HDFS Exersices

1. Verify HDFS access

- Help
  - $ hadoop fs -help

- As hdfs user
  - $ sudo -u hdfs hadoop fs -ls -R /
- As local user
  - $ hadoop fs -ls -R -d
  - $ hadoop fs -mkdir wordcount
  - $ hadoop fs -ls -R

```
[cloudera@elephant ~]$ sudo -u hdfs hadoop fs -ls -R /
drwxr-xr-x   - hdfs supergroup          0 2014-10-07 00:40 /user
drwxr-xr-x   - cloudera supergroup          0 2014-10-07 00:40 /user/cloudera
[cloudera@elephant ~]$
[cloudera@elephant ~]$
[cloudera@elephant ~]$
[cloudera@elephant ~]$ hadoop fs -ls -R -d
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2014-10-07 00:40 .
[cloudera@elephant ~]$
[cloudera@elephant ~]$ hadoop fs -mkdir wordcount
[cloudera@elephant ~]$ hadoop fs -ls -R
drwxr-xr-x   - cloudera supergroup          0 2014-10-07 00:42 wordcount
[cloudera@elephant ~]$
```

2. Create input path
   $ hadoop fs -mkdir /user/cloudera/wordcount/input

3. Download a file example:

   $ wget http://www.gnu.org/licenses/gpl.txt
   $ wget http://www.apache.org/licenses/LICENSE-2.0.txt

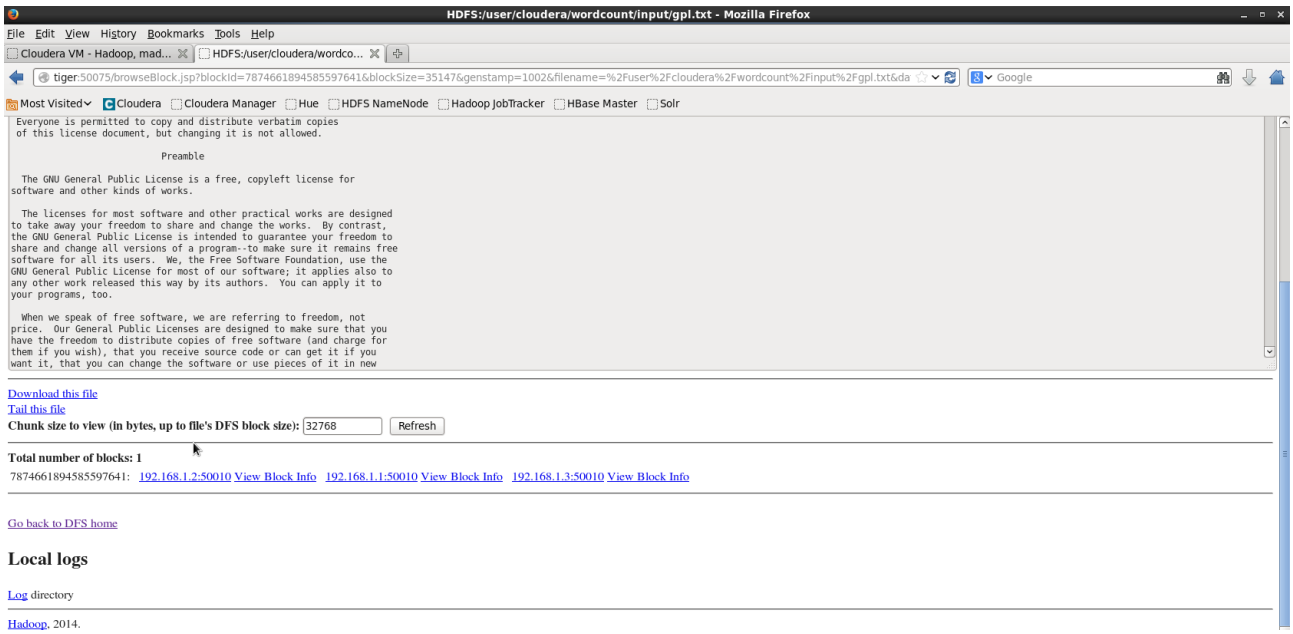
4. Put on Hadoop HDFS
   $ hadoop fs -put gpl.txt LICENSE-2.0.txt wordcount/input

