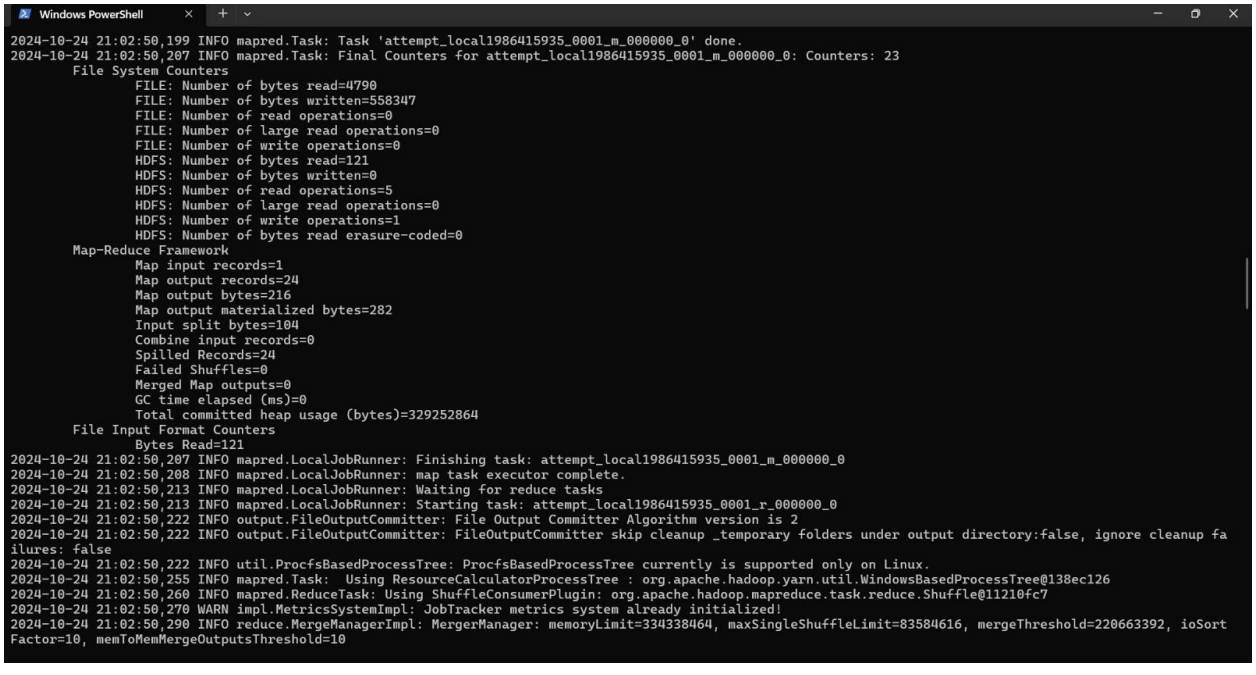
BDA
Practical 5
MapReduce algorithms

21BCE020

```
2024-10-24 20:21:00,463 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory
e cleanup failures: false
2024-10-24 20:21:00,502 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-10-24 20:21:00,506 INFO mapred.LocalJobRunner: Starting task: attempt_local1599803945_0001_m_000000_0
2024-10-24 20:21:00,529 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-10-24 20:21:00,530 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory
e cleanup failures: false
2024-10-24 20:21:00,538 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2024-10-24 20:21:00,571 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTre
2024-10-24 20:21:00,587 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/ahan/Prac5/Input/Prac5test.txt:0+121
2024-10-24 20:21:00,613 INFO mapred.MapTask: numReduceTasks: 1
2024-10-24 20:21:00,649 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-10-24 20:21:00,649 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-10-24 20:21:00,649 INFO mapred.MapTask: soft limit at 83886080
2024-10-24 20:21:00,649 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-10-24 20:21:00,650 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-10-24 20:21:00,653 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-10-24 20:21:00,777 INFO mapred.LocalJobRunner:
2024-10-24 20:21:00,778 INFO mapred.MapTask: Starting flush of map output
2024-10-24 20:21:00,778 INFO mapred.MapTask: Spilling map output
2024-10-24 20:21:00,778 INFO mapred.MapTask: bufstart = 0; bufend = 216; bufvoid = 104857600
2024-10-24 20:21:00,778 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214304(104857216); length = 93/6553600
2024-10-24 20:21:00,793 INFO mapred.MapTask: Finished spill 0
2024-10-24 20:21:00,807 INFO mapred.Task: Task:attempt_local1599803945_0001_m_000000_0 is done. And is in the process of committing
2024-10-24 20:21:00,811 INFO mapred.LocalJobRunner: hdfs://localhost:9000/ahan/Prac5/Input/Prac5test.txt:0+121
2024-10-24 20:21:00,811 INFO mapred.Task: Task 'attempt_local1599803945_0001_m_000000_0' done.
