a desired output for input text data.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop jar solution3.jar solution3 A1T3/grep.txt A1T3OUT
24/05/02 20:07:59 INFO client.RMProxy: Connecting to ResourceManager at bigdata-VirtualBox/10.0.2.15:8032
24/05/02 20:07:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
24/05/02 20:07:59 INFO input.FileInputFormat: Total input paths to process : 1
24/05/02 20:07:59 INFO mapreduce.JobSubmitter: number of splits:1
24/05/02 20:07:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1714635690331_0009
24/05/02 20:07:59 INFO impl.YarnClientImpl: Submitted application application_1714635690331_0009
24/05/02 20:07:59 INFO mapreduce.Job: The url to track the job: http://bigdata-VirtualBox:8088/proxy/application_1714635690331_0009/
24/05/02 20:07:59 INFO mapreduce.Job: Running job: job_1714635690331_0009
24/05/02 20:08:05 INFO mapreduce.Job: Job job_1714635690331_0009 running in uber mode : false
24/05/02 20:08:05 INFO mapreduce.Job:  map 0% reduce 0%
24/05/02 20:08:09 INFO mapreduce.Job:  map 100% reduce 0%
24/05/02 20:08:13 INFO mapreduce.Job:  map 100% reduce 100%
24/05/02 20:08:13 INFO mapreduce.Job: Job job_1714635690331_0009 completed successfully
```

After creating an output file, then using “-cat” command to display the output.

```
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -ls A1T3OUT/*
-rw-r--r--   1 bigdata supergroup          0 2024-05-02 20:08 A1T3OUT/_SUCCESS
-rw-r--r--   1 bigdata supergroup        156 2024-05-02 20:08 A1T3OUT/part-r-00000
bigdata@bigdata-VirtualBox:~$ $HADOOP_HOME/bin/hadoop fs -cat A1T3OUT/part-r-00000
X short:        17 words
short:          24 words
medium:         29 words
long:           18 words
X long:         10 words
XX long:        14 words
bigdata@bigdata-VirtualBox:~$
```