5. Test output files on HDFS

   $ hadoop fs -tail wordcount/input/gpl.txt
   $ hadoop fs -tail wordcount/input/LICENSE-2.0.txt

6. Verify on web service http://elephant:50070/dfshealth.jsp

7. Click on ***Browse the filesystem -> /user/cloudera/wordcount/input -> gpl.txt***



8. Check DataNodes where are the blocks
9. Click on **View Block Info** from DataNode 192.168.1.1:50010 (**elephant**)
10. Search value from parameter "***block_name***" and copy number
    example: *blk_7874661894585597641*

11. Login as root

    ```
    $ sudo su -
    ```

12. Search file in /disk1 or /disk2

```
# find /disk1/dfs/dn/current/ -name blk_7874661894585597641
/disk1/dfs/dn/current/BP-936788720-192.168.1.1-
1412634379960/current/finalized/blk_7874661894585597641
```

13. Go to path from file

```
# cd /disk1/dfs/dn/current/BP-936788720-192.168.1.1-
1412634379960/current/finalized/blk_7874661894585597641

# ls -l
total 40
-rw-r--r-- 1 hdfs hdfs 35147 Oct  7 01:10 blk_7874661894585597641
-rw-r--r-- 1 hdfs hdfs   283 Oct  7 01:10 blk_7874661894585597641_1002.meta
```

14. View file blk_[id]

```
# tail blk_7874661894585597641
```

15. Modify file

```
# cat >> blk_7874661894585597641
new line
[Control + D]
```

16. Exit as root and read from Hadoop HDFS

```
$ hadoop fs -tail wordcount/input/gpl.txt
```

17. What happened?

18. Return to http://elephant:50070/dfshealth.jsp and click on ***Browse the filesystem -> /user/cloudera/wordcount/input -> gpl.txt***

19. Check the health of HDFS

Hadoop includes the dfsadmin command line tool for HDFS administration functionality. This tool allows the user to view the status of the HDFS cluster

```
$ sudo -u hdfs hadoop dfsadmin -report
```

