Practical-5

Name : Yash Marthak

Roll No. : 19BCE122

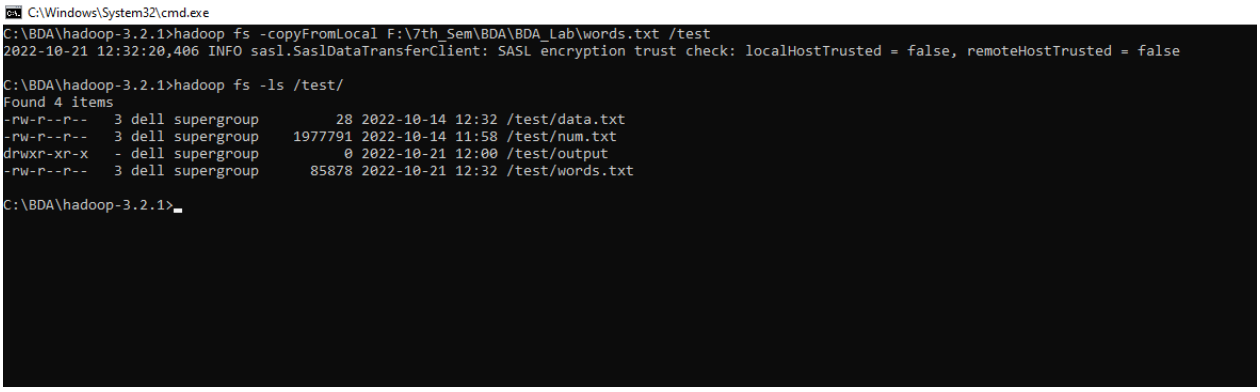
Couse Name : Big Data Analytics

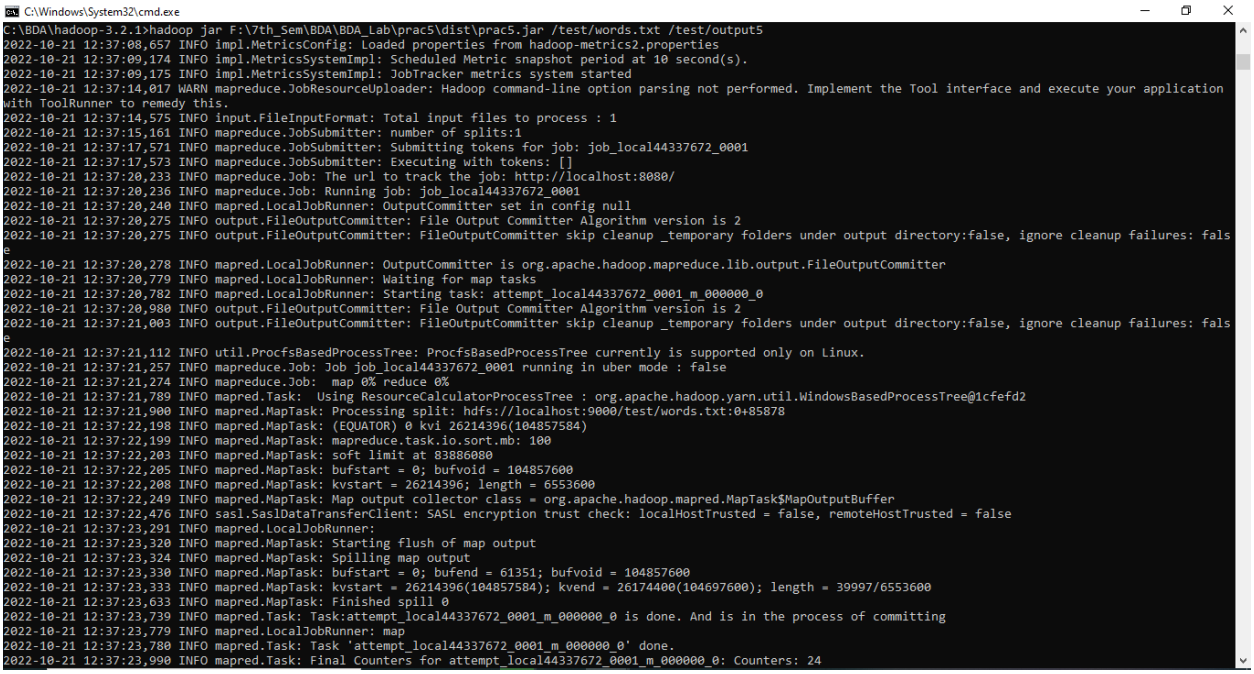
Aim:

