

IMPLEMENT WORD COUNT/FREQUENCY PROGRAMS USING MAPREDUCE

AIM:

To implement the python mapper and reducer programs using MapReduce to count the words in a text file using Hadoop.

PROCEDURE:

1. Open command prompt as administrator and start the Hadoop by using the command:

```
start-all.cmd
```

2. Create a new directory in the Hadoop file systems using the command:

```
hadoop fs -mkdir /wordCount
```

3. Upload the input text file into the wordCount directory using the command:

```
hadoop fs -put C:/DA/wordcount/input.txt /wordCount
```

4. Create the mapper and reducer files.

5. To execute the files with Hadoop streaming run the following command:

```
hadoop jar "C:\hadoop-3.3.6\share\hadoop\tools\lib\hadoop-streaming-3.3.6.jar" ^-input  
/wordCount/input.txt ^-output /wordCount/output/ ^-mapper "python  
C:/wordcount/mapper.py" ^-reducer "python C:/wordcount/reducer.py"
```

MAPPER.PY

```
#!C:/Users/md_aa/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))
```

REDUCER.PY

```
#!C:/Users/md_aa/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))
```

OUTPUT:

The screenshot shows the Hadoop web interface at localhost:9870/explorer.html. The 'Browse Directory' page displays a table of files in the root directory. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, Name, and a trash icon. Three files are listed: 'tmp', 'weather', and 'wordCount'. The 'wordCount' file is highlighted in blue. The page also shows a search bar, a 'Go!' button, and a 'Showing 1 to 3 of 3 entries' message.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	md_aa	supergroup	0 B	Sep 17 23:03	0	0 B	tmp
drwxr-xr-x	md_aa	supergroup	0 B	Sep 17 23:29	0	0 B	weather
drwxr-xr-x	md_aa	supergroup	0 B	Sep 17 23:15	0	0 B	wordCount

[Hadoop](#) [Overview](#) [Datanodes](#) [Datanode Volume Failures](#) [Snapshot](#) [Startup Progress](#) [Utilities](#)

Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	md_aa	supergroup	28 B	Sep 17 23:14	1	128 MB	input.txt	<input type="checkbox"/>
<input type="checkbox"/>	drwxr-xr-x	md_aa	supergroup	0 B	Sep 17 23:16	0	0 B	output	<input type="checkbox"/>

Showing 1 to 2 of 2 entries

localhost:9870/explorer.html#/wordCount/output

[Hadoop](#) [Overview](#) [Datanodes](#)

Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	md_aa	supergroup	28 B	Sep 17 23:14	1	128 MB	input.txt	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	md_aa	supergroup	0 B	Sep 17 23:16	0	0 B	output	<input type="checkbox"/>

Showing 1 to 2 of 2 entries

File information - part-00000

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

Block information -- Block 0

Block ID: 1073741965
Block Pool ID: BP-551977252-172.17.48.1-1724733081844
Generation Stamp: 1141
Size: 24
Availability:

- mohamed-aadhil.mshome.net

File contents

hello 2
hi 2
Aadhil 1

Block Size

Name

MB

part-00000

Previous

1

Next

RESULT:

Thus, the implementation of the python mapper and reducer programs using MapReduce to count the words in a text file using Hadoop is executed successfully.