

Activities Text Editor Apr 22 14:16 reducer.py mapper.py

```
1 import sys
2 import io
3 current_word=None
4 current_count=0
5 word=None
6 for line in sys.stdin:
7     word,count=line.split('\t',1)
8     count=int(count)
9     if current_word==word:
10         current_count+=count
11     else:
12         if current_word:
13             print("%s\t%s"%(current_word,current_count))
14             current_count=count
15             current_word=word
16 if current_word==word:
17     print("%s\t%s"%(current_word,current_count))
```

Activities Text Editor Apr 22 14:16 mapper.py ~/Desktop/3025 reducer.py mapper.py

```
1
2 import sys
3 import io
4 input_stream=io.TextIOWrapper(sys.stdin.buffer)
5
6 for line in input_stream:
7     words=line.split()
8     for word in words:
9         print("%s \t %s"%(word,1))
```

```
Activities Terminal Apr 22 14:11 hduser@student-ThinkCentre-M700: ~/Desktop/3025

For more details about these options:
Use $HADOOP_PREFIX/bin/hadoop jar hadoop-streaming.jar -info

Try -help for more information
Streaming Command Failed!
hduser@student-ThinkCentre-M700:~/Desktop/3025$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.0.jar -input /3025/test.txt -output /3025/pyhadoop/ -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
25/04/22 14:07:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/04/22 14:07:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
25/04/22 14:07:42 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
25/04/22 14:07:42 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
25/04/22 14:07:42 INFO mapred.FileInputFormat: Total input files to process : 1
25/04/22 14:07:42 INFO mapreduce.JobSubmitter: number of splits:1
25/04/22 14:07:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2109197348_0001
25/04/22 14:07:42 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
25/04/22 14:07:42 INFO mapreduce.Job: Running job: job_local2109197348_0001
25/04/22 14:07:42 INFO mapred.LocalJobRunner: OutputCommitter set in config null
25/04/22 14:07:42 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
25/04/22 14:07:42 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
25/04/22 14:07:42 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
25/04/22 14:07:42 INFO mapred.LocalJobRunner: Waiting for map tasks
25/04/22 14:07:42 INFO mapred.LocalJobRunner: Starting task: attempt_local2109197348_0001_m_000000_0
25/04/22 14:07:42 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
25/04/22 14:07:42 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
25/04/22 14:07:42 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
```

```
Activities Terminal Apr 22 14:10 hduser@student-ThinkCentre-M700: ~/Desktop/3025

Reduce input records=268
Reduce output records=7
Spilled Records=536
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=594542592

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=1465
File Output Format Counters
Bytes Written=74
25/04/22 14:07:43 INFO streaming.StreamJob: Output directory: /3025/pyhadoop/
hduser@student-ThinkCentre-M700:~/Desktop/3025$ hdfs dfs -cat /3025/pyhadoop/*
25/04/22 14:09:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bye 19
hello 57
ok 57
shubham 59
sk 40
skshubham 17
world 19
hduser@student-ThinkCentre-M700:~/Desktop/3025$
```