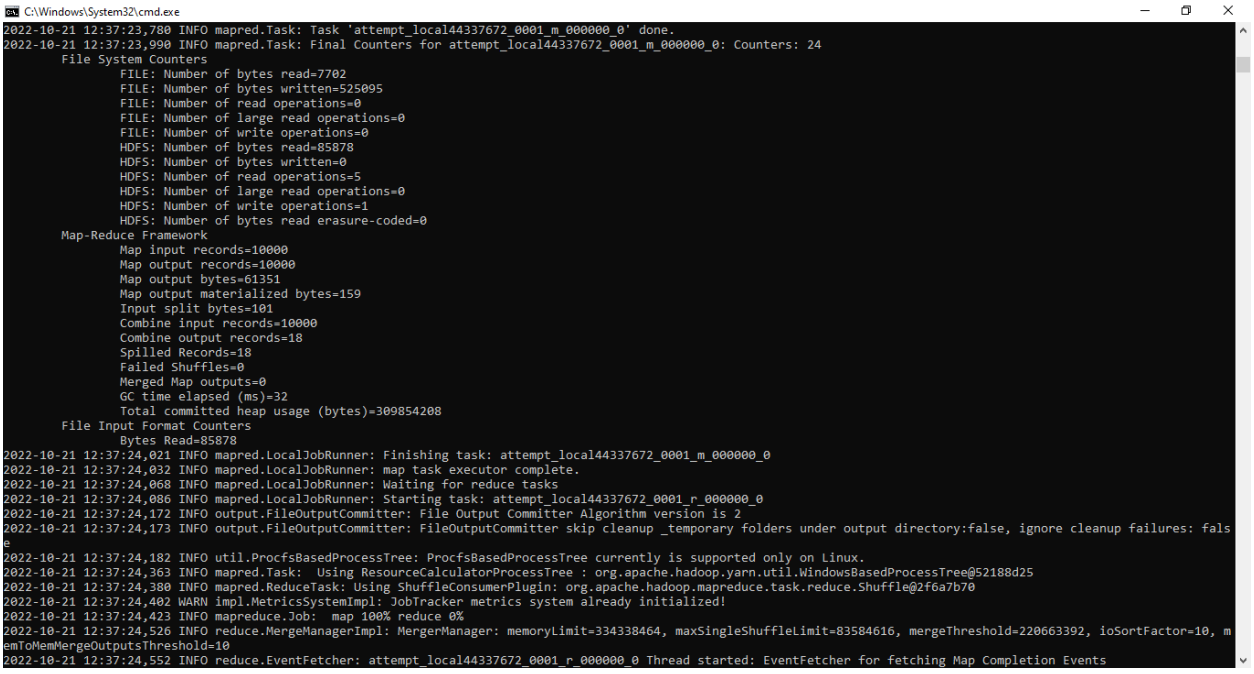
Apply MapReduce algorithms to find phrase frequency from given dataset.

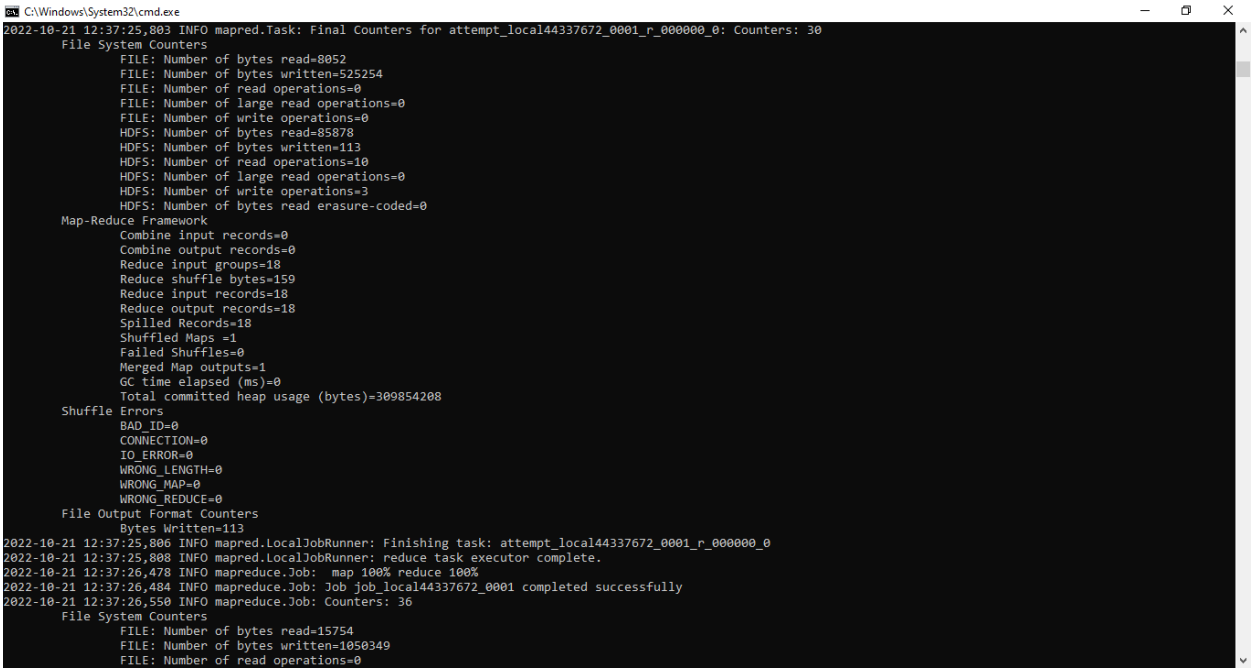
- Prepare a report to guide design of mapper and reducer.

