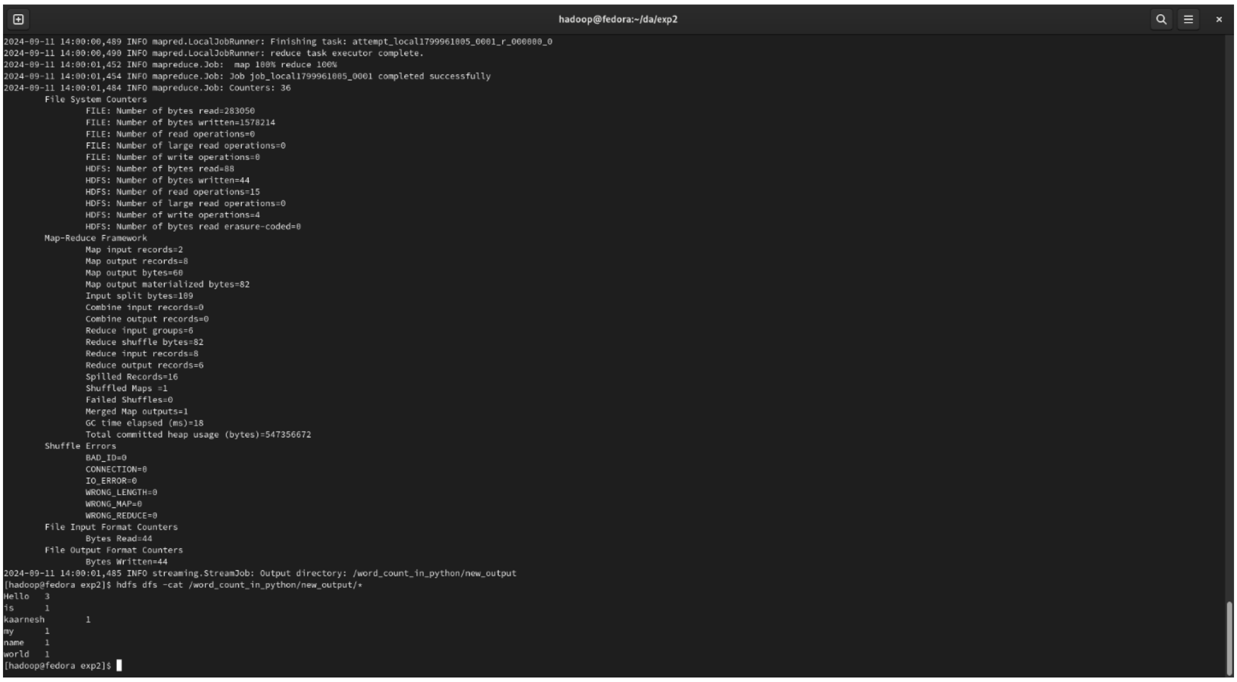
EXP NO: 2 RUN A BASIC WORD COUNT MAP REDUCE PROGRAM TO UNDERSTAND MAP REDUCE PARADIGM

$mkdir DA-Lab
$cd DA-Lab
$mkdir exp2
$cd exp2

$nano word_count.txt

$nano mapper.py

$nano reducer.py



```
2024-09-11 14:00:00,489 INFO mapred.LocalJobRunner: Finishing task: attempt_local1799961005_0001_r_000000_0
2024-09-11 14:00:00,490 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-09-11 14:00:01,452 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-11 14:00:01,454 INFO mapreduce.Job: Job job_local1799961005_0001 completed successfully
2024-09-11 14:00:01,484 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=283050
                FILE: Number of bytes written=1578214
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=88
                HDFS: Number of bytes written=44
                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=2
                Map output records=8
                Map output bytes=60
                Map output materialized bytes=82
                Input split bytes=109
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=82
                Reduce input records=8
                Reduce output records=6
                Spilled Records=16
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=18
                Total committed heap usage (bytes)=547356672
        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=44
        File Output Format Counters
                Bytes Written=44
2024-09-11 14:00:01,485 INFO streaming.StreamJob: Output directory: /word_count_in_python/new_output
[hadoop@fedora exp2]$ hdfs dfs -cat /word_count_in_python/new_output/*
Hello   3
is      1
kaarnesh        1
my      1
name    1
world   1
[hadoop@fedora exp2]$ 
```

