

# Finding the words and their frequencies in given data using HADOOP MAPREDUCE

## 1. Starting all Hadoop daemons

- `cd hadoop-3.2.1/`
- `cd sbin`
- `./start-all.sh`
- `Jps`

```
hadoop@ubuntu:~$ cd hadoop-3.2.1/
hadoop@ubuntu:~/hadoop-3.2.1$ cd sbin
hadoop@ubuntu:~/hadoop-3.2.1/sbin$ ls
distributed-exclude.sh  hadoop-daemons.sh  mr-jobhistory-daemon.sh  start-all.sh  start-dfs.sh  start-yarn.sh  stop-balancer.sh
FederationStateStore  httpfs.sh  refresh-namenodes.sh  start-balancer.sh  start-secure-dns.sh  stop-all.cmd  stop-dfs.cmd
hadoop-daemon.sh  kms.sh  start-all.cmd  start-dfs.cmd  start-yarn.cmd  stop-all.sh  stop-dfs.sh
hadoop@ubuntu:~/hadoop-3.2.1/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
Starting resource manager
Starting node managers
hadoop@ubuntu:~/hadoop-3.2.1/sbin$ jps
8384 SecondaryNameNode
8632 ResourceManager
9116 Jps
8766 NodeManager
8207 DataNode
8079 NameNode
hadoop@ubuntu:~/hadoop-3.2.1/sbin$
```

## 2. Verifying the namenode's status and application status

- Go to browser and type <http://localhost:8088/> and <http://localhost:9870/> and check for active status.

## Overview 'localhost:9000' (active)

Started:	Sat Apr 24 01:05:52 -0700 2021
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 08:56:00 -0700 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-dd52d547-3da0-4b59-a5d6-7e21f7ee68af
Block Pool ID:	BP-1821965271-127.0.1.1-1618670408249

## Summary

Security is off.  
Safemode is off.  
18 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 23 total filesystem object(s).  
Heap Memory used 59.7 MB of 166 MB Heap Memory. Max Heap Memory is 652.5 MB.  
Non Heap Memory used 46.82 MB of 47.96 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	67.91 GB
Configured Remote Capacity:	0 B
DFS Used:	331.87 KB (0%)
Non DFS Used:	10.56 GB
DFS Remaining:	53.85 GB (79.31%)


AirjunAS861/BDA\_1BM1

All Applications

Namenode information

localhost:8088/cluster

Logged in as: dr.who



Cluster

About Nodes Node Labels ApplicationsNEWNEW\_SAVINGSUBMITTEDACCEPTEDRUNNINGFINISHEDFAILEDKILLED Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID

User

Name

Application Type

Queue

Application Priority

StartTime

LaunchTime

FinishTime

State

FinalStatus

Running Containers

Allocated CPU VCoers

Allocated Memory MB

Reserved CPU VCoers

Reserved Memory MB

% of Queue

% of Cluster

Progress

Tracking UI

Blacklisted Nodes

No data available in table

Showing 0 to 0 of 0 entries

First Previous Next Last

### 3. Creating a sample file for input in local system and moving it to dfs

- Create a file in local system
  - cd Downloads
  - nano word\_sample.txt
  - cat word\_sample.txt
- Go to Hadoop-3.2.1 directory and execute following commands
  - Hdfs dfs -ls /
- Create a folder in Hadoop for placing input files.
  - hdfs dfs -mkdir /input
  - hdfs dfs -copyFromLocal ~/Downloads/word\_sample.txt /input
  - hdfs dfs -ls /input
  - hdfs dfs -cat /input/word\_sample.txt
  -

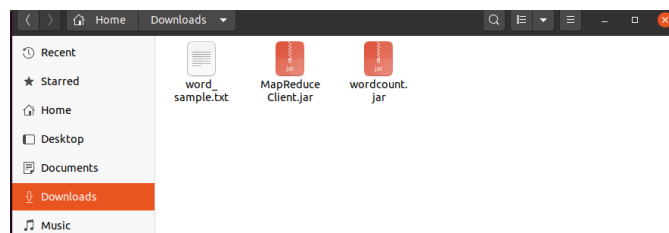
```
hadoop@ubuntu:~/hadoop-3.2.1$ cd ..
hadoop@ubuntu:~$ cd Downloads
hadoop@ubuntu:~/Downloads$ nano word_sample.txt
hadoop@ubuntu:~/Downloads$ cat word_sample.txt
Hello my name is Arjun
I love computer Science

hadoop@ubuntu:~/Downloads$ cd ..
hadoop@ubuntu:~$ cd hadoop-3.2.1/
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -ls /
Found 2 items
drwx-----   - hadoop supergroup          0 2021-04-24 00:31 /tmp
drwxr-xr-x   - hadoop supergroup          0 2021-04-20 20:18 /user
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -mkdir /input
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -copyFromLocal ~/Downloads/word_sample.txt /input
2021-04-24 01:22:54,063 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -ls /input
Found 1 items
-rw-r--r--   1 hadoop supergroup          48 2021-04-24 01:22 /input/word_sample.txt
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -cat /input/word_sample.txt
2021-04-24 01:23:48,253 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
Hello my name is Arjun
I love computer Science

hadoop@ubuntu:~/hadoop-3.2.1$
```