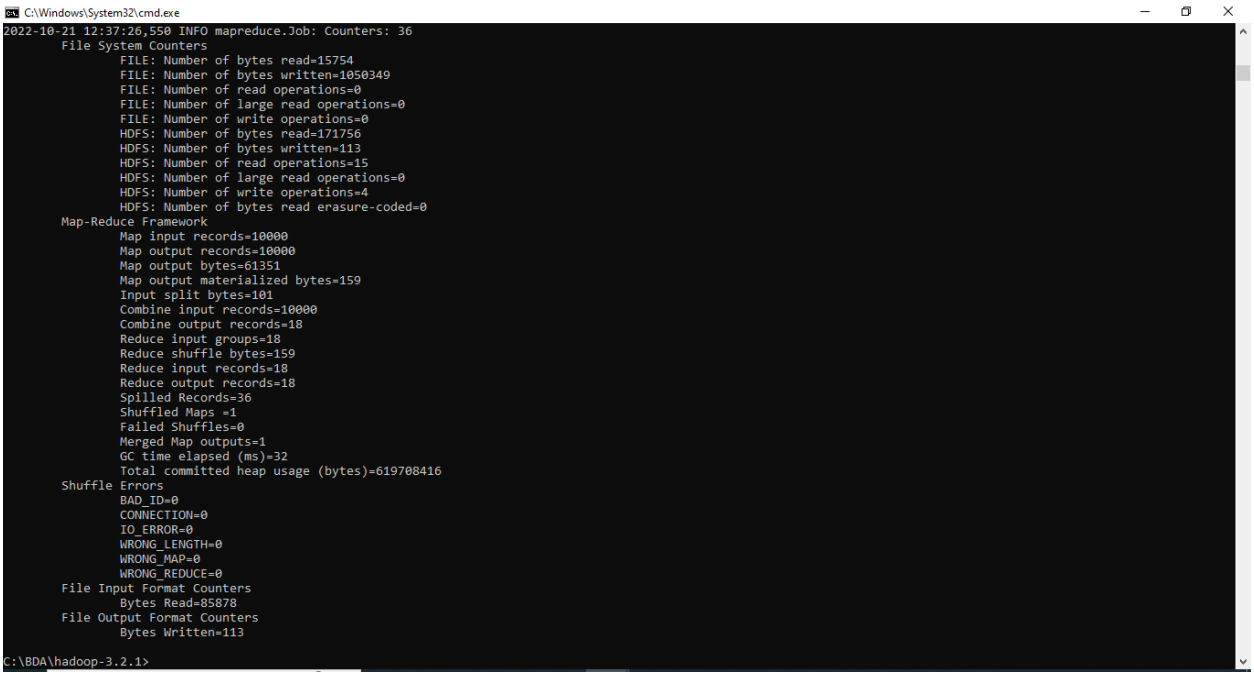
Output Screenshots:

