EX NO: 2 **Implement word count/frequency using mapreduce**

Aim:

To run a Word Count MapReduce program.

Procedure:

Prepare the data file.

The data file contains words which are repeated.

Create program **mapper.py**

```
import sys
for line in sys.stdin:
    line=line.strip()
    words=line.split()
    for word in words:
        print('%s\t%s' % (word,1))
```

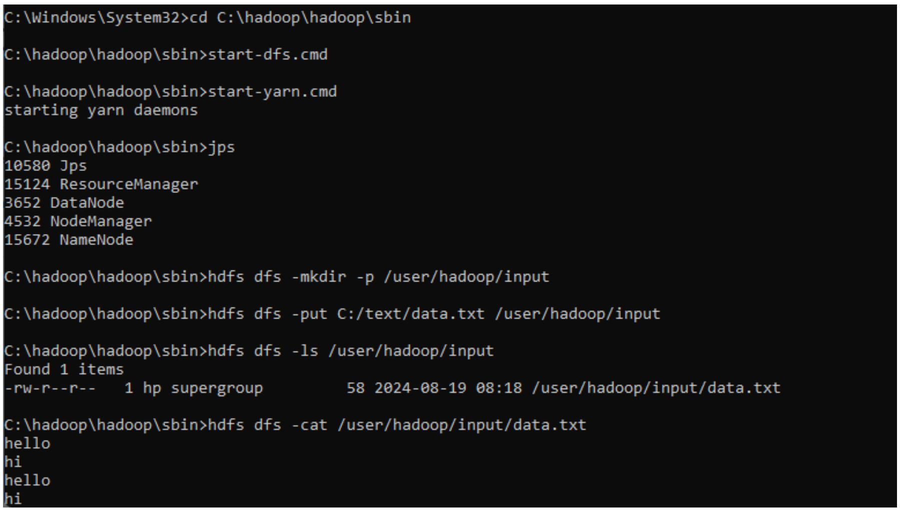
Create program **reducer.py**

```
import sys
prev_word=None
prev_count=0
for line in sys.stdin:
    line=line.strip()
    word,count=line.split('\t')
    count=int(count)
    if prev_word==word:
        prev_count+=count
    else:
        if prev_word:
            print('%s\t%s' % (prev_word, prev_count))
        prev_word=word
        prev_count=count
if prev_word==word:
```

    print('%s\t%s' % (prev_word, prev_count))

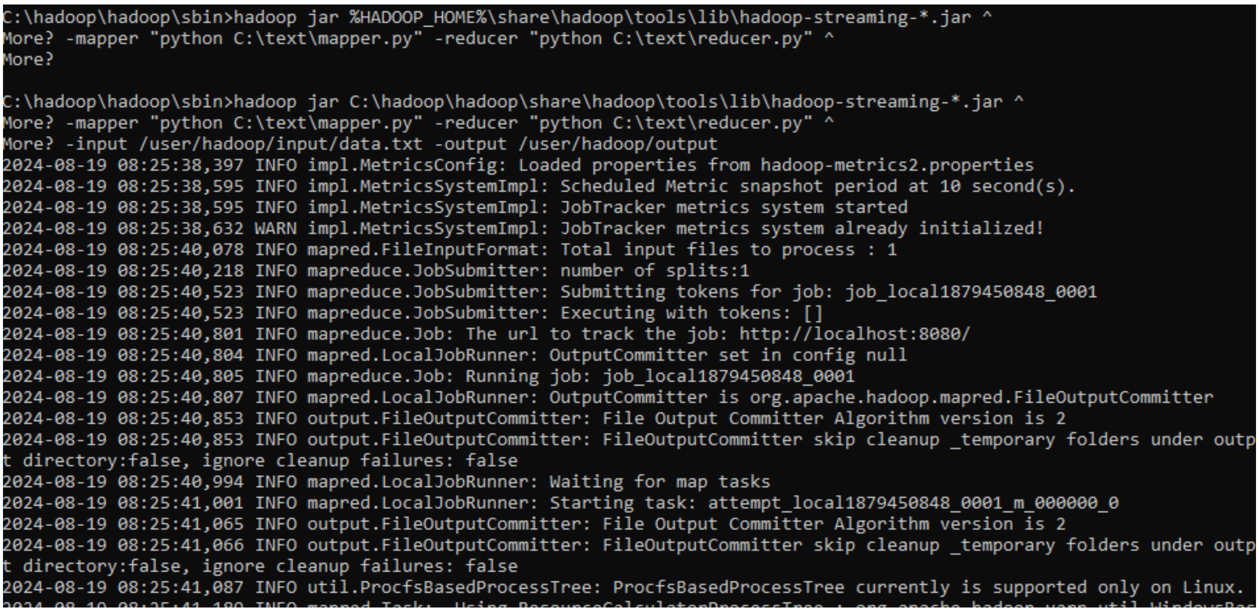
Start the services

Make a directory, put the text file inside it.

