Hadoop Streaming – Wordcount Using Map reducer in Hadoop

Steps:

1. Open command prompt and run as administrator

Go to hadoop sbin directory
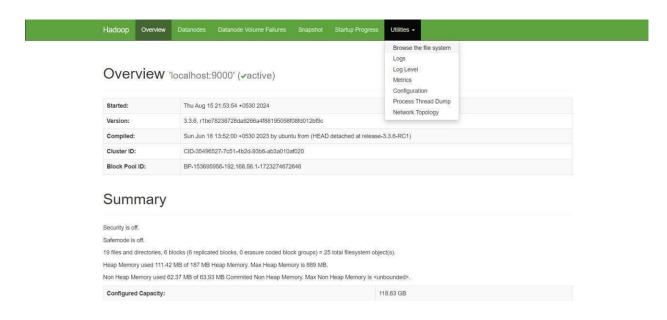
```
C:\>cd C:\Hadoop\sbin

C:\Hadoop\sbin>
```

Note:

1. Check hadoop/data/datanode and hadoop/data/namenode and if both folders are empty, type "hdfs namenode -format".

2. Check python version with "python --version".

3. Check "C:\Python39\" is added in Environment variables > System variables > Path, if not add your python path.

4. Check Environment variables > System variables > HADOOP_HOME is set as "C:\Hadoop".

```
C:\Hadoop\sbin>echo %HADOOP_HOME%
C:\Hadoop

C:\Hadoop\sbin>python --version
Python 3.11.4
```

2. Start Hadoop Services start-dfs.cmd start-yarn.cmd

```
C:\Hadoop\sbin>start-dfs.cmd

C:\Hadoop\sbin>start-yarn.cmd
starting yarn daemons

C:\Hadoop\sbin>jps
13120 NameNode
2384 NodeManager
4100 DataNode
7956 ResourceManager
9124 Jps
```

3. Open the browser and go to the URL localhost:9870

| Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities ▾ |

Browse the file system
Logs
Log Level
Metrics
Configuration
Process Thread Dump
Network Topology

Overview 'localhost:9000' (✔active)

Started:	Thu Aug 15 21:53:54 +0530 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 13:52:00 +0530 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-35496527-7c51-4b2d-93b6-ab3a010af020
Block Pool ID:	BP-153695956-192.168.56.1-1723274672646

Summary

Security is off.

Safemode is off.

19 files and directories, 6 blocks (6 replicated blocks, 0 erasure coded block groups) = 25 total filesystem object(s).

Heap Memory used 111.42 MB of 187 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 62.37 MB of 63.93 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

| Configured Capacity: | 118.63 GB |

4. Create a Directory in HDFS hdfs dfs -mkdir -p /user/hadoop/input

```
C:\Hadoop\sbin>hdfs dfs -mkdir -p /user/hadoop/input

C:\Hadoop\sbin>_
```

5. Copy the Input File to HDFS hdfs dfs -put C:/Users/Admin/input.txt /user/hadoop/input

```
C:\Hadoop\sbin>hdfs dfs -put C:/Users/Admin/input.txt /user/hadoop/input

C:\Hadoop\sbin>hdfs dfs -ls /user/hadoop/input
Found 1 items
-rw-r--r--   1 Admin supergroup          42 2024-08-18 15:15 /user/hadoop/input/input.txt

C:\Hadoop\sbin>hdfs dfs -cat /user/hadoop/input/input.txt
hello world
hi all
hello all
all the best

C:\Hadoop\sbin>_
```

Note: mapper.py:

```python
#! /usr/bin/env python
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:

```python
#! /usr/bin/env python
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_word=word
        prev_count=count
if prev_word==word:
    print('%s\t%s' % (prev_word, prev_count))
```

6. Run the Hadoop Streaming Job hadoop jar

hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.3.1.jar ^
   -files
/Users/monid/OneDrive/Documents/DataAnalytics/mapper.py,/Users/monid/OneDrive/Document
s/DataAnalytics/reducer.py ^
   -input /user/hadoop/input/data.txt ^

   -output /user/output ^
   -mapper "python C:/Users/monid/OneDrive/Documents/DataAnalytics/mapper.py" ^
   -reducer "python C:/Users/monid/OneDrive/Documents/DataAnalytics/reducer.py "

