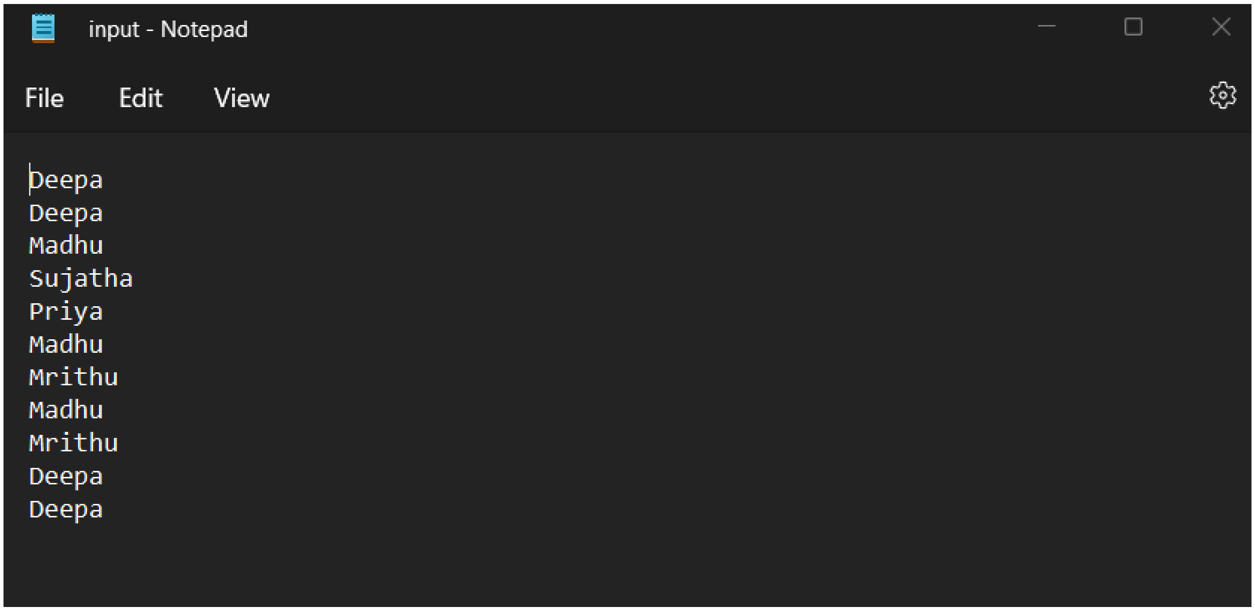
EXP 2: Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm

AIM:

To run a basic Word Count MapReduce program using Hadoop.

PROCEDURE:

Step 1: Create Data File: Create a file named "input.txt" and populate it with text data that you wish to analyse.



```
Deepa
Deepa
Madhu
Sujatha
Priya
Madhu
Mrithu
Madhu
Mrithu
Deepa
Deepa
```

Step 2: Mapper Logic - mapper.py:

Create a file named "mapper.py" to implement the logic for the mapper. The mapper will read input data from STDIN, split lines into words, and output each word with its count.

mapper.py:

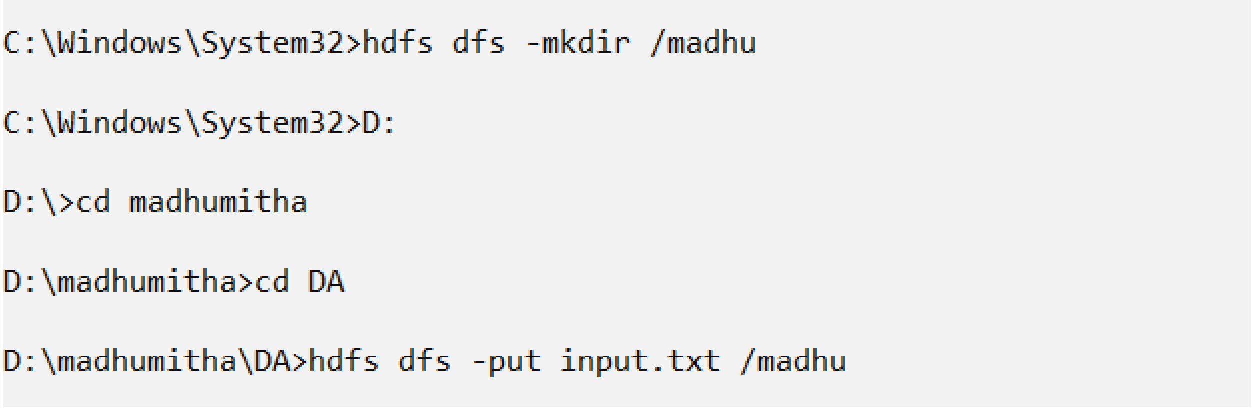
```
#!/usr/bin/python
import sys
sys.stderr.write("Logging info: Mapper started\n")
for line in sys. stdin:
        line=line.strip()
        words=line.split()
        for word in words:
                print ('%s\t%s'%(word, 1))
```