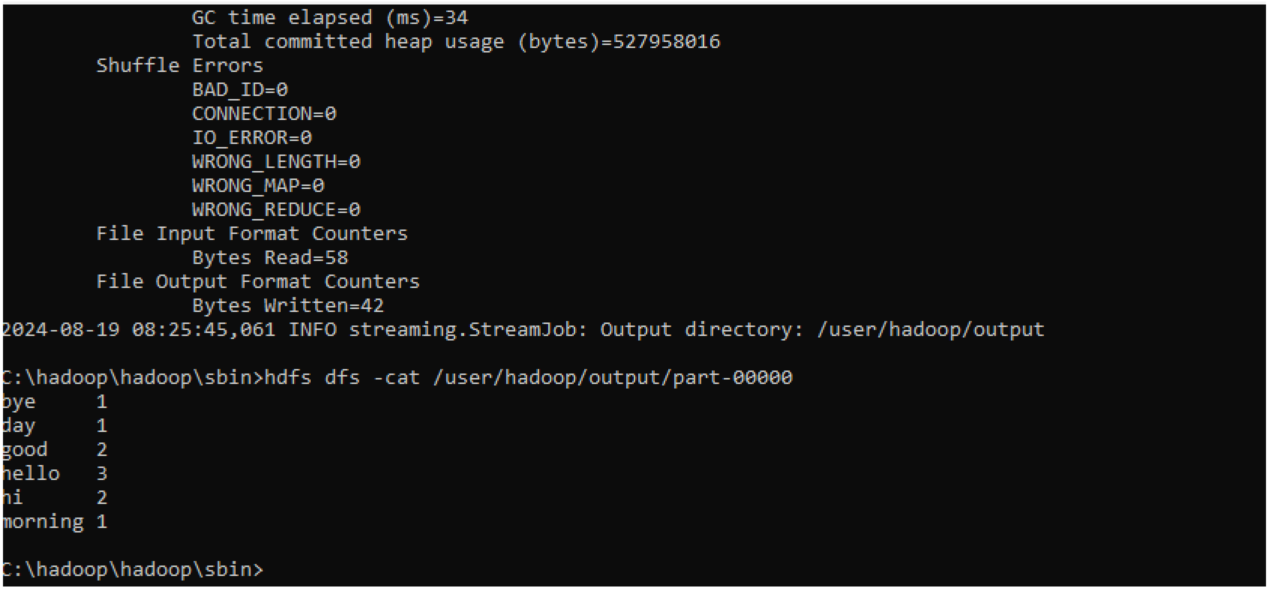


```
C:\Windows\System32>cd C:\hadoop\hadoop\sbin

C:\hadoop\hadoop\sbin>start-dfs.cmd

C:\hadoop\hadoop\sbin>start-yarn.cmd
starting yarn daemons

C:\hadoop\hadoop\sbin>jps
10580 Jps
15124 ResourceManager
3652 DataNode
4532 NodeManager
15672 NameNode

C:\hadoop\hadoop\sbin>hdfs dfs -mkdir -p /user/hadoop/input

C:\hadoop\hadoop\sbin>hdfs dfs -put C:/text/data.txt /user/hadoop/input

C:\hadoop\hadoop\sbin>hdfs dfs -ls /user/hadoop/input
Found 1 items
-rw-r--r--   1 hp supergroup         58 2024-08-19 08:18 /user/hadoop/input/data.txt

C:\hadoop\hadoop\sbin>hdfs dfs -cat /user/hadoop/input/data.txt
hello
hi
hello
hi
```