```
C:\Hadoop\sbin>hadoop jar %HADOOP_HOME%\share\hadoop\tools\lib\hadoop-streaming-*.jar ^
More? -mapper "python C:\\Users\\Admin\\mapper.py" -reducer "python C:\\Users\\Admin\\reducer.py" ^
More? -input /user/hadoop/input/input.txt -output /user/hadoop/output
packageJobJar: [/C:/Users/Admin/AppData/Local/Temp/hadoop-unjar4352040893517806187/] [] C:\Users\Admin\AppData\Local\Tem
p\streamjob1481680013776791488.jar tmpDir=null
2024-08-18 15:21:35,517 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-18 15:21:35,949 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-18 15:21:37,279 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
Admin/.staging/job_1723973693127_0001
2024-08-18 15:21:38,430 INFO mapred.FileInputFormat: Total input files to process : 1
2024-08-18 15:21:38,990 INFO mapreduce.JobSubmitter: number of splits:2
2024-08-18 15:21:39,415 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1723973693127_0001
2024-08-18 15:21:39,416 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-18 15:21:39,723 INFO conf.Configuration: resource-types.xml not found
2024-08-18 15:21:39,724 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-08-18 15:21:41,112 INFO impl.YarnClientImpl: Submitted application application_1723973693127_0001
2024-08-18 15:21:41,196 INFO mapreduce.Job: The url to track the job: http://DESKTOP-TF65P79:8088/proxy/application_1723
973693127_0001/
2024-08-18 15:21:41,202 INFO mapreduce.Job: Running job: job_1723973693127_0001
2024-08-18 15:22:04,875 INFO mapreduce.Job: Job job_1723973693127_0001 running in uber mode : false
2024-08-18 15:22:04,905 INFO mapreduce.Job:  map 0% reduce 0%
2024-08-18 15:22:20,569 INFO mapreduce.Job:  map 100% reduce 0%
2024-08-18 15:22:32,773 INFO mapreduce.Job:  map 100% reduce 100%
```

```
        File Input Format Counters
              Bytes Read=63
        File Output Format Counters
              Bytes Written=40
2024-08-18 15:22:34,120 INFO streaming.StreamJob: Output directory: /user/hadoop/output

C:\Hadoop\sbin>_
```
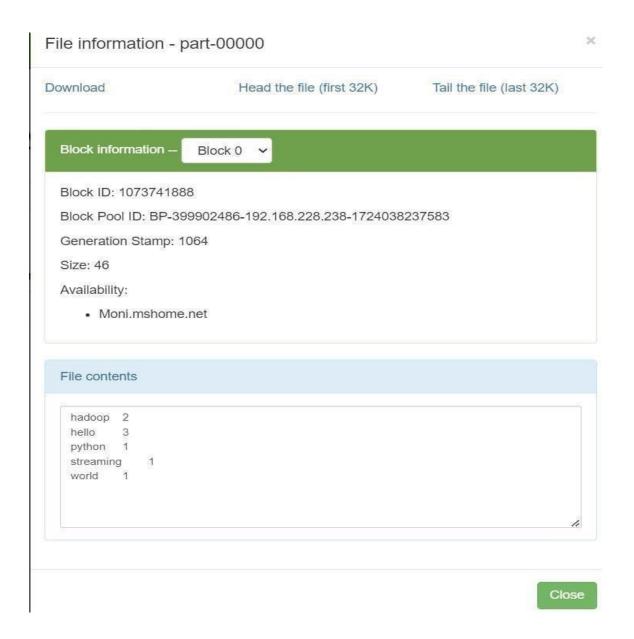
7. View the Output

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

```
:\hadoop\sbin>hadoop dfs -cat /user/output/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
hadoop  2
hello   3
python  1
streaming       1
world   1

:\hadoop\sbin>
```

8. Once the map reduce operations are performed successfully, the output will be present in the
   specified directory.

"/user/output/part-00000"

File information - part-00000   ×

Download   Head the file (first 32K)   Tail the file (last 32K)

Block information — Block 0 ∨

Block ID: 1073741888

Block Pool ID: BP-399902486-192.168.228.238-1724038237583

Generation Stamp: 1064

Size: 46

Availability:

- Moni.mshome.net

File contents

```
hadoop    2
hello     3
python    1
streaming     1
world     1
```

Close

9. Stop Hadoop Services stop-dfs.cmd stop-yarn.cmd

```
C:\Hadoop\sbin>stop-dfs.cmd
SUCCESS: Sent termination signal to the process with PID 6248.
SUCCESS: Sent termination signal to the process with PID 8616.

C:\Hadoop\sbin>stop-yarn.cmd
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 16904.
SUCCESS: Sent termination signal to the process with PID 15344.

INFO: No tasks running with the specified criteria.

C:\Hadoop\sbin>
```

10.   Stop Hadoop Services stop-dfs.cmd stop-yarn.cmd

```
C:\Hadoop\sbin>stop-dfs.cmd
SUCCESS: Sent termination signal to the process with PID 6248.
SUCCESS: Sent termination signal to the process with PID 8616.

C:\Hadoop\sbin>stop-yarn.cmd
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 16904.
SUCCESS: Sent termination signal to the process with PID 15344.

INFO: No tasks running with the specified criteria.

C:\Hadoop\sbin>
```

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.