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

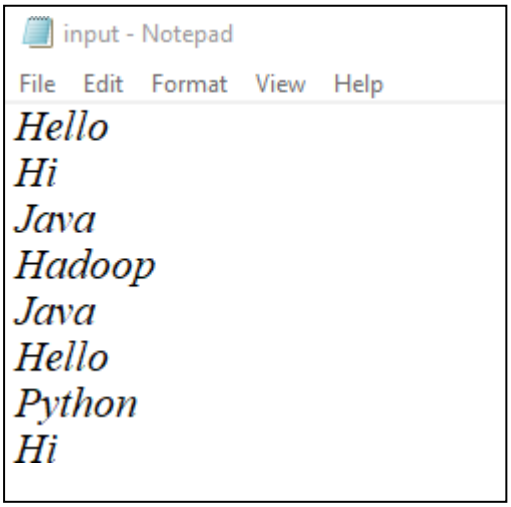
AIM:

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.



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.

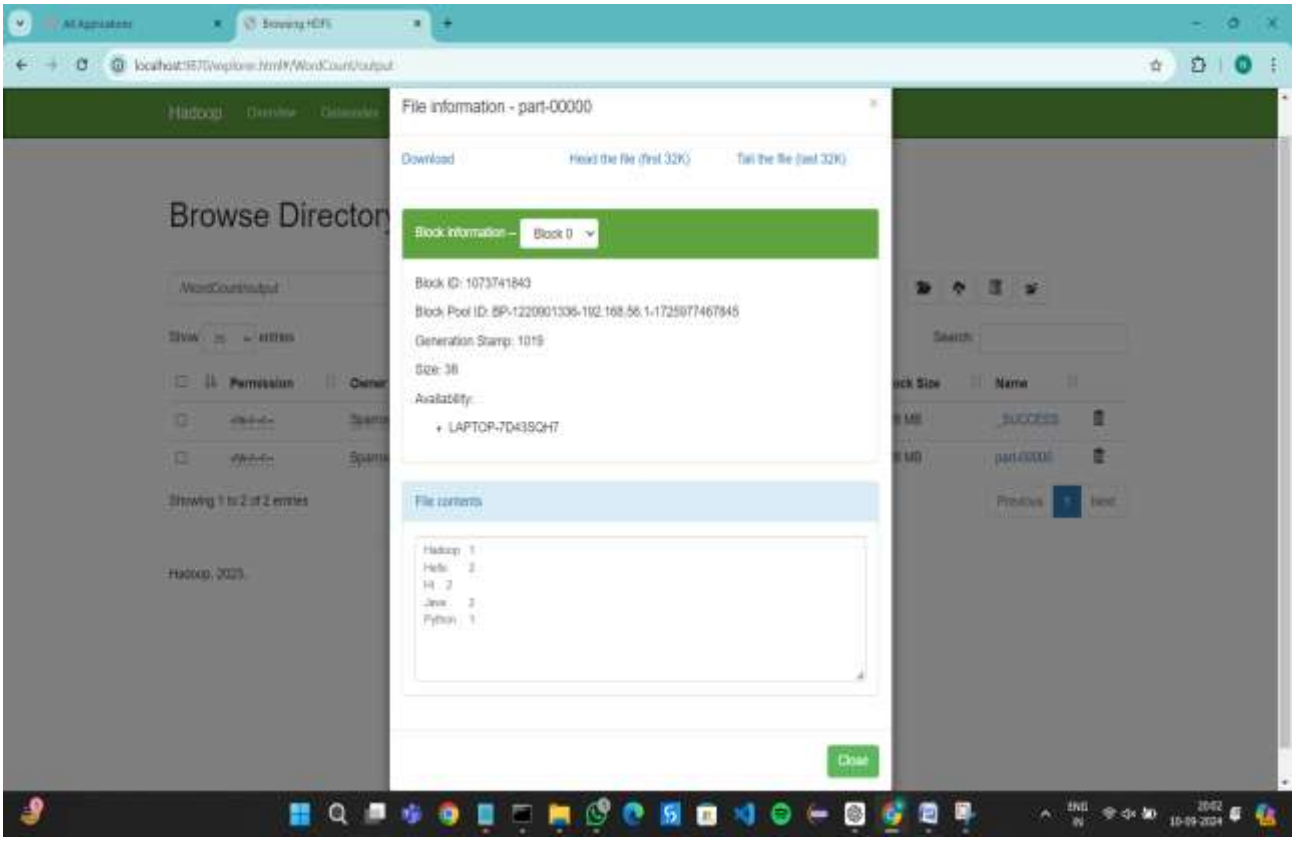
start-all.cmd
cd C:/Hadoop/sbin
hdfs dfs -mkdir /WordCount
hdfs dfs -put C:/Users/user/Documents/DataAnalytics/input.txt /WordCount
hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar ^
-input /WordCount/input.txt ^
-output /WordCount/output ^
-mapper "python C:/ Users/user/Documents/DataAnalytics/mapper.py"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


File information - part-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information –
Block 0
Block ID: 1073741843
Block Pool ID: BP-1220901336-192.168.56.1-1725977467845
Generation Stamp: 1019
Size: 38
Availability:
LAPTOP-7D43SQH7
File contents
Hadoop 1
Hello 2
Hi 2
Java 2
Python 1
Close
Browse Directory

RESULT:

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