Step 3: Reducer Logic - reducer.py: Create a file named "reducer.py" to implement the logic for the reducer. The reducer will aggregate the occurrences of each word and generate the final output.

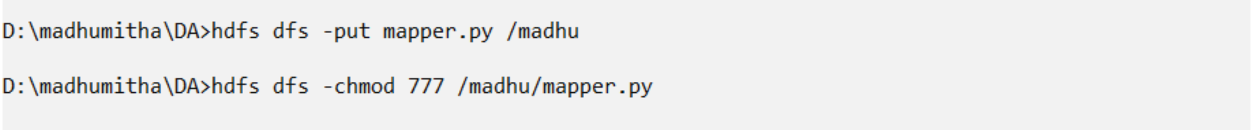
reducer.py

```
#!/usr/bin/python
import sys
sys.stderr.write("Logging info: Reducer started\n")
#sys.path.append('.')
prevw=None
prevc=0
for line in sys.stdin:
    line = line.strip()
    word, count= line.split('\t')
    count=int(count)
    if prevw == word:
        prevc += count
    else:
        if prevw:
            print ('%s\t%s' % (prevw, prevc))
        prevc = count
        prevw = word
if prevw == word:
    print ('%s\t%s'% (prevw, prevc) )
```

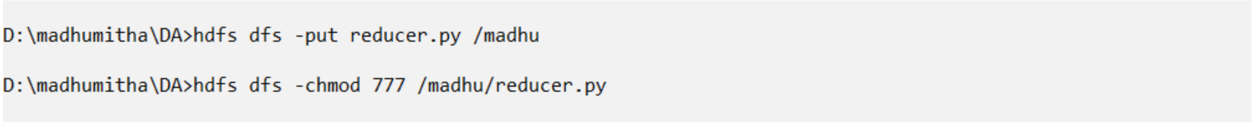
Step 4: Prepare Hadoop Environment: Start the Hadoop daemons and create a directory in HDFS to store your data. Run the following commands to store the data in the WordCount Directory.



```
C:\Windows\System32>hdfs dfs -mkdir /madhu

C:\Windows\System32>D:

D:\>cd madhumitha

D:\madhumitha>cd DA

D:\madhumitha\DA>hdfs dfs -put input.txt /madhu
```



```
D:\madhumitha\DA>hdfs dfs -put mapper.py /madhu

D:\madhumitha\DA>hdfs dfs -chmod 777 /madhu/mapper.py
```



