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:

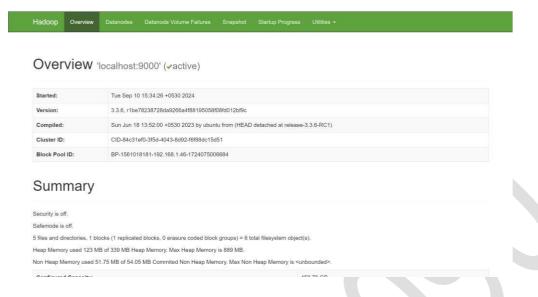
Open command prompt and run as administrator

Start Hadoop services by typing in the following commands:

- start-dfs.cmd
- start-yarn.cmd

```
C:\Windows\System32>jps
14212 Jps
C:\Windows\System32>start-dfs.cmd
C:\Windows\System32>jps
12000 DataNode
16488 Jps
24904 NameNode
C:\Windows\System32>start-yarn.cmd
starting yarn daemons
C:\Windows\System32>jps
12000 DataNode
6384 NodeManager
31300 Jps
24904 NameNode
29036 ResourceManager
C:\Windows\System32>
```

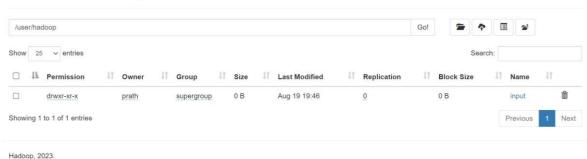
Open the browser and go to the URL localhost:9870



Create a directory in HDFS using the command:

hdfs dfs -mkdir -p /user/hadoop/input





Copy the input file to HDFS using the command:

hdfs dfs -put C:/Semester7/DataAnalytics/Lab/input.txt /user/hadoop/input

```
C:\Windows\System32>hdfs dfs -put C:/Semester7/DataAnalytics/Lab/input.txt /user/hadoop/input
```

Display the contents of the file using this command:

hdfs dfs -cat /user/hadoop/input/input.txt

```
C:\Windows\System32>hdfs dfs -cat /user/hadoop/input/input.txt
Hello world
Welcome to the world of programming
Have fun
Bye
```

Create mapper.py and reducer.py files 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))
```

reducer.py

```
import sys
previous word=None
previous count=0
for line in sys.stdin:
   line=line.strip()
   word, count=line.split("\t")
   count=int(count)
    if previous_word==word:
       previous count+=count
   else:
        if prev word:
           print("%s\t%s" %(previous word, previous count))
        previous word=word
        previous count=count
if previous word==word:
   print("%s\t%s" %(previous word, previous count))
```

Run the Hadoop Streaming Job and give the file paths to the input, mapper and reducer using the following command:

 $hadoop \quad jar \quad \% HADOOP_HOME\% \\ \share \\ \hadoop \\ \tools \\ \lib \\ \hadoop-streaming-*.jar^{\wedge}$

- -mapper "python C:\Semester7\DataAnalytics\Lab\Ex.2\mapper.py" -reducer "python C:\Semester7\DataAnalytics\Lab\Ex.2\reducer.py"^
- -input /user/hadoop/input/input.txt -output /user/hadoop/output

```
C:\Windows\System32:hadoop jar %HADOOP_EOMEX\share\hadoop\tools\lib\hadoop\streaming=".jar\
More\" = input (\).\Seeseter/\DataAnalytics\lab\Ex.2\mapper.pyr\" reducer "pyrhon C:\Seeseter/\DataAnalytics\lab\Ex.2\reducer.pyr\" ^
More\" = input (\).\Seeseter/\DataAnalytics\lab\Ex.2\mapper.pyr\" reducer "pyrhon C:\Seeseter/\DataAnalytics\lab\Ex.2\reducer.pyr\" ^
Move\" = input (\).\share\hadoop\tauta (\).\share\hado
```

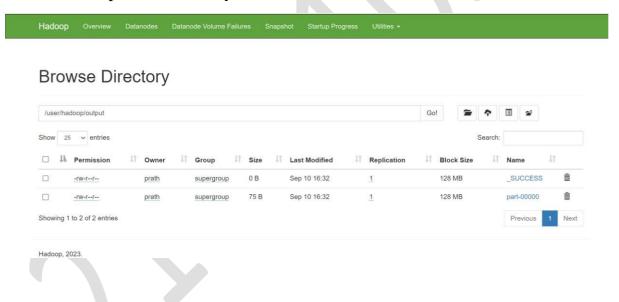
```
Map-Reduce Framework
                Map input records=4
                Map output records=11
                Map output bytes=83
                Map output materialized bytes=117
                Input split bytes=202
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                 Reduce shuffle bytes=117
                Reduce input records=11
                 Reduce output records=10
                Spilled Records=22
                Shuffled Maps =2
                 Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=146
                CPU time spent (ms)=421
Physical memory (bytes) snapshot=976105472
                Virtual memory (bytes) snapshot=1553080320
                 Total committed heap usage (bytes)=861405184
                Peak Map Physical memory (bytes)=351379456
Peak Map Virtual memory (bytes)=535887872
                 Peak Reduce Physical memory (bytes)=273371136
                 Peak Reduce Virtual memory (bytes)=489156608
        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=95
        File Output Format Counters
                Bytes Written=75
2024-09-10 16:32:21,532 INFO streaming.StreamJob: Output directory: /user/hadoop/output
```

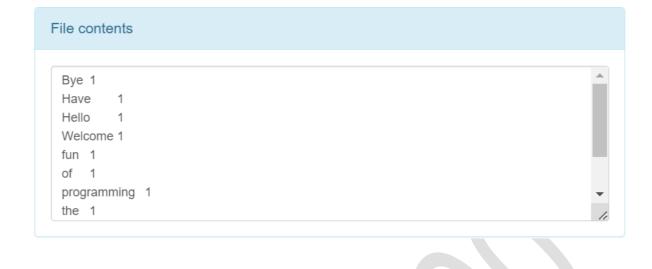
View the output using the command:

hdfs dfs -cat /user/hadoop/output/part-00000

```
C:\Windows\System32>hdfs dfs -cat /user/hadoop/output/part-00000
Bye
        1
Have
Hello
        1
Welcome 1
fun
        1
of
        1
programming
                 1
the
to
        1
world
        2
```

Check the output on the file system in the browser





RESULT:

Thus, to demonstrate the MAP REDUCE programming model for counting the number of words in a file was completed successfully.