$start-all.sh

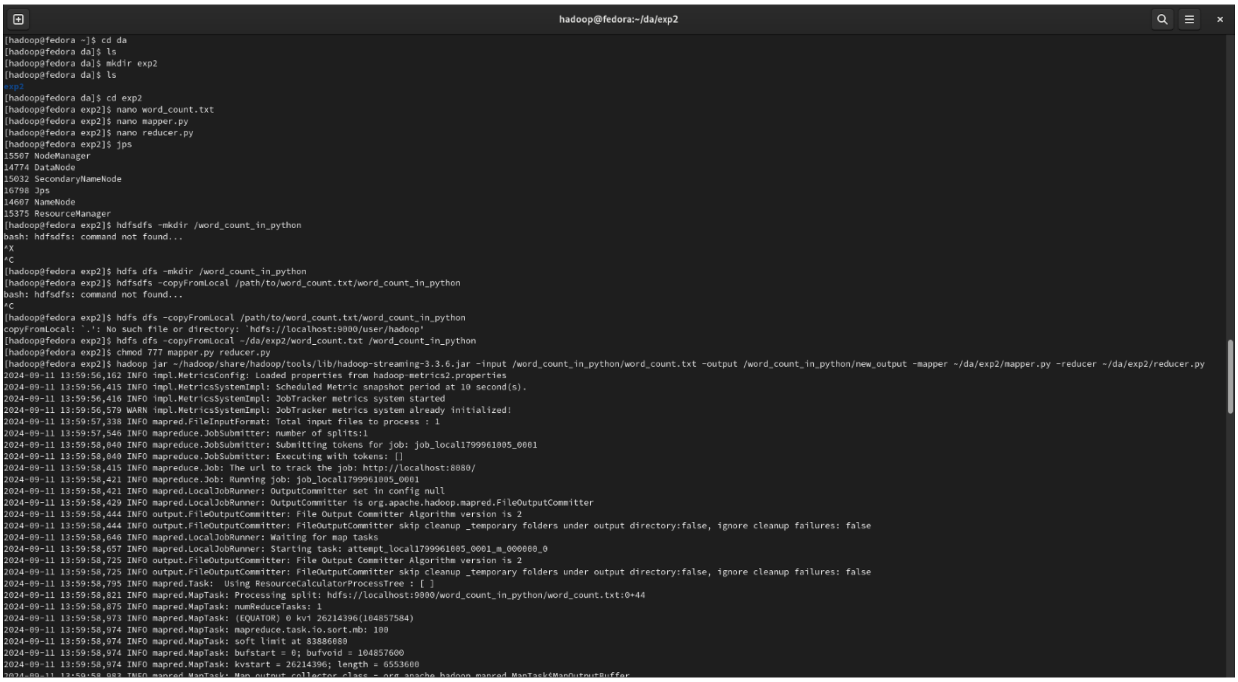
$ jps

$hdfs dfs -mkdir /exp2

$hdfs dfs -copyFromLocal ~/DA-Lab/exp2/word_count.txt /exp2

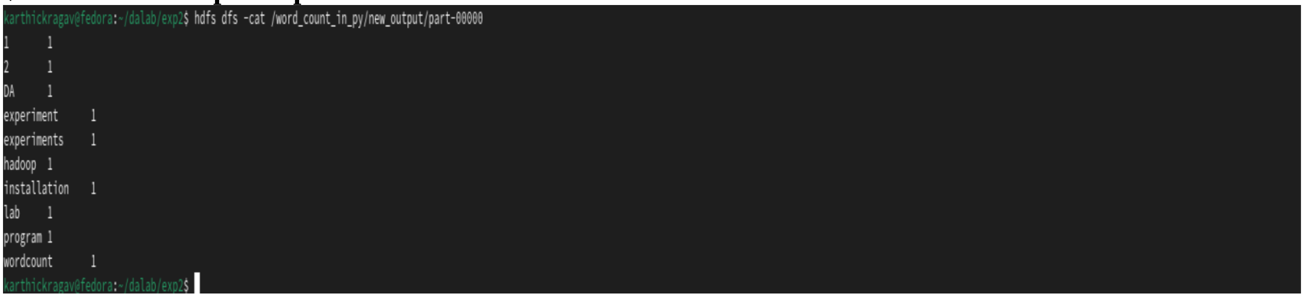
$chmod 777 mapper.py reducer.py

$hadoop jar $HADOOP_STREAMING -input /exp2/word_count.txt -output /exp2/output -mapper ~/DA-Lab/exp2/mapper.py -reducer ~/DA-Lab/exp2/reducer.py



```
[hadoop@fedora ~]$ cd da
[hadoop@fedora da]$ ls
[hadoop@fedora da]$ mkdir exp2
[hadoop@fedora da]$ ls
exp2
[hadoop@fedora da]$ cd exp2
[hadoop@fedora exp2]$ nano word_count.txt
[hadoop@fedora exp2]$ nano mapper.py
[hadoop@fedora exp2]$ nano reducer.py
[hadoop@fedora exp2]$ jps
15507 NodeManager
14774 DataNode
15032 SecondaryNameNode
16798 Jps
14607 NameNode
15375 ResourceManager
[hadoop@fedora exp2]$ hdfsdfs -mkdir /word_count_in_python
bash: hdfsdfs: command not found...
^X
^C
[hadoop@fedora exp2]$ hdfs dfs -mkdir /word_count_in_python
[hadoop@fedora exp2]$ hdfsdfs -copyFromLocal /path/to/word_count.txt/word_count_in_python
bash: hdfsdfs: command not found...
^C
[hadoop@fedora exp2]$ hdfs dfs -copyFromLocal /path/to/word_count.txt/word_count_in_python
copyFromLocal: `.': No such file or directory: `hdfs://localhost:9000/user/hadoop'
[hadoop@fedora exp2]$ hdfs dfs -copyFromLocal ~/da/exp2/word_count.txt /word_count_in_python
[hadoop@fedora exp2]$ chmod 777 mapper.py reducer.py
[hadoop@fedora exp2]$ hadoop jar ~/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -input /word_count_in_python/word_count.txt -output /word_count_in_python/new_output -mapper ~/da/exp2/mapper.py -reducer ~/da/exp2/reducer.py
2024-09-11 13:59:56,162 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-09-11 13:59:56,415 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-09-11 13:59:56,416 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-09-11 13:59:56,579 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-09-11 13:59:57,338 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-11 13:59:57,546 INFO mapreduce.JobSubmitter: number of splits:1