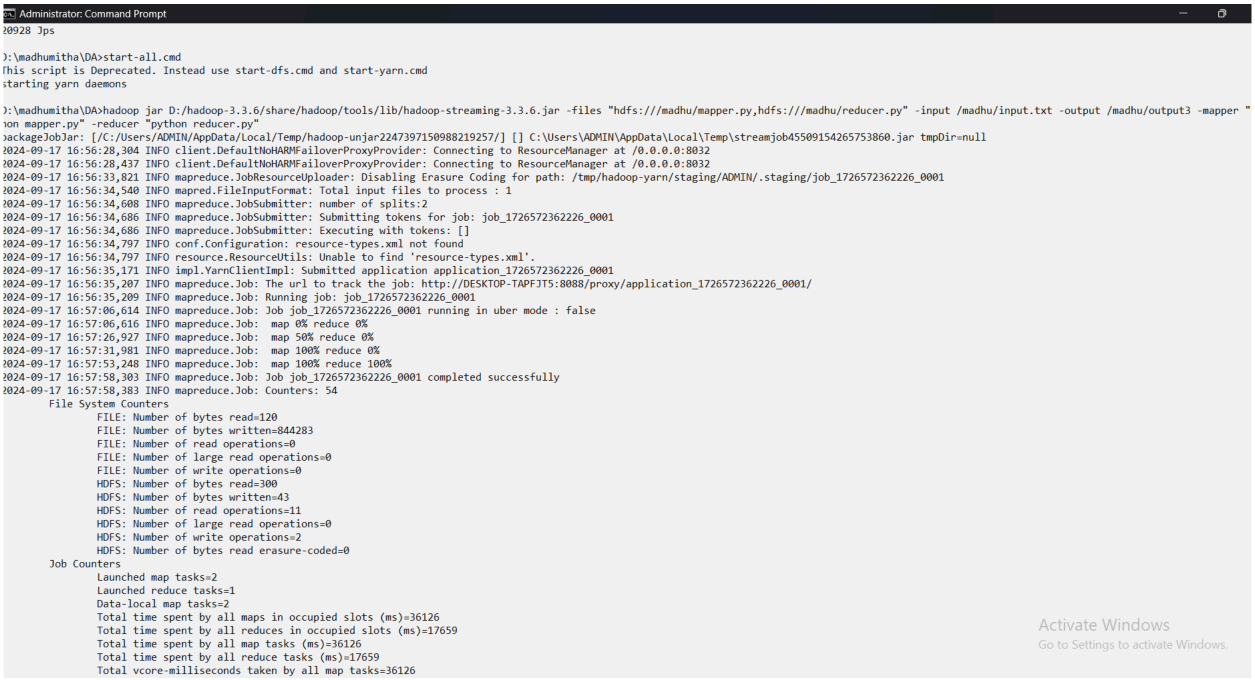
```
D:\madhumitha\DA>hdfs dfs -put reducer.py /madhu

D:\madhumitha\DA>hdfs dfs -chmod 777 /madhu/reducer.py
```

```
D:\madhumitha\DA>hadoop jar
D:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -files
"hdfs:///madhu/mapper.py,hdfs:///madhu/reducer.py" -input /madhu/input.txt
-output /madhu/output3 -mapper "python mapper.py" -reducer "python reducer.py"
```

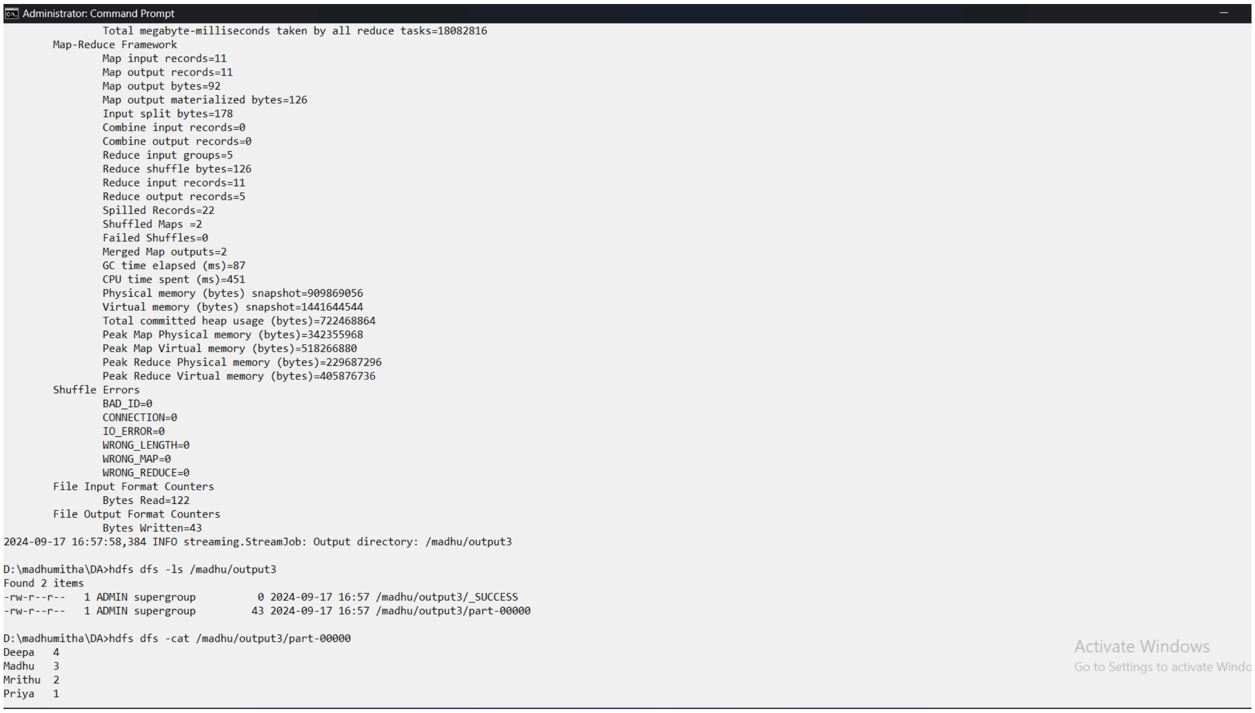
Step 5: Check Output:
Check the output of the Word Count program in the specified HDFS output directory.

```
hdfs dfs -cat /madhu/output3/part-00000
```

