

**Exp .No : 9**

**Date :30.08.24**

## **DEMONSTRATE THE MAP REDUCE PROGRAMMING MODEL BY COUNTING THE NUMBER OF WORDS IN A FILE**

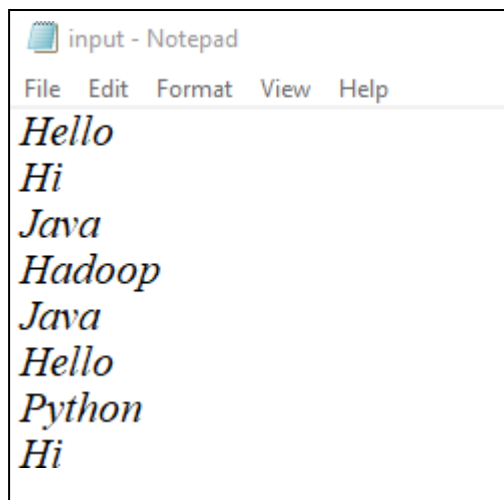
### **AIM:**

To demonstrate the MAP REDUCE programming model for counting the number of words in a file.

### **PROCEDURE:**

#### **Step 1: Create Data File:**

Create a file named "input.txt" and populate it with text data that you wish to analyse.



#### **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:**

```
#!/C:/Users/user/AppData/Local/Microsoft/WindowsApps/python.exe
import sys
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:**

```
#!/C:/Users/user/AppData/Local/Microsoft/WindowsApps/python.exe
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_count = count
    prev_word = word
if prev_word == word:
    print('%s\t%s' % (prev_word, prev_count))
```

### 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.

```
start-all.cmd
cd C:/Hadoop/sbin
hdfs dfs -mkdir /WordCount
hdfs dfs -put C:/Users/user/Documents/DataAnalytics/input.txt /WordCount
hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar ^
-input /WordCount/input.txt ^
-output /WordCount/output ^
-mapper "python C:/Users/user/Documents/DataAnalytics/mapper.py" ^
-reducer "python C:/Users/user/Documents/DataAnalytics/reducer.py"
```

### Step 5: Check Output:

Check the output of the Word Count program in the specified HDFS output directory.

```
hdfs dfs -cat /WordCount/output/part-00000
```

**OUTPUT:**

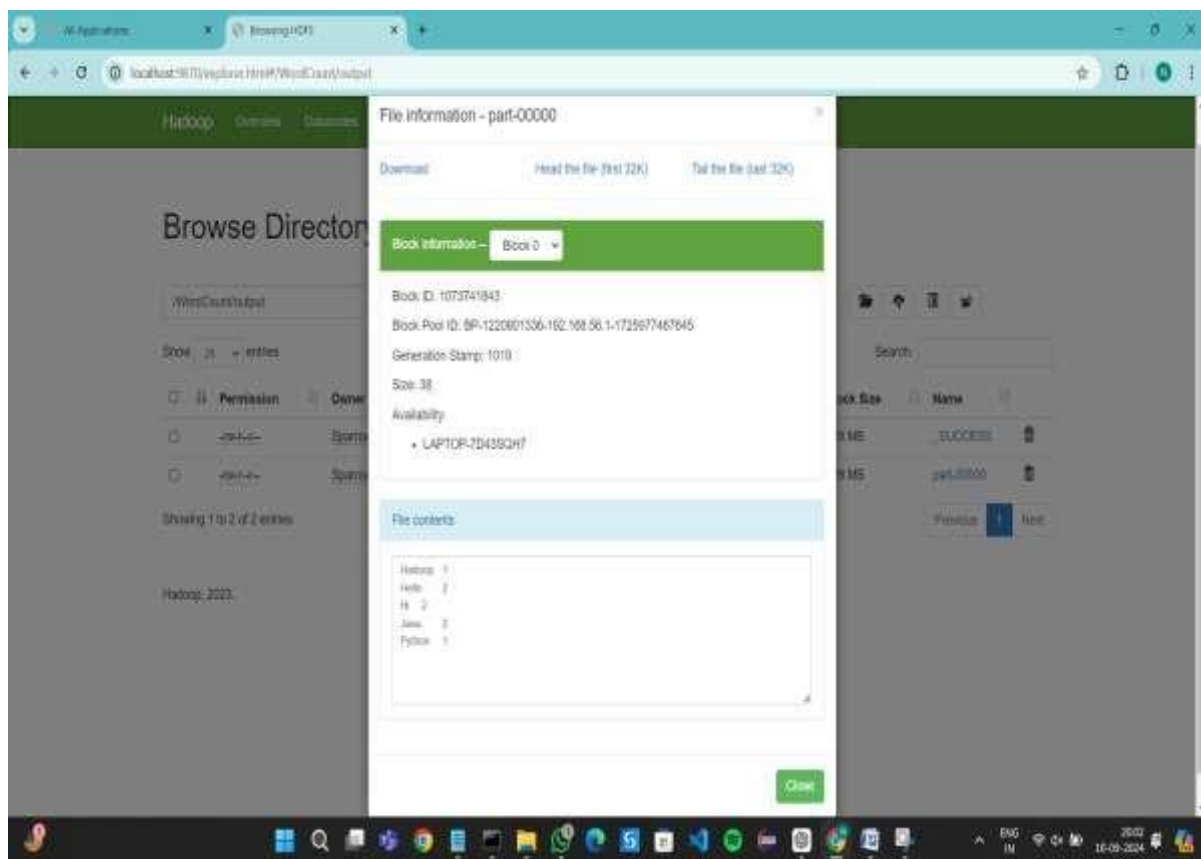
[illegible]

```
Administrator Command Prompt

Job: WordCount
Launched map tasks: 2
Launched reduce tasks: 1
Data local map tasks: 2
Data local map tasks: 2
Total time spent by all maps in occupied slots (ms): 22832
Total time spent by all reducers in occupied slots (ms): 54296
Total time spent by all map tasks (ms): 22832
Total time spent by all reduce tasks (ms): 54296
Total system-wide seconds taken by all map tasks: 22832
Total system-wide seconds taken by all reduce tasks: 54296
Total mapwrite-milliseconds taken by all map tasks: 2440728
Total mapwrite-milliseconds taken by all reduce tasks: 18676720

Map-Reduce framework
Map task records: 2
Map output records: 2
Map output bytes: 1024
Map output materialized bytes: 71
Input split bytes: 1024
Combine input records: 2
Combine output records: 2
Reduce input records: 2
Reduce shuffle bytes: 2
Reduce input records: 2
Reduce output records: 2
Spilled records: 2
Spilled bytes: 2
Failed shuffle: 0
Reported map output: 2
GC time elapsed (ms): 187
CPU time spent (ms): 188
Physical memory (bytes) allocated: 2440728
Virtual memory (bytes) allocated: 14703840
Total committed heap usage (bytes): 2440728
Peak map physical memory (bytes): 4426400
Peak map virtual memory (bytes): 4426400
Peak reduce physical memory (bytes): 25027808
Peak reduce virtual memory (bytes): 4426400

Shuffle errors:
GDP_10-0
COMMITMENT
TD_1000-0
WIPING_1000-0
WIPING_1000-0
WIPING_1000-0
WIPING_1000-0
File Input Format Counters:
Bytes Read: 0
File Output Format Counters:
Bytes Written: 0
2024-09-09 10:11:32,080 INFO chaching.Shredder: Output directory: /home/centos/output
C:\hadoop-3.3.6\bin\hadoop.exe
```



## RESULT:

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