2024-10-24 20:21:00,819 INFO mapred.Task: Final Counters for attempt_local1599803945_0001_m_000000_0: Counters: 23
```

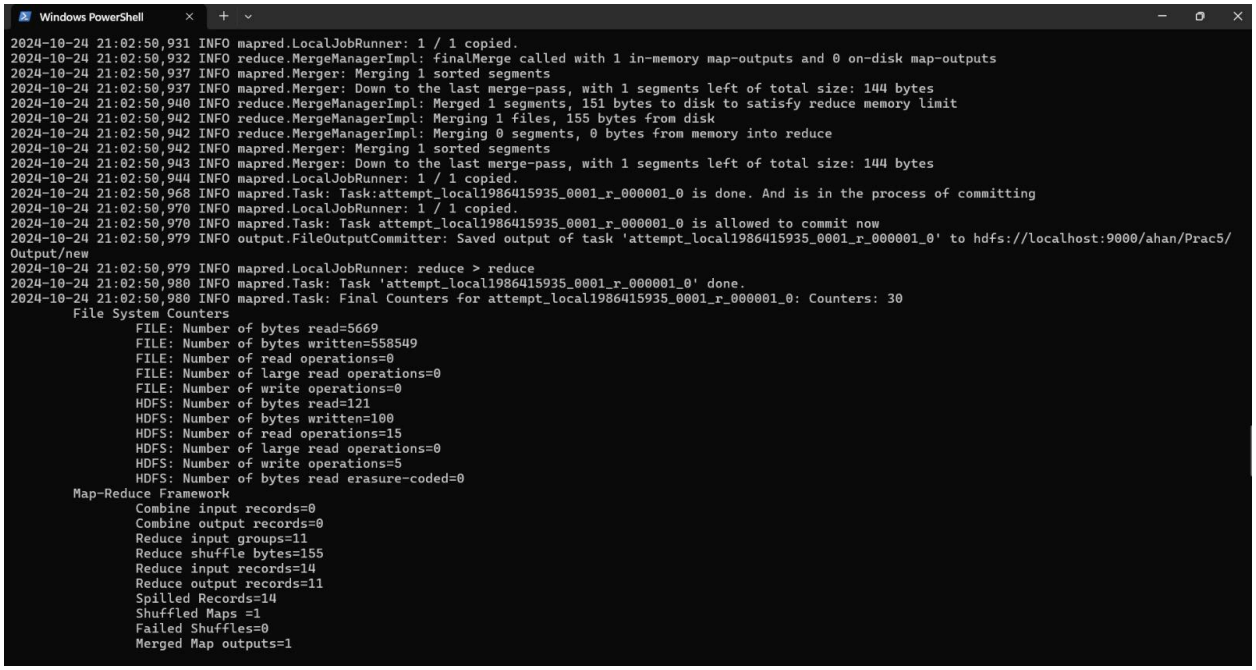
Taking a count of 3 reducers

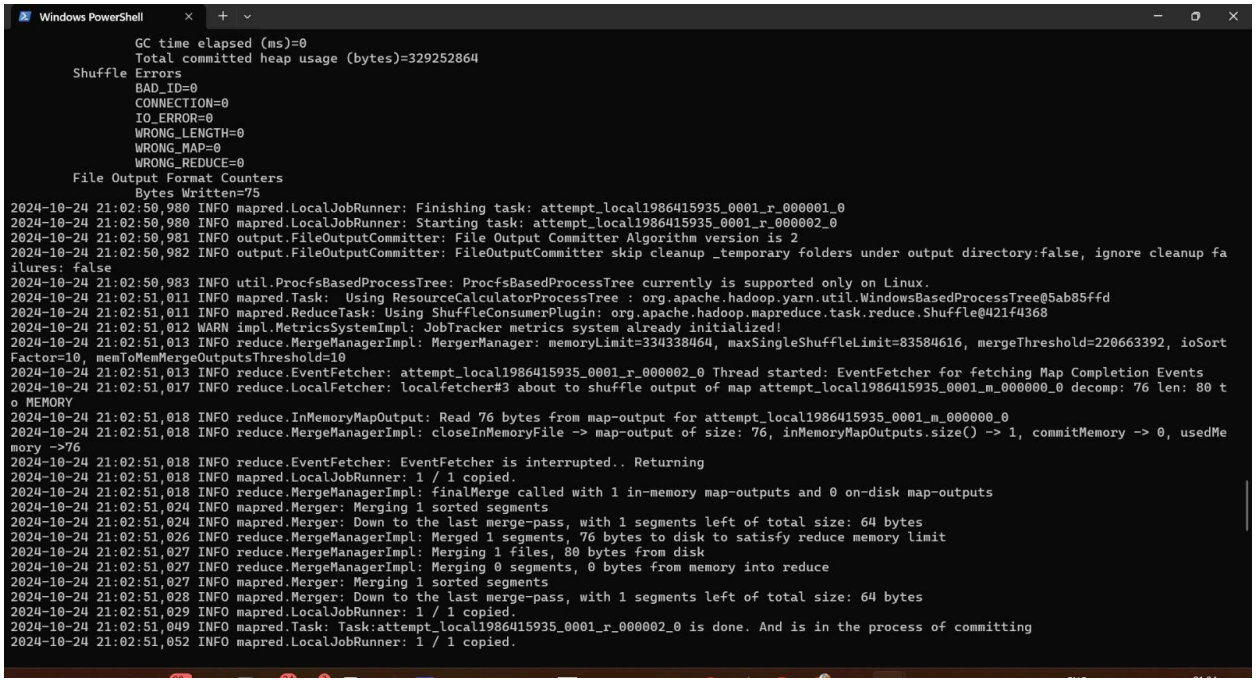


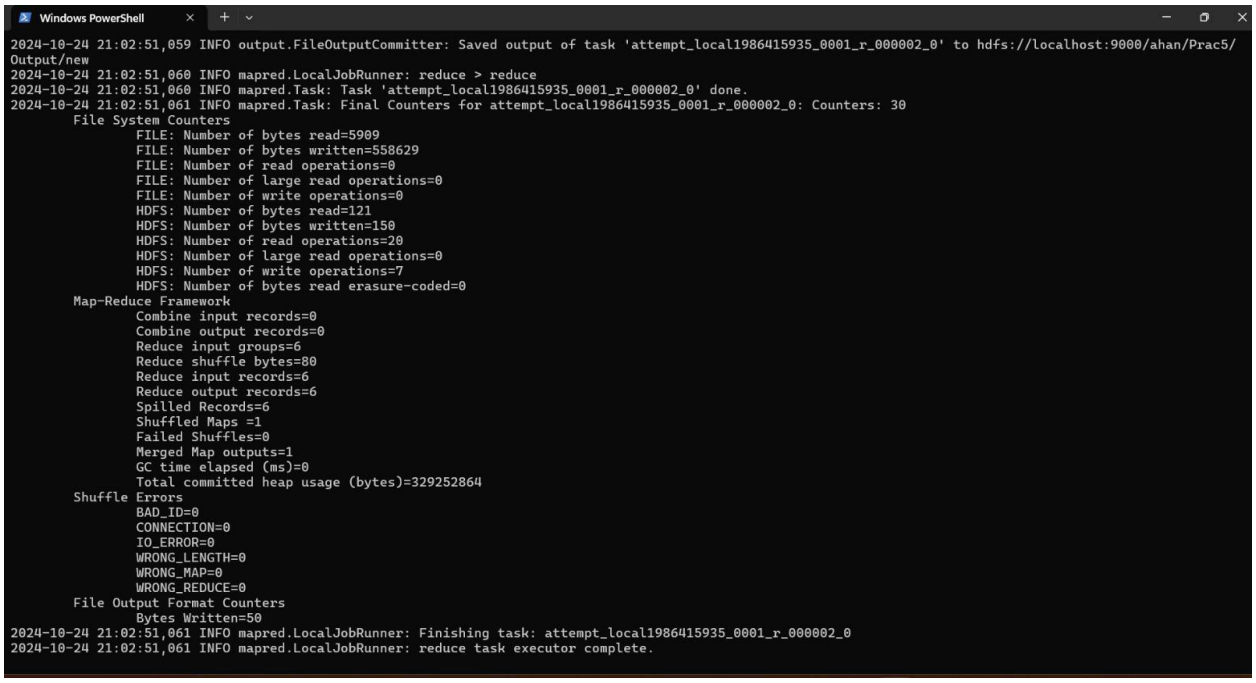
```
2024-10-24 21:02:50,255 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@138ec126
2024-10-24 21:02:50,260 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@11210fc7
2024-10-24 21:02:50,270 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-10-24 21:02:50,290 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSort
Factor=10, memToMemMergeOutputsThreshold=10
2024-10-24 21:02:50,294 INFO reduce.EventFetcher: attempt_local1986415935_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2024-10-24 21:02:50,320 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1986415935_0001_m_000000_0 decomp: 43 len: 47 t
o MEMORY
2024-10-24 21:02:50,326 INFO reduce.InMemoryMapOutput: Read 43 bytes from map-output for attempt_local1986415935_0001_m_000000_0
2024-10-24 21:02:50,327 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 43, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMe
mory ->43
2024-10-24 21:02:50,329 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2024-10-24 21:02:50,330 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,331 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2024-10-24 21:02:50,340 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:50,341 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 38 bytes
2024-10-24 21:02:50,343 INFO reduce.MergeManagerImpl: Merged 1 segments, 43 bytes to disk to satisfy reduce memory limit
2024-10-24 21:02:50,344 INFO reduce.MergeManagerImpl: Merging 1 files, 47 bytes from disk
2024-10-24 21:02:50,345 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2024-10-24 21:02:50,345 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:50,346 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 38 bytes
2024-10-24 21:02:50,347 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,692 INFO mapreduce.Job: Job job_local1986415935_0001 running in uber mode : false