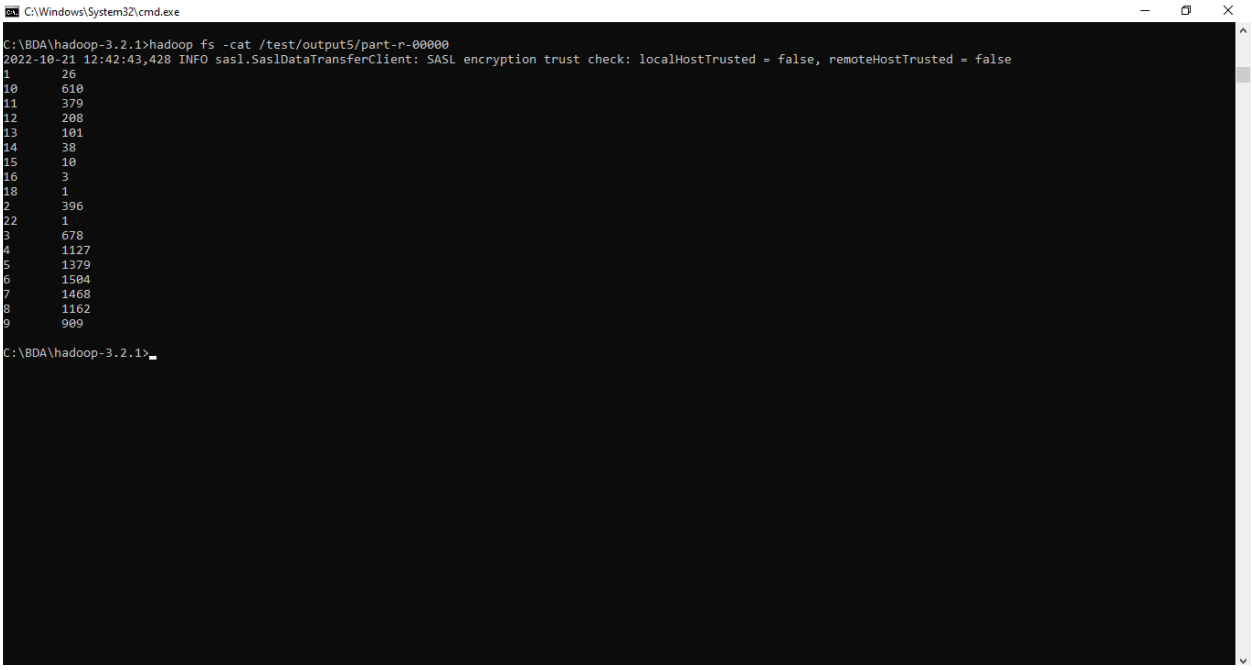



C:\Windows\System32\cmd.exe
C:\BDA\hadoop-3.2.1>hadoop jar F:\7th_Sem\BDA\BDA_Lab\prac5\dist\prac5.jar /test/words.txt /test/output5
2022-10-21 12:37:08,657 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-10-21 12:37:09,174 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-10-21 12:37:09,175 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2022-10-21 12:37:14,017 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-10-21 12:37:14,575 INFO input.FileInputFormat: Total input files to process : 1
2022-10-21 12:37:15,161 INFO mapreduce.JobSubmitter: number of splits:1
2022-10-21 12:37:17,571 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local44337672_0001
2022-10-21 12:37:17,573 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-10-21 12:37:20,233 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-10-21 12:37:20,236 INFO mapreduce.Job: Running job: job_local44337672_0001
2022-10-21 12:37:20,240 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-10-21 12:37:20,275 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-21 12:37:20,275 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-10-21 12:37:20,278 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2022-10-21 12:37:20,779 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-10-21 12:37:20,782 INFO mapred.LocalJobRunner: Starting task: attempt_local44337672_0001_m_000000_0
2022-10-21 12:37:20,980 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-21 12:37:21,003 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-10-21 12:37:21,112 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2022-10-21 12:37:21,257 INFO mapreduce.Job: Job job_local44337672_0001 running in uber mode : false
2022-10-21 12:37:21,274 INFO mapreduce.Job: map 0% reduce 0%
2022-10-21 12:37:21,789 INFO mapred.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@1cfefd2
2022-10-21 12:37:21,900 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/test/words.txt:0+85878
2022-10-21 12:37:22,198 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2022-10-21 12:37:22,199 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2022-10-21 12:37:22,203 INFO mapred.MapTask: soft limit at 83886080
2022-10-21 12:37:22,205 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2022-10-21 12:37:22,208 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2022-10-21 12:37:22,249 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2022-10-21 12:37:22,476 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-10-21 12:37:23,291 INFO mapred.LocalJobRunner:
2022-10-21 12:37:23,320 INFO mapred.MapTask: Starting flush of map output
2022-10-21 12:37:23,324 INFO mapred.MapTask: Spilling map output
2022-10-21 12:37:23,330 INFO mapred.MapTask: bufstart = 0; bufend = 61351; bufvoid = 104857600
2022-10-21 12:37:23,333 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26174400(104697600); length = 39997/6553600
2022-10-21 12:37:23,633 INFO mapred.MapTask: Finished spill 0
2022-10-21 12:37:23,739 INFO mapred.Task: Task:attempt_local44337672_0001_m_000000_0 is done. And is in the process of committing
2022-10-21 12:37:23,779 INFO mapred.LocalJobRunner: map
2022-10-21 12:37:23,780 INFO mapred.Task: Task 'attempt_local44337672_0001_m_000000_0' done.
2022-10-21 12:37:23,990 INFO mapred.Task: Final Counters for attempt_local44337672_0001_m_000000_0: Counters: 24


