Experiment 3

Aim: Implement the classic Map Reduce word count algorithm to count the frequency of words in a large text corpus stored in HDFS.

Directory:

```
C:\Windows\System32>hadoop dfs -mkdir /user/rawat
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
mkdir: `hdfs://localhost:9000/user': No such file or directory

C:\Windows\System32>hdfs dfs -mkdir /user/rawat
mkdir: `hdfs://localhost:9000/user': No such file or directory

C:\Windows\System32>hdfs dfs -mkdir /user

C:\Windows\System32>hdfs dfs -mkdir /user/rawat
```

Fig 1: Making Directory

Map Reduce Job:

```
C:\Windows\System32>hadoop jar "C:\Hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar" -input /user/rawat/test_input/test.txt -output /user/rawat/test_output -mapper
"python3 mapper.py" -reducer "python3 reducer.py" -file C:\Users\rawat\Desktop\mapper.py -file C:\Users\rawat\Desktop\reducer.py
2024-09-03 22:19:49,341 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [C:\Users\rawat\Desktop\mapper.py, C:\Users\rawat\Desktop\reducer.py, /C:/Users/rawat/AppData/Local/Temp/hadoop-unjar243489378834739534/] [] C:\Users\rawat\A
ppData\Local\Temp\streamjob8393153507915007475.jar tmpDir=null
2024-09-03 22:19:51,291 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-03 22:19:51,651 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-03 22:19:52,317 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/rawat/.staging/job_1725380464454_0004
2024-09-03 22:19:52,824 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-03 22:19:52,903 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-03 22:19:53,109 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725380464454_0004
2024-09-03 22:19:53,109 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-03 22:19:53,411 INFO conf.Configuration: resource-types.xml not found
2024-09-03 22:19:53,416 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-03 22:19:53,570 INFO impl.YarnClientImpl: Submitted application application_1725380464454_0004
2024-09-03 22:19:53,652 INFO mapreduce.Job: The url to track the job: http://MADHURIMA-RAWAT:8088/proxy/application_1725380464454_0004/
2024-09-03 22:19:53,656 INFO mapreduce.Job: Running job: job_1725380464454_0004
2024-09-03 22:20:03,949 INFO mapreduce.Job: Job job_1725380464454_0004 running in uber mode : false
2024-09-03 22:20:03,949 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-03 22:20:12,156 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-03 22:20:20,265 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-03 22:20:21,287 INFO mapreduce.Job: Job job_1725380464454_0004 completed successfully
2024-09-03 22:20:21,421 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=74
                FILE: Number of bytes written=843588
```

Fig 2: Executing Map Reduce Job

```
        File Input Format Counters
                Bytes Read=59
        File Output Format Counters
                Bytes Written=54
2024-09-03 22:20:21,422 INFO streaming.StreamJob: Output directory: /user/rawat/test_output

C:\Windows\System32>hadoop fs -ls /user/rawat/test_output
Found 2 items
-rw-r--r--   3 rawat supergroup          0 2024-09-03 22:20 /user/rawat/test_output/_SUCCESS
-rw-r--r--   3 rawat supergroup         54 2024-09-03 22:20 /user/rawat/test_output/part-00000

C:\Windows\System32>
C:\Windows\System32>hadoop fs -cat /user/rawat/test_output/part-00000
Hello   1
World   1
file    1
haddop  1
is      1
testing 1
this    1
```

Fig 3: Output of Map Reduce Job