WordCount using MapReduce in Hadoop

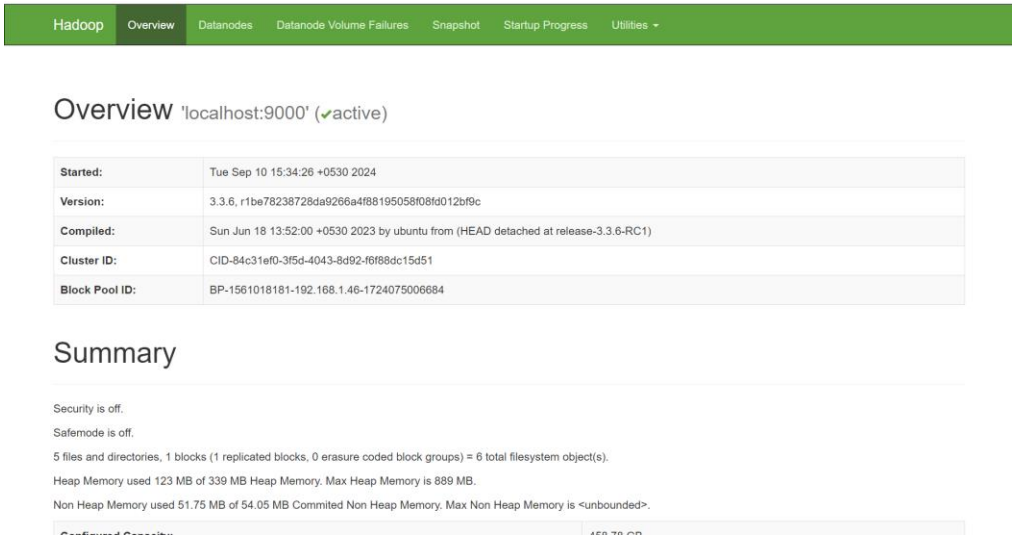
Open command prompt and run as administrator

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>
```

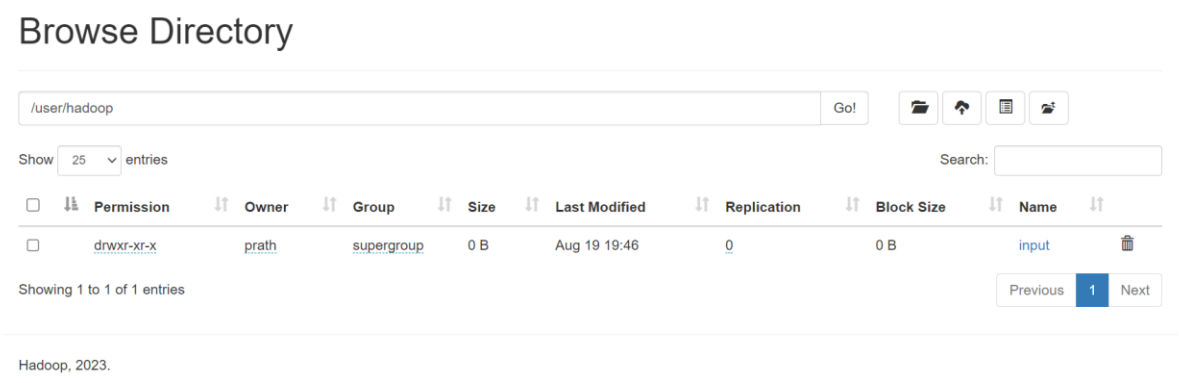
Open the browser and go to the URL localhost:9870



Create a directory in HDFS using the command:

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

