

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

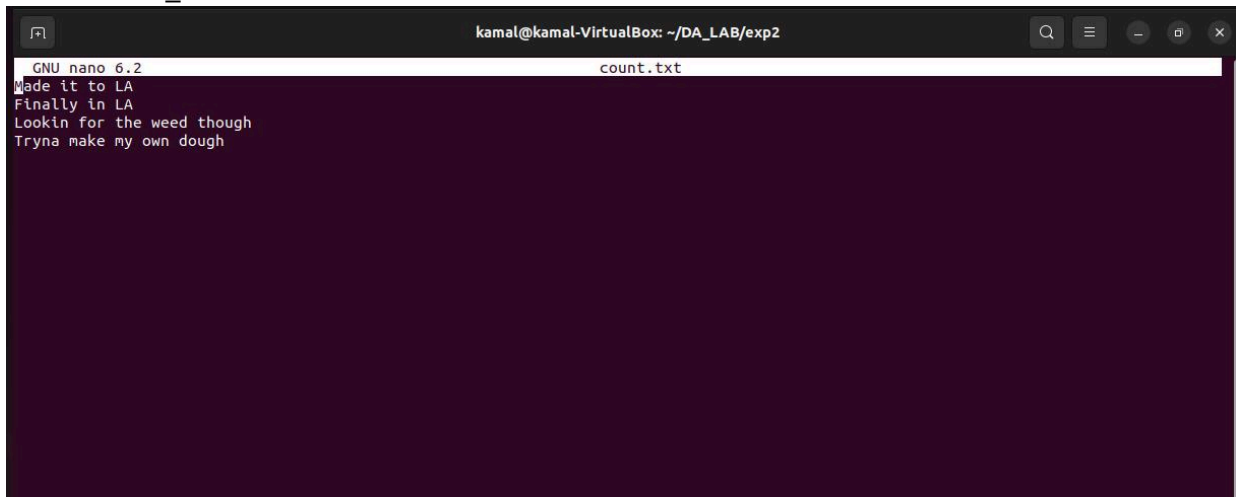
\$mkdir DA-Lab

\$cd DA-Lab

\$mkdir exp2

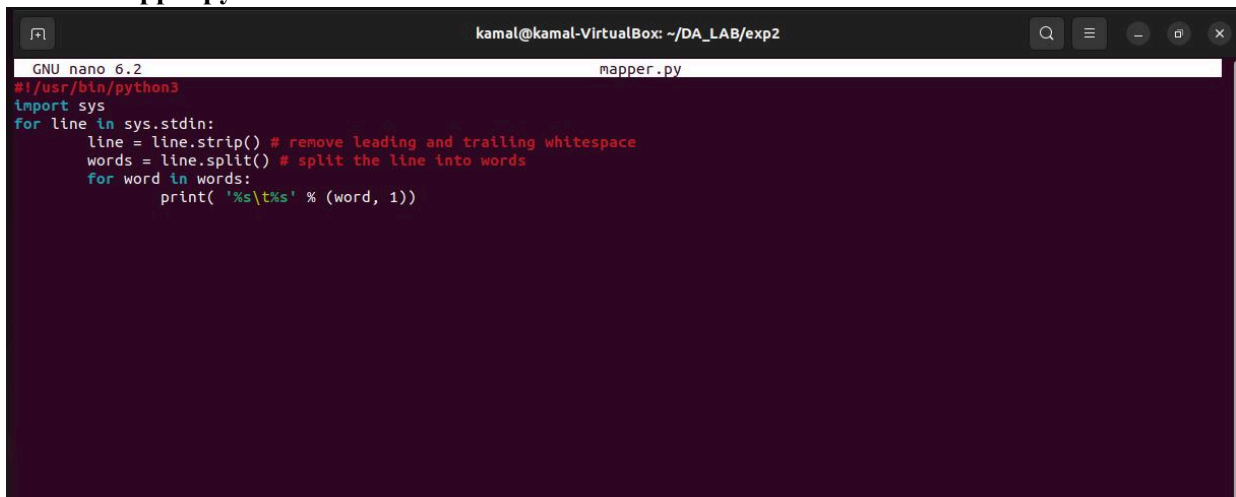
\$cd exp2

\$nano word_count.txt




```
kamal@kamal-VirtualBox: ~/DA_LAB/exp2
GNU nano 6.2 count.txt
Made it to LA
Finally in LA
Lookin for the weed though
Tryna make my own dough
```