### 4. Using MapreduceClient.jar file to calculate wordcount

- The java code to calculate word count is in wordcount.jar and also use MapReduceClient.jar



- Execute `hadoop jar /home/hadoop/Downloads/MapReduceClient.jar wordcount /input /output`

```
hadoop@ubuntu:~/hadoop-3.2.1$ hadoop jar /home/hadoop/Downloads/MapReduceClient.jar wordcount /input /output
2021-04-24 01:46:29,722 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-04-24 01:46:31,180 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1619251589497_0001
2021-04-24 01:46:31,400 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2021-04-24 01:46:31,761 INFO input.FileInputFormat: Total input files to process : 1
2021-04-24 01:46:31,888 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2021-04-24 01:46:31,955 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2021-04-24 01:46:31,991 INFO mapreduce.JobSubmitter: number of splits:1
2021-04-24 01:46:32,477 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
2021-04-24 01:46:32,579 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1619251589497_0001
2021-04-24 01:46:32,582 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-04-24 01:46:33,240 INFO conf.Configuration: resource-types.xml not found
2021-04-24 01:46:33,241 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-04-24 01:46:34,069 INFO Impl.YarnClientImpl: Submitted application application_1619251589497_0001
2021-04-24 01:46:34,288 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1619251589497_0001/
2021-04-24 01:46:34,389 INFO mapreduce.Job: Running job: job_1619251589497_0001
2021-04-24 01:47:00,352 INFO mapreduce.Job: Job job_1619251589497_0001 running in uber mode : false
2021-04-24 01:47:00,360 INFO mapreduce.Job: map 0% reduce 0%
2021-04-24 01:47:11,759 INFO mapreduce.Job: map 100% reduce 0%
2021-04-24 01:47:22,960 INFO mapreduce.Job: map 100% reduce 100%
2021-04-24 01:47:24,012 INFO mapreduce.Job: Job job_1619251589497_0001 completed successfully
2021-04-24 01:47:24,260 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=107
    FILE: Number of bytes written=451489
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=156
    HDFS: Number of bytes written=65
    HDFS: Number of read operations=8
    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=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=9376
    Total time spent by all reduces in occupied slots (ms)=7920
    Total time spent by all map tasks (ms)=9376
    Total time spent by all reduce tasks (ms)=7920
    Total vcore-milliseonds taken by all map tasks=9376
    Total vcore-milliseonds taken by all reduce tasks=7920
    Total megabyte-milliseonds taken by all map tasks=9601024
    Total megabyte-milliseonds taken by all reduce tasks=8110080
```

- A new application will be running with final status as undefined and after completion of the task finalStatus is set to Succeeded in <http://localhost:8088/>

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue	% of Clus
application_1619251589497_0001	hadoop	word count	MAPREDUCE	default	0	Sat Apr 24 01:46:33 -0700 2021	Sat Apr 24 01:46:37 -0700 2021	N/A	ACCEPTED	UNDEFINED	1	1	2048	0	0	25.0	25.0

Showing 1 to 1 of 1 entries

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB	Reserved CPU VCoers	Reserved Memory MB	% of Queue	% of Clus
application_1619251589497_0001	hadoop	word count	MAPREDUCE	default	0	Sat Apr 24 01:46:33 -0700 2021	Sat Apr 24 01:46:37 -0700 2021	Sat Apr 24 01:47:22 -0700 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0

Showing 1 to 1 of 1 entries

## 5. Viewing the output

- `hdfs dfs -cat /output/*`

```
hadoop@ubuntu:~/hadoop-3.2.1$ hdfs dfs -cat /output/*
2021-04-24 01:53:37,986 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
Arjun 1
Hello 1
I 1
Science 1
computer 1
is 1
love 1
my 1
name 1
hadoop@ubuntu:~/hadoop-3.2.1$
```

## 6. Stop all daemons

```
hadoop@ubuntu:~/hadoop-3.2.1/sbin$ ./stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [ubuntu]
Stopping nodemanagers
localhost: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying
to kill with kill -9
Stopping resourcemanager
hadoop@ubuntu:~/hadoop-3.2.1/sbin$
```