C:\Windows\System32\cmd.exe
2022-10-21 12:37:23,780 INFO mapred.Task: Task 'attempt_local44337672_0001_m_000000_0' done.
2022-10-21 12:37:23,990 INFO mapred.Task: Final Counters for attempt_local44337672_0001_m_000000_0: Counters: 24
File System Counters
FILE: Number of bytes read=7702
FILE: Number of bytes written=525095
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=85878
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=10000
Map output bytes=61351
Map output materialized bytes=159
Input split bytes=101
Combine input records=10000
Combine output records=18
Spilled Records=18
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=32
Total committed heap usage (bytes)=309854208
File Input Format Counters
Bytes Read=85878
2022-10-21 12:37:24,021 INFO mapred.LocalJobRunner: Finishing task: attempt_local44337672_0001_m_000000_0
2022-10-21 12:37:24,032 INFO mapred.LocalJobRunner: map task executor complete.
2022-10-21 12:37:24,068 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2022-10-21 12:37:24,086 INFO mapred.LocalJobRunner: Starting task: attempt_local44337672_0001_r_000000_0
2022-10-21 12:37:24,172 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-21 12:37:24,173 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-10-21 12:37:24,182 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2022-10-21 12:37:24,363 INFO mapred.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@52188d25
2022-10-21 12:37:24,380 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2f6a7b70
2022-10-21 12:37:24,402 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2022-10-21 12:37:24,423 INFO mapreduce.Job: map 100% reduce 0%
2022-10-21 12:37:24,526 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2022-10-21 12:37:24,552 INFO reduce.EventFetcher: attempt_local44337672_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events


C:\Windows\System32\cmd.exe
2022-10-21 12:37:25,803 INFO mapred.Task: Final Counters for attempt_local44337672_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=8052
FILE: Number of bytes written=525254
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=85878
HDFS: Number of bytes written=113
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=18
Reduce shuffle bytes=159
Reduce input records=18
Reduce output records=18
Spilled Records=18
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=309854208
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=113
2022-10-21 12:37:25,806 INFO mapred.LocalJobRunner: Finishing task: attempt_local44337672_0001_r_000000_0
2022-10-21 12:37:25,808 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-10-21 12:37:26,478 INFO mapreduce.Job: map 100% reduce 100%
2022-10-21 12:37:26,484 INFO mapreduce.Job: Job job_local44337672_0001 completed successfully
2022-10-21 12:37:26,550 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=15754
FILE: Number of bytes written=1050349
FILE: Number of read operations=0


C:\Windows\System32\cmd.exe
2022-10-21 12:37:26,550 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=15754
FILE: Number of bytes written=1050349
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=171756
HDFS: Number of bytes written=113
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=10000
Map output records=10000
Map output bytes=61351
Map output materialized bytes=159
Input split bytes=101
Combine input records=10000
Combine output records=18
Reduce input groups=18
Reduce shuffle bytes=159
Reduce input records=18
Reduce output records=18
Spilled Records=36
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=32
Total committed heap usage (bytes)=619708416
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=85878
File Output Format Counters
Bytes Written=113
C:\BDA\hadoop-3.2.1>


C:\Windows\System32\cmd.exe
C:\BDA\hadoop-3.2.1>hadoop fs -cat /test/output5/part-r-00000
2022-10-21 12:42:43,428 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
1 26
10 610
11 379
12 208
13 101
14 38
15 10
16 3
18 1
2 396
22 1
3 678
4 1127
5 1379
6 1504
7 1468
8 1162
9 909
C:\BDA\hadoop-3.2.1>