

Ex No: 2

Roll No: 210701173

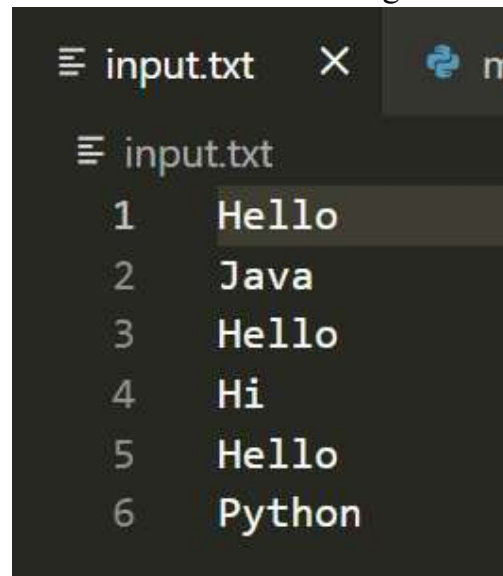
Run a basic Word Count Map Reduce program to understand Map Reduce Paradigm.

AIM:

To run a basic Word Count MapReduce program using Hadoop.

PROCEDURE:

1. Create a text file containing necessary information.

A screenshot of a text editor window titled 'input.txt'. The window shows a list of six lines of text: 'Hello', 'Java', 'Hello', 'Hi', 'Hello', and 'Python'. Each line is numbered from 1 to 6 on the left side of the editor. The text is displayed in a monospaced font on a dark background.

```
1 Hello
2 Java
3 Hello
4 Hi
5 Hello
6 Python
```

2. Create a mapper.py file that will read the input data and split lines into words.

A screenshot of a Python script named 'mapper.py'. The script is shown in a text editor with a dark background. It contains a loop that reads lines from standard input, strips leading and trailing whitespace, splits the line into words, and prints each word on a new line.

```
2 import sys
3 for line in sys.stdin:
4     line = line.strip() # remove leading and trailing whitespace
5     words = line.split() # split the line into words
6     for word in words:
7         print( '%s\t%s' % (word, 1))
8
```

3. Create a mapper.py file that will be used to implement the reducer.py file.

```
2 import sys
3
4 current_word = None
5 current_count = 0
6
7 for line in sys.stdin:
8     line = line.strip()
9     word, count = line.split('\t', 1)
10    count = int(count)
11
12    if current_word == word:
13        current_count += count
14    else:
15        if current_word:
16            print(f'{current_word}\t{current_count}')
17            current_word = word
18            current_count = count
19
20 if current_word == word:
21     print(f'{current_word}\t{current_count}')
```

4. Start the Hadoop environment and create a directory to store values into HDFS.

```
C:\>cd C:\hadoop-3.3.6\sbin
C:\hadoop-3.3.6\sbin>start-dfs.cmd
C:\hadoop-3.3.6\sbin>start-yarn.cmd
starting yarn daemons
C:\hadoop-3.3.6\sbin>jps
15968 NodeManager
33264 Jps
23876 NameNode
20728 ResourceManager
17500 DataNode
```

Commands to create directory, upload files and execute.

```
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"
```