2024-10-24 21:02:50,692 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-24 21:02:50,874 INFO mapred.Task: Task:attempt_local1986415935_0001_r_000000_0 is done. And is in the process of committing
2024-10-24 21:02:50,875 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,876 INFO mapred.Task: Task attempt_local1986415935_0001_r_000000_0 is allowed to commit now
2024-10-24 21:02:50,889 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1986415935_0001_r_000000_0' to hdfs://localhost:9000/ahan/Prac5/
Output/new
2024-10-24 21:02:50,890 INFO mapred.LocalJobRunner: reduce > reduce
2024-10-24 21:02:50,890 INFO mapred.Task: Task 'attempt_local1986415935_0001_r_000000_0' done.
2024-10-24 21:02:50,890 INFO mapred.Task: Final Counters for attempt_local1986415935_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=5199
                FILE: Number of bytes written=558394
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=121
                HDFS: Number of bytes written=25
                HDFS: Number of read operations=10
```

Windows PowerShell

```
                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=4
                Reduce shuffle bytes=47
                Reduce input records=4
                Reduce output records=4
                Spilled Records=4
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=329252864
        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=25
2024-10-24 21:02:50,890 INFO mapred.LocalJobRunner: Finishing task: attempt_local1986415935_0001_r_000000_0
2024-10-24 21:02:50,892 INFO mapred.LocalJobRunner: Starting task: attempt_local1986415935_0001_r_000001_0
2024-10-24 21:02:50,893 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-10-24 21:02:50,893 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup fa
ilures: false
2024-10-24 21:02:50,893 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2024-10-24 21:02:50,924 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@33089255
2024-10-24 21:02:50,924 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7bfc687a
2024-10-24 21:02:50,924 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-10-24 21:02:50,926 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSort
Factor=10, memToMemMergeOutputsThreshold=10
2024-10-24 21:02:50,926 INFO reduce.EventFetcher: attempt_local1986415935_0001_r_000001_0 Thread started: EventFetcher for fetching Map Completion Events
2024-10-24 21:02:50,930 INFO reduce.LocalFetcher: localfetcher#2 about to shuffle output of map attempt_local1986415935_0001_m_000000_0 decomp: 151 len: 155
 to MEMORY
2024-10-24 21:02:50,930 INFO reduce.InMemoryMapOutput: Read 151 bytes from map-output for attempt_local1986415935_0001_m_000000_0
2024-10-24 21:02:50,931 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 151, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedM
emory ->151
```


Windows PowerShell
2024-10-24 21:02:50,931 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,932 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2024-10-24 21:02:50,937 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:50,937 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 144 bytes
2024-10-24 21:02:50,940 INFO reduce.MergeManagerImpl: Merged 1 segments, 151 bytes to disk to satisfy reduce memory limit
2024-10-24 21:02:50,942 INFO reduce.MergeManagerImpl: Merging 1 files, 155 bytes from disk
2024-10-24 21:02:50,942 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2024-10-24 21:02:50,942 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:50,943 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 144 bytes
2024-10-24 21:02:50,944 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,968 INFO mapred.Task: Task:attempt_local1986415935_0001_r_000001_0 is done. And is in the process of committing
2024-10-24 21:02:50,970 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:50,970 INFO mapred.Task: Task attempt_local1986415935_0001_r_000001_0 is allowed to commit now
2024-10-24 21:02:50,979 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1986415935_0001_r_000001_0' to hdfs://localhost:9000/ahan/Prac5/
Output/new
2024-10-24 21:02:50,979 INFO mapred.LocalJobRunner: reduce > reduce
2024-10-24 21:02:50,980 INFO mapred.Task: Task 'attempt_local1986415935_0001_r_000001_0' done.