2024-09-11 13:59:58,040 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1799961005_0001
2024-09-11 13:59:58,040 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-11 13:59:58,415 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-09-11 13:59:58,421 INFO mapreduce.Job: Running job: job_local1799961005_0001
2024-09-11 13:59:58,421 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-09-11 13:59:58,429 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-09-11 13:59:58,444 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-11 13:59:58,444 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-11 13:59:58,646 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-09-11 13:59:58,657 INFO mapred.LocalJobRunner: Starting task: attempt_local1799961005_0001_m_000000_0
2024-09-11 13:59:58,725 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-09-11 13:59:58,725 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-09-11 13:59:58,795 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-09-11 13:59:58,821 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/word_count_in_python/word_count.txt:0+44
2024-09-11 13:59:58,875 INFO mapred.MapTask: numReduceTasks: 1
2024-09-11 13:59:58,973 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-09-11 13:59:58,974 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-09-11 13:59:58,974 INFO mapred.MapTask: soft limit at 83886080
2024-09-11 13:59:58,974 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-09-11 13:59:58,974 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-09-11 13:59:58,983 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
```

$hdfs dfs -cat /exp2/output/*



```
karthickragav@fedora:~/dalab/exp2$ hdfs dfs -cat /word_count_in_py/new_output/part-00000
1       1
2       1
DA      1
experiment      1
experiments     1
hadoop  1
installation    1
lab     1
program 1
wordcount       1
karthickragav@fedora:~/dalab/exp2$
```