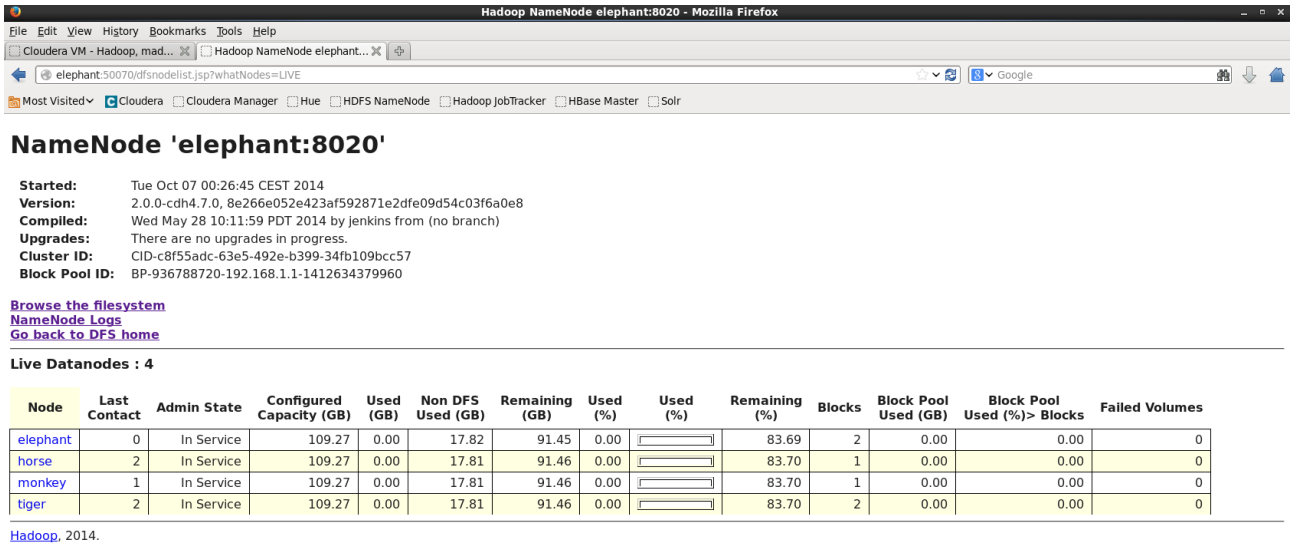
```
DFS Used: 94208 (92 KB)
Non DFS Used: 19127259136 (17.81 GB)
DFS Remaining: 98204123136 (91.46 GB)
DFS Used%: 0.00%
DFS Remaining%: 83.70%
Last contact: Tue Oct 07 01:34:56 CEST 2014


Name: 192.168.1.1:50010 (elephant)
Hostname: elephant
Decommission Status : Normal
Configured Capacity: 117331476480 (109.27 GB)
DFS Used: 110592 (108 KB)
Non DFS Used: 19135475712 (17.82 GB)
DFS Remaining: 98195890176 (91.45 GB)
DFS Used%: 0.00%
DFS Remaining%: 83.69%
Last contact: Tue Oct 07 01:34:55 CEST 2014


Name: 192.168.1.2:50010 (tiger)
Hostname: tiger
Decommission Status : Normal
Configured Capacity: 117331476480 (109.27 GB)
DFS Used: 110592 (108 KB)
Non DFS Used: 19127955456 (17.81 GB)
DFS Remaining: 98203410432 (91.46 GB)
DFS Used%: 0.00%
DFS Remaining%: 83.70%
Last contact: Tue Oct 07 01:34:56 CEST 2014


Name: 192.168.1.4:50010 (monkey)
Hostname: monkey
Decommission Status : Normal
Configured Capacity: 117331476480 (109.27 GB)
DFS Used: 69632 (68 KB)
Non DFS Used: 19127488512 (17.81 GB)
DFS Remaining: 98203918336 (91.46 GB)
DFS Used%: 0.00%
DFS Remaining%: 83.70%
Last contact: Tue Oct 07 01:34:56 CEST 2014
```

20. Go to http://elephant:50070/dfsnodelist.jsp and click on **Live Nodes**



NameNode 'elephant:8020'

Started: Tue Oct 07 00:26:45 CEST 2014
Version: 2.0.0-cdh4.7.0, 8e266e052e423af592871e2dfe09d54c03f6a0e8
Compiled: Wed May 28 10:11:59 PDT 2014 by jenkins from (no branch)
Upgrades: There are no upgrades in progress.
Cluster ID: CID-c8f55adc-63e5-492e-b399-34fb109bcc57
Block Pool ID: BP-936788720-192.168.1.1-1412634379960

Browse the filesystem
NameNode Logs
Go back to DFS home

Live Datanodes : 4

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)	Remaining (%)	Blocks	Block Pool Used (GB)	Block Pool Used (%)> Blocks	Failed Volumes
elephant	0	In Service	109.27	0.00	17.82	91.45	0.00		83.69	2	0.00	0.00	0
horse	2	In Service	109.27	0.00	17.81	91.46	0.00		83.70	1	0.00	0.00	0
monkey	1	In Service	109.27	0.00	17.81	91.46	0.00		83.70	1	0.00	0.00	0
tiger	2	In Service	109.27	0.00	17.81	91.46	0.00		83.70	2	0.00	0.00	0

Hadoop, 2014.

21. Check size from terminal, 109Gb or 55Gb?

```
$ df -h
```

We configured tow paths (/disk1 and /disk2) as DN and NN devices for Hadoop. But the device has only 55Gb size.