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

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

OUTPUT:

```
Administrator Command Prompt
C:\>cd hadoop-3.3.6
C:\hadoop-3.3.6>cd sbin
C:\hadoop-3.3.6\sbin>hdfs dfs -mkdir /WordCount
C:\hadoop-3.3.6\sbin>hdfs dfs -put C:/Users/vedav/OneDrive/Documents/WordCount/input.txt /WordCount
C:\hadoop-3.3.6\sbin>hadoop jar C:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar ^
More? file input /WordCount/input.txt ^
More? file output /WordCount/output ^
More? mapper "python C:/Users/vedav/OneDrive/Documents/WordCount/mapper.py" ^
More? reducer "python C:/Users/vedav/OneDrive/Documents/WordCount/reducer.py"
Found 10 unexpected arguments on the command line [file, input, /WordCount/input.txt, file, output, /WordCount/output, mapper, python C:/Users/vedav/OneDrive/Documents/WordCount/mapper.py, reducer, python C:/Users/vedav/OneDrive/Documents/WordCount/reducer.py]
Try -help for more information
Streaming Command Failed!
C:\hadoop-3.3.6\sbin>hadoop jar C:/hadoop-3.3.6/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar ^
More? -input /WordCount/input.txt ^
More? -output /WordCount/output ^
More? -mapper "python C:/Users/vedav/OneDrive/Documents/WordCount/mapper.py" ^
More? -reducer "python C:/Users/vedav/OneDrive/Documents/WordCount/reducer.py"
packageJobJar: [/C:/Users/vedav/AppData/Local/Temp/hadoop-unjar2194537253408/17232/] [] C:/Users/vedav/AppData/Local/Temp/streamJob1340729418656478856.jar tmpDir=null
2024-09-03 06:54:30,130 INFO client.DefaultHadoopFileOverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-03 06:54:30,198 INFO client.DefaultHadoopFileOverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-03 06:54:30,200 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/vedav/.staging/job_1725325536302_0001
2024-09-03 06:54:32,149 INFO mapreduce.JobSubmitter: Total input files to process : 1
2024-09-03 06:54:32,190 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-03 06:54:37,139 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725325536302_0001
2024-09-03 06:54:37,340 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-03 06:54:37,562 INFO conf.Configuration: resource-types.xml not found
2024-09-03 06:54:37,563 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-03 06:54:37,924 INFO impl.YarnClientImpl: Submitted application application_1725325536302_0001/
2024-09-03 06:54:37,984 INFO mapreduce.Job: The url to track the job: http://hmdha:8088/proxy/application_1725325536302_0001/
2024-09-03 06:54:37,987 INFO mapreduce.Job: Running job: job_1725325536302_0001
2024-09-03 06:55:14,457 INFO mapreduce.Job: Job job_1725325536302_0001 running in uber mode : false
2024-09-03 06:55:14,458 INFO mapreduce.Job: map 0% reduce 0%
2024-09-03 06:55:30,712 INFO mapreduce.Job: map 50% reduce 0%
2024-09-03 06:55:35,708 INFO mapreduce.Job: map 100% reduce 0%
2024-09-03 06:55:52,004 INFO mapreduce.Job: map 100% reduce 100%
2024-09-03 06:55:57,067 INFO mapreduce.Job: Job job_1725325536302_0001 completed successfully
2024-09-03 06:55:57,104 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=69
    FILE: Number of bytes written=830745
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=240
    HDFS: Number of bytes written=35
    HDFS: Number of read operations=11
```

```
Administrator: Command Prompt

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)=27822
  Total time spent by all reduces in occupied slots (ms)=14530
  Total time spent by all map tasks (ms)=27822
  Total time spent by all reduce tasks (ms)=14530
  Total vcore-milliseconds taken by all map tasks=27822
  Total vcore-milliseconds taken by all reduce tasks=14530
  Total megabyte-milliseconds taken by all map tasks=20489728
  Total megabyte-milliseconds taken by all reduce tasks=14878720

Map-Reduce Framework
  Map input records=3
  Map output records=7
  Map output bytes=49
  Map output materialized bytes=75
  Input split bytes=188
  Combine input records=0
  Combine output records=0
  Reduce input groups=5
  Reduce shuffle bytes=75
  Reduce input records=7
  Reduce output records=5
  Spilled Records=14
  Shuffled Maps=2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=187
  CPU time spent (ms)=1201
  Physical memory (bytes) snapshot=935358464
  Virtual memory (bytes) snapshot=1470304256
  Total committed heap usage (bytes)=793772032
  Peak Map Physical memory (bytes)=142470056
  Peak Map Virtual memory (bytes)=488009184
  Peak Reduce Physical memory (bytes)=255193088
  Peak Reduce Virtual memory (bytes)=471677824

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=54

File Output Format Counters
  Bytes Written=35

2024-09-03 06:55:57,164 INFO streaming.StreamJob: Output directory: /wordcount/output
C:\hadoop-3.3.6\sbin>
```

File information - part-00000

[Download](#)[Head the file \(first 32K\)](#)[Tail the file \(last 32K\)](#)**Block information**

Block 0 ▾

Block ID: 1073741848

Block Pool ID: BP-2024779555-192.168.56.1-1724921847714

Generation Stamp: 1024

Size: 29

Availability:

- envy24

File contents

Hello	3
Hi	1
Java	1
Python	1

RESULT:

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