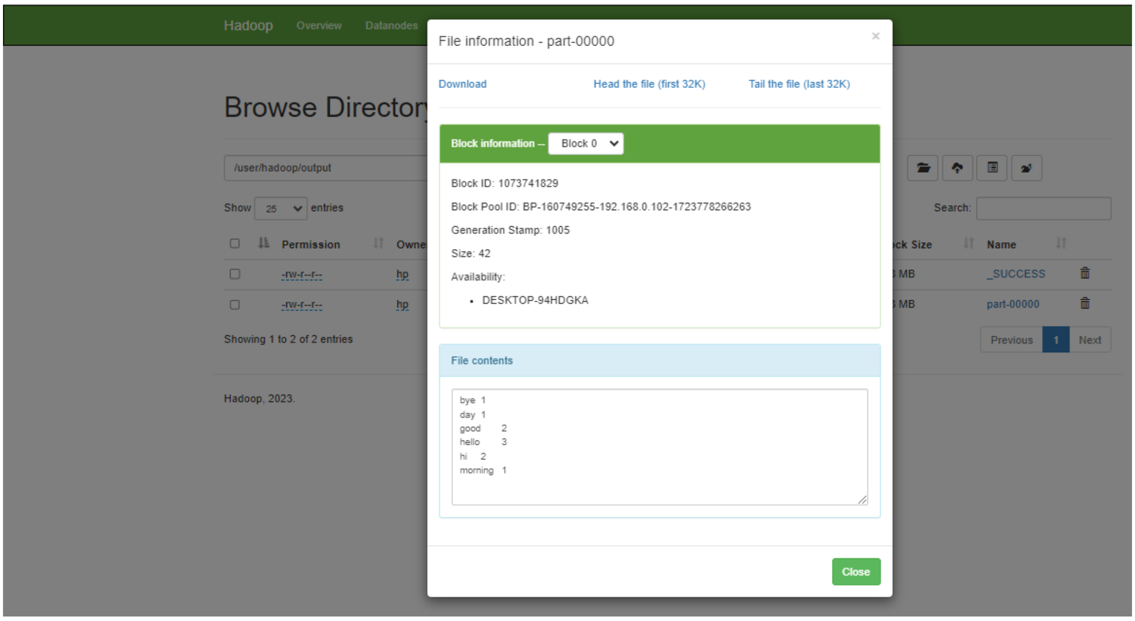
Run the MapReduce program in hadoop environment:



```
C:\hadoop\hadoop\sbin>hadoop jar %HADOOP_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
More? -mapper "python C:\text\mapper.py" -reducer "python C:\text\reducer.py" ^
More?

C:\hadoop\hadoop\sbin>hadoop jar C:\hadoop\hadoop\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
More? -mapper "python C:\text\mapper.py" -reducer "python C:\text\reducer.py" ^
More? -input /user/hadoop/input/data.txt -output /user/hadoop/output
2024-08-19 08:25:38,397 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-08-19 08:25:38,595 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-08-19 08:25:38,595 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-08-19 08:25:38,632 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2024-08-19 08:25:40,078 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-19 08:25:40,218 INFO mapreduce.JobSubmitter: number of splits:1
2024-08-19 08:25:40,523 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1879450848_0001
2024-08-19 08:25:40,523 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-19 08:25:40,801 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-08-19 08:25:40,804 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-08-19 08:25:40,805 INFO mapreduce.Job: Running job: job_local1879450848_0001
2024-08-19 08:25:40,807 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2024-08-19 08:25:40,853 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-08-19 08:25:40,853 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outp
t directory:false, ignore cleanup failures: false
2024-08-19 08:25:40,994 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-08-19 08:25:41,001 INFO mapred.LocalJobRunner: Starting task: attempt_local1879450848_0001_m_000000_0
2024-08-19 08:25:41,065 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-08-19 08:25:41,066 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under outp
t directory:false, ignore cleanup failures: false
2024-08-19 08:25:41,087 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2024-08-19 08:25:41,180 INFO mapred.Task:  Using ResourceCalculatorProcessTree : org.apache.hadoop.yarn.util.WindowsBas
```



```
                GC time elapsed (ms)=34
                Total committed heap usage (bytes)=527958016
        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=58
        File Output Format Counters
                Bytes Written=42
2024-08-19 08:25:45,061 INFO streaming.StreamJob: Output directory: /user/hadoop/output

C:\hadoop\hadoop\sbin>hdfs dfs -cat /user/hadoop/output/part-00000
bye     1
day     1
good    2
hello   3
hi      2
morning 1

C:\hadoop\hadoop\sbin>
```



Browse Directory

/user/hadoop/output

□	Permission	Owne	ck Size	Name
□	-rw-r--r--	hp	MB	_SUCCESS
□	-rw-r--r--	hp	MB	part-00000

Showing 1 to 2 of 2 entries

Hadoop, 2023.

File information - part-00000

Block information --

Block ID: 1073741829

Block Pool ID: BP-160749255-192.168.0.102-1723778266263

Generation Stamp: 1005

Size: 42

Availability:

- DESKTOP-94HDGKA

File contents

```
bye 1
day 1
good    2
hello   3
hi  2
morning 1
```