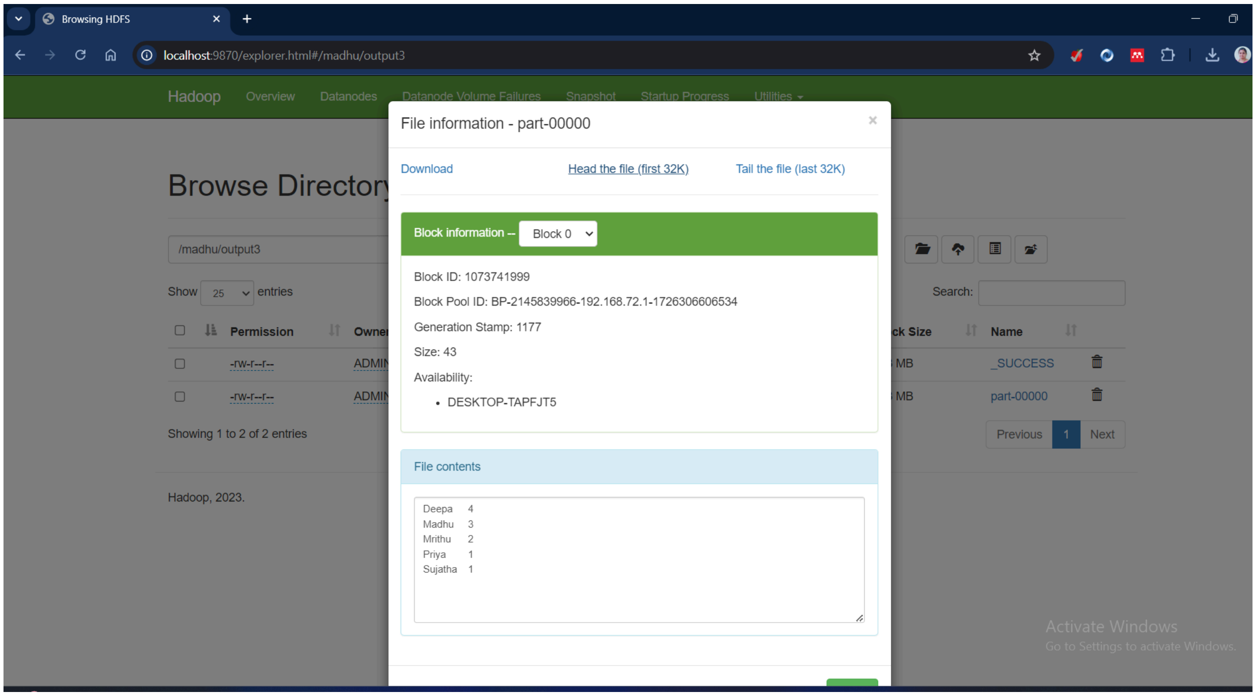
OUTPUT:



```
20928 Jps

):\madhumitha\DA>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

):\madhumitha\DA>hadoop jar D:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -files "hdfs:///madhu/mapper.py,hdfs:///madhu/reducer.py" -input /madhu/input.txt -output /madhu/output3 -mapper "
hon mapper.py" -reducer "python reducer.py"
packageJobJar: [/C:/Users/ADMIN/AppData/Local/Temp/hadoop-unjar2247397150988219257/] [] C:\Users\ADMIN\AppData\Local\Temp\streamjob45509154265753860.jar tmpDir=null
2024-09-17 16:56:28,304 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-17 16:56:28,437 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-17 16:56:33,821 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ADMIN/.staging/job_1726572362226_0001
2024-09-17 16:56:34,540 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-17 16:56:34,608 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-17 16:56:34,686 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1726572362226_0001
2024-09-17 16:56:34,686 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-17 16:56:34,797 INFO conf.Configuration: resource-types.xml not found
2024-09-17 16:56:34,797 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-17 16:56:35,171 INFO impl.YarnClientImpl: Submitted application application_1726572362226_0001
2024-09-17 16:56:35,207 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TAPFJT5:8088/proxy/application_1726572362226_0001/
2024-09-17 16:56:35,209 INFO mapreduce.Job: Running job: job_1726572362226_0001
2024-09-17 16:57:06,614 INFO mapreduce.Job: Job job_1726572362226_0001 running in uber mode : false
2024-09-17 16:57:06,616 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-17 16:57:26,927 INFO mapreduce.Job:  map 50% reduce 0%
2024-09-17 16:57:31,981 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-17 16:57:53,248 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-17 16:57:58,303 INFO mapreduce.Job: Job job_1726572362226_0001 completed successfully
2024-09-17 16:57:58,383 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=120
                FILE: Number of bytes written=844283
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=300
