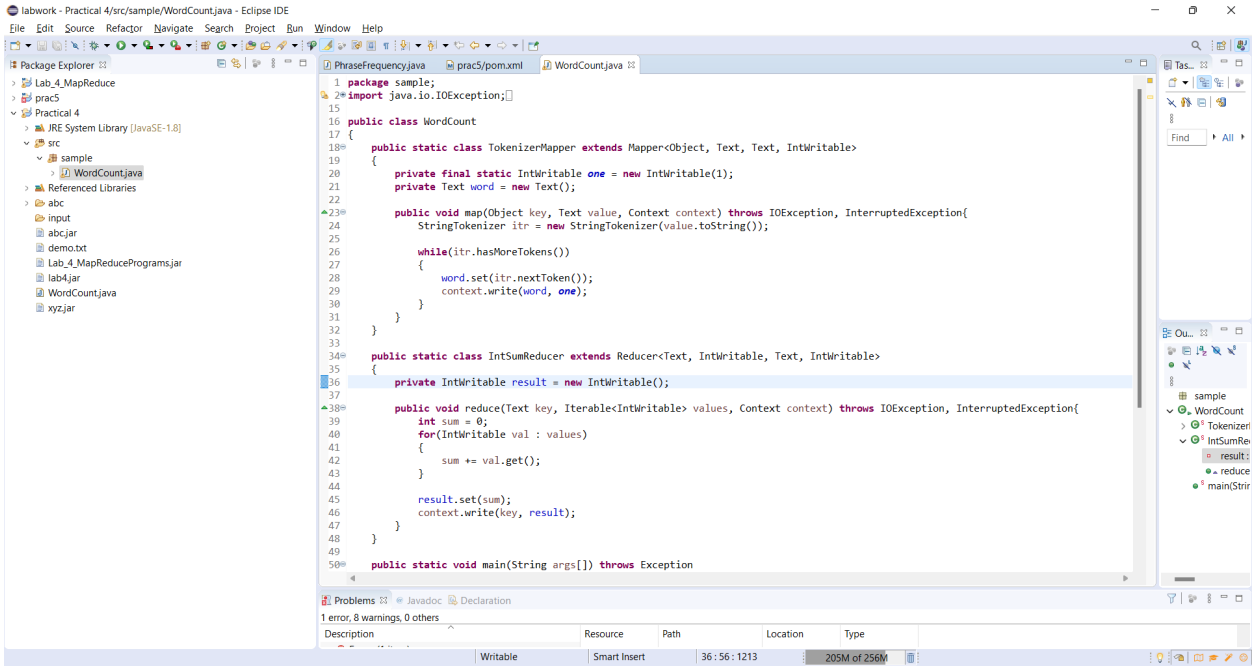
2CS702 Big Data Analytics

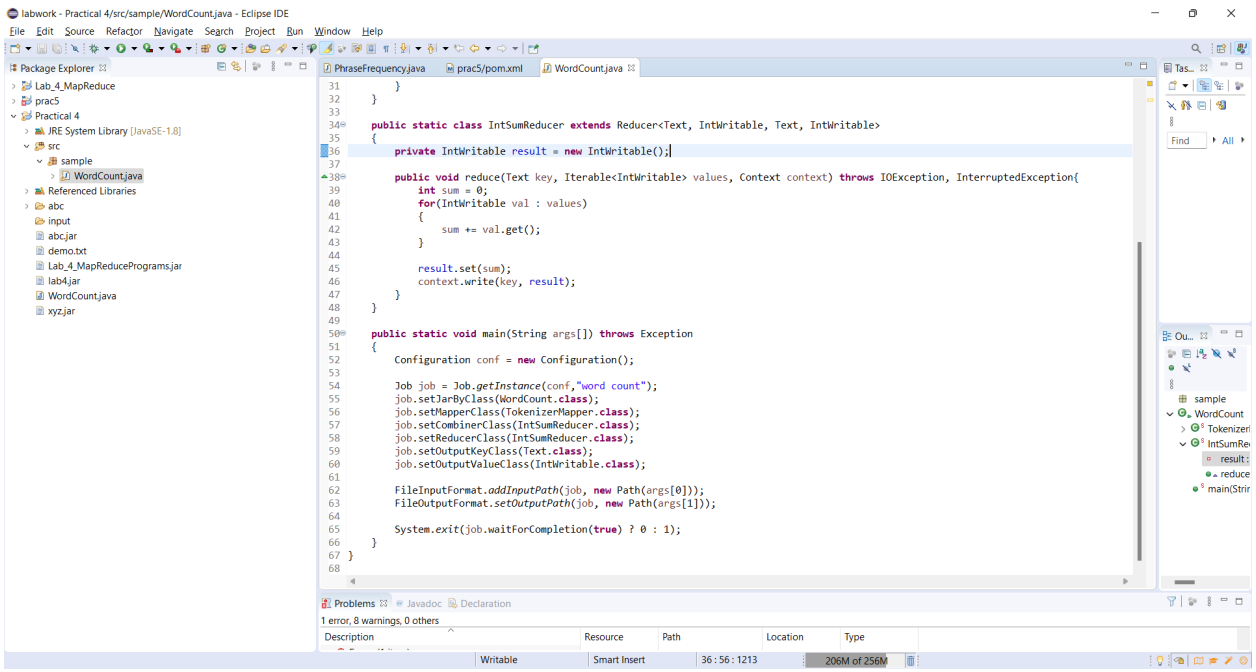
Lab-6 Task

Submitted by: Labdhi Sheth 18BCE101

Aim: **Analyse impact of different numbers of mapper and reducers on same definition as practical 4.**

Code:





Changes to be made in code:

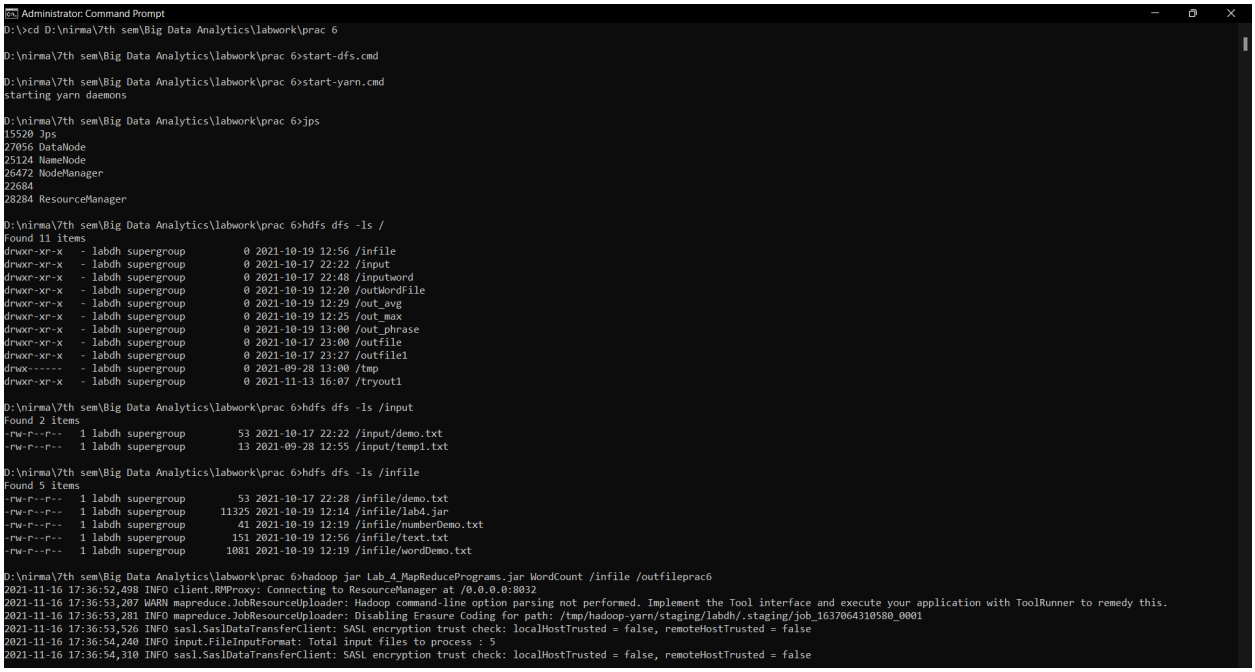
1. As mentioned to reduce the number of reducers to 0 we must edit the following sentence
   in the code: job.setNumReduceTasks(0);.

2. And to increase the number of reducers we need to edit the above sentence and instead of 0, we need to put in (n) the number we want to substitute. In my instance, I changed it to job.setNumReduceTasks(5);.

Changes observed:

1. So, it is usually preferred to keep the number of reducers such that it is a multiple of the block size, does not take a long time to execute, doesn't require a lot of file creations.
2. The number of reducers can't exceed the number of partitions, so we also need to take that into consideration when deciding the number of reducers. As for the instance, we make the number of reducers 10 but we have a single partition system then the output will be divided into 10 shards but those 10 outputs will be consolidated by a single reducer.
3. The benefits of having more reducers are that it increases load balancing, lowers the cost of failure.
4. But on the other hand, having too many reducers can also increase the framework overhead, slower start-up time, and increase the number of input for the next tasks.