2024-10-24 21:02:50,980 INFO mapred.Task: Final Counters for attempt_local1986415935_0001_r_000001_0: Counters: 30
File System Counters
FILE: Number of bytes read=5669
FILE: Number of bytes written=558549
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=121
HDFS: Number of bytes written=100
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=5
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=11
Reduce shuffle bytes=155
Reduce input records=14
Reduce output records=11
Spilled Records=14
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1


Windows PowerShell
GC time elapsed (ms)=0
Total committed heap usage (bytes)=329252864
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=75
2024-10-24 21:02:50,980 INFO mapred.LocalJobRunner: Finishing task: attempt_local1986415935_0001_r_000001_0
2024-10-24 21:02:50,980 INFO mapred.LocalJobRunner: Starting task: attempt_local1986415935_0001_r_000002_0
2024-10-24 21:02:50,981 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-10-24 21:02:50,982 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-10-24 21:02:50,983 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2024-10-24 21:02:51,011 INFO mapred.Task: Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBasedProcessTree@5ab85ffd
2024-10-24 21:02:51,011 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@421f4368
2024-10-24 21:02:51,012 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-10-24 21:02:51,013 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2024-10-24 21:02:51,013 INFO reduce.EventFetcher: attempt_local1986415935_0001_r_000002_0 Thread started: EventFetcher for fetching Map Completion Events
2024-10-24 21:02:51,017 INFO reduce.LocalFetcher: localfetcher#3 about to shuffle output of map attempt_local1986415935_0001_m_000000_0 decomp: 76 len: 80 to MEMORY
2024-10-24 21:02:51,018 INFO reduce.InMemoryMapOutput: Read 76 bytes from map-output for attempt_local1986415935_0001_m_000000_0
2024-10-24 21:02:51,018 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 76, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->76
2024-10-24 21:02:51,018 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2024-10-24 21:02:51,018 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:51,018 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2024-10-24 21:02:51,024 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:51,024 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 64 bytes
2024-10-24 21:02:51,026 INFO reduce.MergeManagerImpl: Merged 1 segments, 76 bytes to disk to satisfy reduce memory limit
2024-10-24 21:02:51,027 INFO reduce.MergeManagerImpl: Merging 1 files, 80 bytes from disk
2024-10-24 21:02:51,027 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2024-10-24 21:02:51,027 INFO mapred.Merger: Merging 1 sorted segments
2024-10-24 21:02:51,028 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 64 bytes
2024-10-24 21:02:51,029 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-10-24 21:02:51,049 INFO mapred.Task: Task:attempt_local1986415935_0001_r_000002_0 is done. And is in the process of committing
2024-10-24 21:02:51,052 INFO mapred.LocalJobRunner: 1 / 1 copied.


Windows PowerShell
2024-10-24 21:02:51,059 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1986415935_0001_r_000002_0' to hdfs://localhost:9000/ahan/Prac5/Output/new
2024-10-24 21:02:51,060 INFO mapred.LocalJobRunner: reduce > reduce
2024-10-24 21:02:51,060 INFO mapred.Task: Task 'attempt_local1986415935_0001_r_000002_0' done.
2024-10-24 21:02:51,061 INFO mapred.Task: Final Counters for attempt_local1986415935_0001_r_000002_0: Counters: 30
File System Counters
FILE: Number of bytes read=5909
FILE: Number of bytes written=558629
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=121
HDFS: Number of bytes written=150
HDFS: Number of read operations=20
HDFS: Number of large read operations=0
HDFS: Number of write operations=7
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=6
Reduce shuffle bytes=80
Reduce input records=6
Reduce output records=6
Spilled Records=6
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=329252864
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=50
2024-10-24 21:02:51,061 INFO mapred.LocalJobRunner: Finishing task: attempt_local1986415935_0001_r_000002_0
2024-10-24 21:02:51,061 INFO mapred.LocalJobRunner: reduce task executor complete.