\$nano mapper.py



```
kamal@kamal-VirtualBox: ~/DA_LAB/exp2
GNU nano 6.2 mapper.py
#!/usr/bin/python3
import sys
for line in sys.stdin:
    line = line.strip() # remove leading and trailing whitespace
    words = line.split() # split the line into words
    for word in words:
        print( '%s\t%s' % (word, 1))
```

\$nano reducer.py



```
kamal@kamal-VirtualBox: ~/DA_LAB/exp2
GNU nano 6.2 reducer.py
#!/usr/bin/python3
from operator import itemgetter
import sys
current_word = None
current_count = 0
word = None
for line in sys.stdin:
    line = line.strip()
    word, count = line.split('\t', 1)
    try:
        count = int(count)
    except ValueError:
        continue
    if current_word == word:
        current_count += count
    else:
        if current_word:
            print( '%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word
if current_word == word:
    print( '%s\t%s' % (current_word, current_count))
```

\$start-all.sh

```
kamal@kamal-VirtualBox:~/DA_LAB/exp2$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kamal in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [kamal-VirtualBox]
Starting resourcemanager
Starting nodemanagers
```

\$jps

```
Starting nodemanagers
kamal@kamal-VirtualBox:~/DA_LAB/exp2$ jps
4916 DataNode
5828 Jps
5333 ResourceManager
4790 NameNode
5112 SecondaryNameNode
5455 NodeManager
```

\$hdfs dfs -mkdir /exp2

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

```
karthickragav@fedora:~/dalab/exp2$ hdfs dfs -ls /word_count_in_py
Found 2 items
-rw-r--r-- 1 karthickragav supergroup 83 2024-09-01 21:13 /word_count_in_py/count.txt
drwxr-xr-x - karthickragav supergroup 0 2024-09-01 21:14 /word_count_in_py/new_output
karthickragav@fedora:~/dalab/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

```

Activities Terminal Sep 2 13:40
kamal@kamal-VirtualBox: ~/DA_LAB/exp3

kamal@kamal-VirtualBox:~$ ls
DA_LAB  Documents  hadoop-3.4.0  Pictures  snap  Videos
kamal@kamal-VirtualBox:~$ cd DA_LAB
kamal@kamal-VirtualBox:~/DA_LAB$ ls
exp2  exp3
kamal@kamal-VirtualBox:~/DA_LAB$ cd exp3
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ nano exp3
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ nano mapper.py
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ nano reducer.py
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as kamal in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [kamal-VirtualBox]
Starting resourcemanager
Starting nodemanagers
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ chmod 777 mapper.py reducer.py
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ jps
4993 Jps
4531 ResourceManager
4358 SecondaryNameNode
4647 NodeManager
4186 DataNode
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as kamal in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [kamal-VirtualBox]
Stopping nodemanagers
Stopping resourcemanager
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ nano hadoop/etc/hadoop/hdfs-site.xml
kamal@kamal-VirtualBox:~/DA_LAB/exp3$ cd
kamal@kamal-VirtualBox:~$ cd hadoop/etc/hadoop
kamal@kamal-VirtualBox:~$ cd hadoop-3.4.0/etc/hadoop

```

\$hdfs dfs -cat /exp2/output/*

```
kamal@kamal-VirtualBox: ~  
Physical memory (bytes) snapshot=775946240  
Virtual memory (bytes) snapshot=7596027904  
Total committed heap usage (bytes)=634388480  
Peak Map Physical memory (bytes)=284135424  
Peak Map Virtual memory (bytes)=2530283520  
Peak Reduce Physical memory (bytes)=211079168  
Peak Reduce Virtual memory (bytes)=2536181760  
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=122  
File Output Format Counters  
Bytes Written=108  
2024-09-01 23:20:01,320 INFO streaming.StreamJob: Output directory: /word_count_in_python/new_output  
kamal@kamal-VirtualBox:~$ hdfs dfs -cat /word_count_in_python/new_output/part-*  
Finally 1  
LA 1  
Lookin 1  
Made 1  
Tryna 1  
dough 1  
for 1  
in 1  
it 1  
make 1  
my 1  
own 1  
the 1  
the 1  
though 1  
to 1  
weed 1  
kamal@kamal-VirtualBox:~$
```