Output:


Administrator: Command Prompt
D:\>cd D:\nirma\7th sem\Big Data Analytics\labwork\prac 6
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>start-dfs.cmd
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>start-yarn.cmd
starting yarn daemons
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>jps
15520 Jps
27056 DataNode
25124 NameNode
26472 NodeManager
22684
28284 ResourceManager
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hdfs dfs -ls /
Found 11 items
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:56 /infile
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:22 /input
drwxr-xr-x - labdh supergroup 0 2021-10-17 22:48 /inputword
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:20 /outWordFile
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:29 /out_avg
drwxr-xr-x - labdh supergroup 0 2021-10-19 12:25 /out_max
drwxr-xr-x - labdh supergroup 0 2021-10-19 13:00 /out_phrase
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:00 /outfile
drwxr-xr-x - labdh supergroup 0 2021-10-17 23:27 /outfile1
drwx------ - labdh supergroup 0 2021-09-28 13:00 /tmp
drwxr-xr-x - labdh supergroup 0 2021-11-13 16:07 /tryout1
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hdfs dfs -ls /input
Found 2 items
-rw-r--r-- 1 labdh supergroup 53 2021-10-17 22:22 /input/demo.txt
-rw-r--r-- 1 labdh supergroup 13 2021-09-28 12:55 /input/temp1.txt
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hdfs dfs -ls /infile
Found 5 items
-rw-r--r-- 1 labdh supergroup 53 2021-10-17 22:28 /infile/demo.txt
-rw-r--r-- 1 labdh supergroup 11325 2021-10-19 12:14 /infile/lab4.jar
-rw-r--r-- 1 labdh supergroup 41 2021-10-19 12:19 /infile/numberDemo.txt
-rw-r--r-- 1 labdh supergroup 151 2021-10-19 12:56 /infile/text.txt
-rw-r--r-- 1 labdh supergroup 1081 2021-10-19 12:19 /infile/wordDemo.txt
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hadoop jar Lab_4_MapReducePrograms.jar WordCount /infile /outfileprac6
2021-11-16 17:36:52,498 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 17:36:53,207 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-11-16 17:36:53,281 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1637064310580_0001
2021-11-16 17:36:53,526 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-11-16 17:36:54,240 INFO input.FileInputFormat: Total input files to process : 5
2021-11-16 17:36:54,310 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

```
Administrator: Command Prompt

D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hadoop jar Lab_4_MapReducePrograms.jar WordCount /infile /outfileprac6
2021-11-16 17:36:52,498 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-16 17:36:53,207 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-11-16 17:36:53,281 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/labdh/.staging/job_1637064310580_0001
2021-11-16 17:36:53,526 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-11-16 17:36:54,240 INFO input.FileInputFormat: Total input files to process : 5
2021-11-16 17:36:54,310 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-11-16 17:36:54,379 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-11-16 17:36:54,401 INFO mapreduce.JobSubmitter: number of splits:5
2021-11-16 17:36:54,557 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-11-16 17:36:54,591 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1637064310580_0001
2021-11-16 17:36:54,592 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-16 17:36:54,750 INFO conf.Configuration: resource-types.xml not found
2021-11-16 17:36:54,751 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-16 17:36:54,962 INFO impl.YarnClientImpl: Submitted application application_1637064310580_0001
2021-11-16 17:36:55,002 INFO mapreduce.Job: The url to track the job: http://LAPTOP-0Q3JVTL1:8088/proxy/application_1637064310580_0001/
2021-11-16 17:36:55,002 INFO mapreduce.Job: Running job: job_1637064310580_0001
2021-11-16 17:37:03,167 INFO mapreduce.Job: Job job_1637064310580_0001 running in uber mode : false
2021-11-16 17:37:03,169 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-16 17:37:15,913 INFO mapreduce.Job:  map 80% reduce 0%
2021-11-16 17:37:16,927 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-16 17:37:24,007 INFO mapreduce.Job:  map 100% reduce 100%
2021-11-16 17:37:25,032 INFO mapreduce.Job: Job job_1637064310580_0001 completed successfully
2021-11-16 17:37:25,116 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=23194
                FILE: Number of bytes written=1405845
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=13171
                HDFS: Number of bytes written=20822
                HDFS: Number of read operations=20
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=53309
                Total time spent by all reduces in occupied slots (ms)=5054
                Total time spent by all map tasks (ms)=53309
                Total time spent by all reduce tasks (ms)=5054
                Total vcore-milliseconds taken by all map tasks=53309
                Total vcore-milliseconds taken by all reduce tasks=5054
                Total megabyte-milliseconds taken by all map tasks=54588416
                Total megabyte-milliseconds taken by all reduce tasks=5175296
        Map-Reduce Framework
```