                HDFS: Number of bytes written=43
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36126
                Total time spent by all reduces in occupied slots (ms)=17659
                Total time spent by all map tasks (ms)=36126
                Total time spent by all reduce tasks (ms)=17659
                Total vcore-milliseconds taken by all map tasks=36126
```



```
                Total megabyte-milliseconds taken by all reduce tasks=18082816
        Map-Reduce Framework
                Map input records=11
                Map output records=11
                Map output bytes=92
                Map output materialized bytes=126
                Input split bytes=178
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=126
                Reduce input records=11
                Reduce output records=5
                Spilled Records=22
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=87
                CPU time spent (ms)=451
                Physical memory (bytes) snapshot=909869056
                Virtual memory (bytes) snapshot=1441644544
                Total committed heap usage (bytes)=722468864
                Peak Map Physical memory (bytes)=342355968
                Peak Map Virtual memory (bytes)=518266880
                Peak Reduce Physical memory (bytes)=229687296
                Peak Reduce Virtual memory (bytes)=405876736
        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=43
2024-09-17 16:57:58,384 INFO streaming.StreamJob: Output directory: /madhu/output3

D:\madhumitha\DA>hdfs dfs -ls /madhu/output3
Found 2 items
-rw-r--r--   1 ADMIN supergroup          0 2024-09-17 16:57 /madhu/output3/_SUCCESS
-rw-r--r--   1 ADMIN supergroup         43 2024-09-17 16:57 /madhu/output3/part-00000

D:\madhumitha\DA>hdfs dfs -cat /madhu/output3/part-00000
Deepa   4
Madhu   3
Mrithu  2
Priya   1
```



File information - part-00000

Block information -- Block 0

Block ID: 1073741999

Block Pool ID: BP-2145839966-192.168.72.1-1726306606534

Generation Stamp: 1177

Size: 43

Availability:

- DESKTOP-TAPFJT5

File contents

```
Deepa    4
Madhu    3
Mrithu   2
Priya    1
Sujatha  1
```

RESULT:

Thus, the program for basic Word Count Map Reduce has been executed successfully.