```
C:\Windows\System32>hdfs dfs -mkdir -p /user/hadoop/input

C:\Windows\System32>
```



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 jar %HADOOP_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar^

-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_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar^
More?  -mapper "python C:\Semester7\DataAnalytics\Lab\Ex.2\mapper.py" -reducer "python C:\Semester7\DataAnalytics\Lab\Ex.2\reducer.py" ^
More?  -input /user/hadoop/input/input.txt -output /user/hadoop/output
packageJobJar: [/C:/Users/prath/AppData/Local/Temp/hadoop-unjar7189392901373763060/] [] C:\Users\prath\AppData\Local\Temp\streamjob8037495271429660861.jar tmpDir=null
2024-09-10 16:32:00,707 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-10 16:32:00,855 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-09-10 16:32:01,325 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/prath/.staging/job_1725962680547_0004
2024-09-10 16:32:01,570 INFO mapred.FileInputFormat: Total input files to process : 1
2024-09-10 16:32:01,651 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-10 16:32:01,794 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1725962680547_0004
2024-09-10 16:32:01,794 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-10 16:32:01,934 INFO conf.Configuration: resource-types.xml not found
2024-09-10 16:32:01,934 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-09-10 16:32:01,996 INFO impl.YarnClientImpl: Submitted application application_1725962680547_0004
2024-09-10 16:32:02,032 INFO mapreduce.Job: The url to track the job: http://ANURAJ:8088/proxy/application_1725962680547_0004/
2024-09-10 16:32:02,034 INFO mapreduce.Job: Running job: job_1725962680547_0004
2024-09-10 16:32:09,189 INFO mapreduce.Job: Job job_1725962680547_0004 running in uber mode : false
2024-09-10 16:32:09,191 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-10 16:32:15,324 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-10 16:32:20,394 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-10 16:32:21,417 INFO mapreduce.Job: Job job_1725962680547_0004 completed successfully
2024-09-10 16:32:21,532 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=111
                FILE: Number of bytes written=839411
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=297
                HDFS: Number of bytes written=75
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        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)=6817
                Total time spent by all reduces in occupied slots (ms)=3259
                Total time spent by all map tasks (ms)=6817
                Total time spent by all reduce tasks (ms)=3259
                Total vcore-milliseconds taken by all map tasks=6817
                Total vcore-milliseconds taken by all reduce tasks=3259
                Total megabyte-milliseconds taken by all map tasks=6980608
```

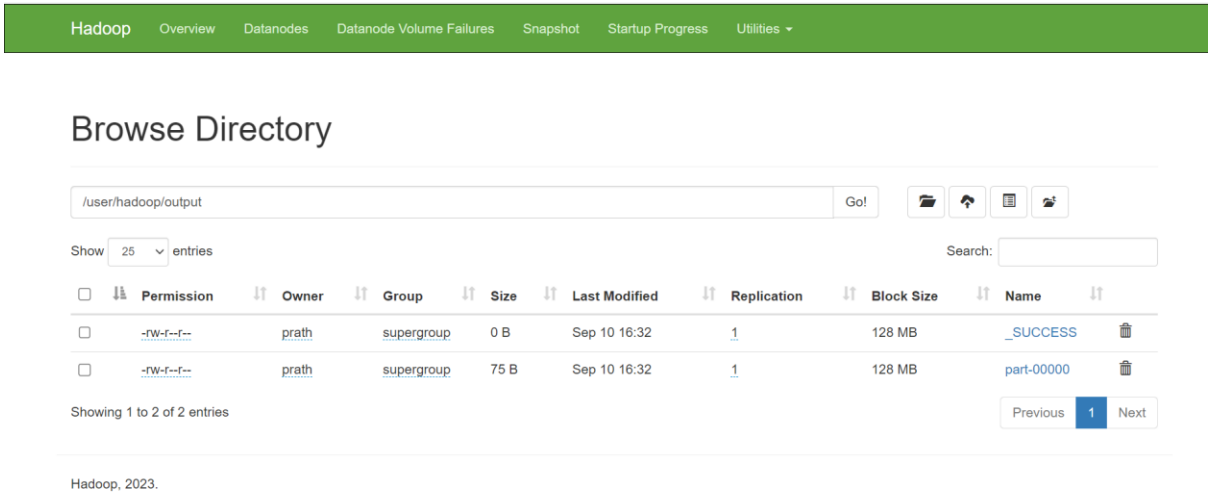
```
        Map-Reduce Framework
                Map input records=4
                Map output records=11
                Map output bytes=83
                Map output materialized bytes=117
                Input split bytes=202
                Combine input records=0
                Combine output 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    1
Hello   1
Welcome 1
fun     1
of      1
programming     1
the     1
to      1
world   2
```

Check the output on the file system in the browser



File contents

```
Bye  1
Have      1
Hello     1
Welcome 1
fun   1
of    1
programming   1
the   1
```