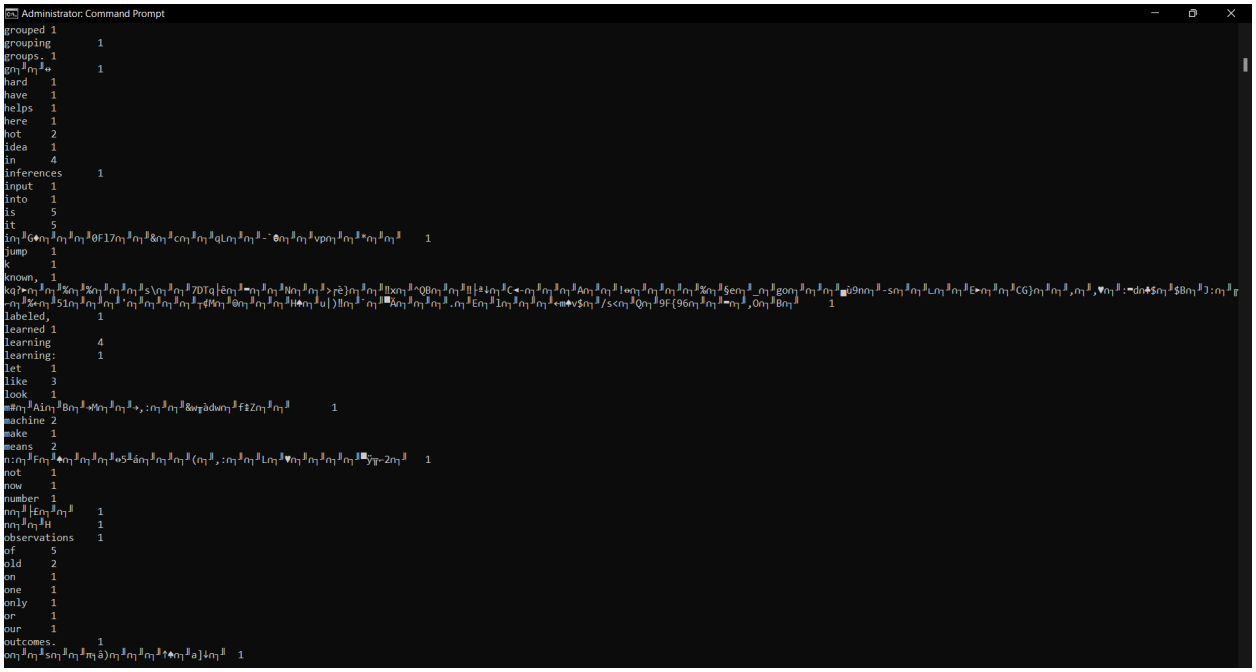
```
Administrator: Command Prompt

                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=53309
                Total time spent by all reduces in occupied slots (ms)=5054
                Total time spent by all map tasks (ms)=53309
                Total time spent by all reduce tasks (ms)=5054
                Total vcore-millliseconds taken by all map tasks=53309
                Total vcore-millliseconds taken by all reduce tasks=5054
                Total megabyte-milliseconds taken by all map tasks=54588416
                Total megabyte-milliseconds taken by all reduce tasks=5175296
        Map-Reduce Framework
                Map input records=90
                Map output records=429
                Map output bytes=22265
                Map output materialized bytes=23218
                Input split bytes=520
                Combine input records=0
                Combine output records=0
                Reduce input groups=359
                Reduce shuffle bytes=23218
                Reduce input records=429
                Reduce output records=359
                Spilled Records=858
                Shuffled Maps =5
                Failed Shuffles=0
                Merged Map outputs=5
                GC time elapsed (ms)=411
                CPU time spent (ms)=4256
                Physical memory (bytes) snapshot=1650475008
                Virtual memory (bytes) snapshot=2513129472
                Total committed heap usage (bytes)=1409286144
                Peak Map Physical memory (bytes)=293937152
                Peak Map Virtual memory (bytes)=443097088
                Peak Reduce Physical memory (bytes)=199962624
                Peak Reduce Virtual memory (bytes)=352751616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=12651
        File Output Format Counters
                Bytes Written=20822

D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hadoop fs -ls /outfileprac6
```


Administrator: Command Prompt
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hadoop fs -ls /outfileprac6
Found 2 items
-rw-r--r--   1 labdh supergroup   0 2021-11-16 17:37 /outfileprac6/_SUCCESS
-rw-r--r--   1 labdh supergroup   20822 2021-11-16 17:37 /outfileprac6/part-r-00000
D:\nirma\7th sem\Big Data Analytics\labwork\prac 6>hadoop fs -cat /outfileprac6/part*
2021-11-16 17:39:37,090 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false


Administrator: Command Prompt



Conclusion:

Running the above map-reduce program we analyzed that when the number of reducers were set to 0 the output was not grouped and but when the number of reducers were set to 5 we got the output divided into 5 groups and it was all consolidated.