

Data System Architecture Assignment 2.

HDFS and Map Reduce

Name : Bhavesh Waghela

Student ID : 200532173

Table of Content

1. Transfer a file to HDP and run few HDFS commands
 - a. Create a test file which is more than 500 MB (for example zip a movie) then upload it to HDP then run the command mentioned in the lecture (HDFS -Part 2 Lecture - Slide 48 -53)
 - i. Use WinSCP File Transfer to Local System.
 - ii. HDFS create a directory and transfer the RAR file to the server.
 - b. Use the following link and explore the commands for fsck and dfs . Choose five commands from fsk and five commands from dfs. make sure you explain the command along with screen shots)
 - i. **DFS Commands**
 1. dfs -du
 2. dfs appendToFile
 3. dfs -ls
 4. dfs -mkdir
 5. dfs -put

6. `dfs -rmr`
- ii. FSCK Commands
 1. `fsck block`
 2. `fsck file`
 3. `fsck location`
 4. `fsck racks`
 5. `fsck openforwrite`

2. Run Map Reduce Jar File

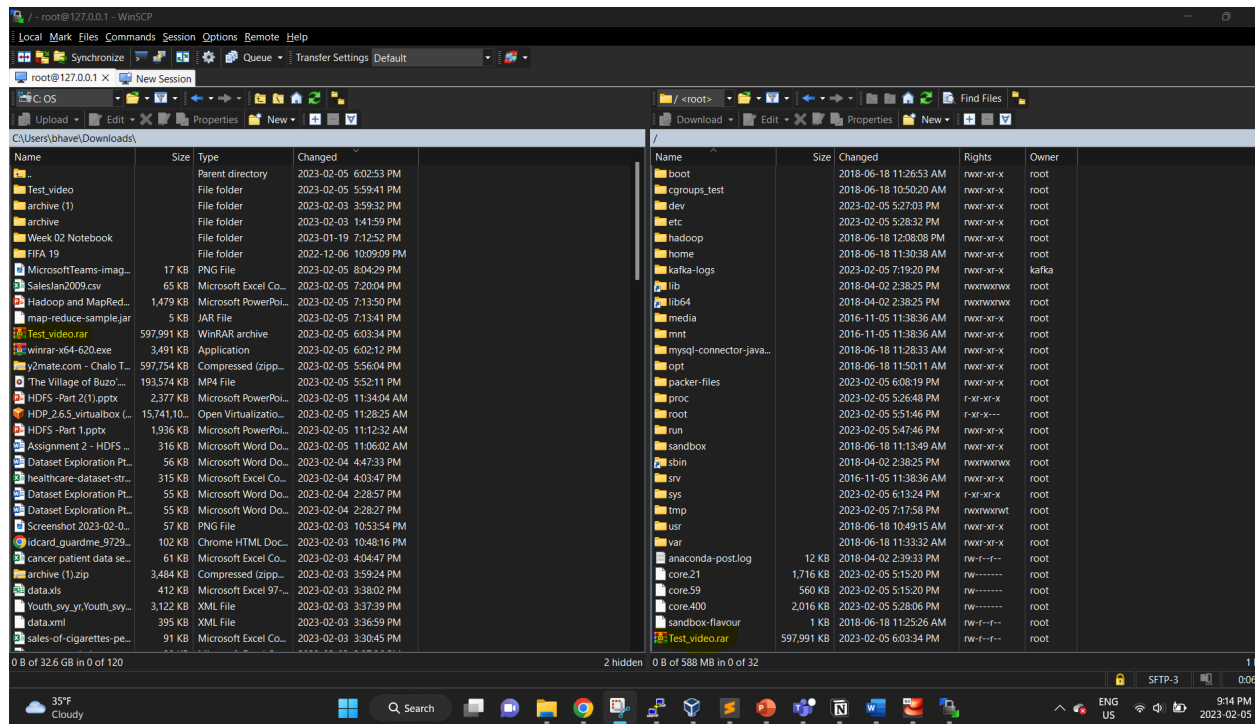
- a. Follow the slides (Week 7 – Hadoop Map Reduce Framework page 73-77). Copy salesjan2009 and jar file to your Hadoop cluster then run the following command.
 - i. Use WinSCP File Transfer to Local System.
 - ii. Login to putty.
 - iii. hadoop command execution for map reduce.

Part 1 - Transfer a file to HDP and run few HDFS commands

A. Create a test file which is more than 500 MB (for example zip a movie) then upload it to HDP then run the command mentioned in the lecture (HDFS -Part 2 Lecture - Slide 48 -53)

1. Use WinSCP File Transfer to Local System.

- Download a video of above 500MB and save the same in a .rar file.
- Connect to WinSCP with the local IP address and transfer the file to linux system.
- Connect to the local system and check if the file is available where it was dropped.



- HDFS create a directory and transfer the RAR file to the server.
 - Log in to putty with the hdfs root user and password.
 - Switch the user to hdfs user and create a directory to copy the file into the same.
 - Copy file from the local machine to hdfs for further processing the file into chunks.

```
[hdfs@sandbox-hdp /]$ hdfs dfs -mkdir /test_data
mkdir: '/test_data': File exists
[hdfs@sandbox-hdp /]$ cd /test_data
-bash: cd: /test_data: No such file or directory
[hdfs@sandbox-hdp /]$ hdfs dfs -mkdir /test_data
mkdir: '/test_data': File exists
[hdfs@sandbox-hdp /]$ hdfs dfs -mkdir /test_data2
[hdfs@sandbox-hdp /]$ hdfs dfs -ls /
Found 13 items
drwxrwxrwx - yarn  hadoop      0 2018-06-18 15:18 /app-logs
drwxr-xr-x - hdfs  hdfs        0 2018-06-18 16:19 /apps
drwxr-xr-x - yarn  hadoop      0 2018-06-18 14:52 /ats
drwxr-xr-x - hdfs  hdfs        0 2018-06-18 14:52 /hdp
drwx----- liiv  hdfs        0 2018-06-18 15:11 /liiv2-recovery
drwxr-xr-x - mapred hadoop      0 2018-06-18 14:52 /mapred
drwxrwxrwx - mapred hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs  hdfs        0 2018-06-18 15:59 /ranger
drwxrwxrwx - spark  hadoop      0 2023-02-05 23:26 /spark2-history
drwxr-xr-x - root   hdfs        0 2023-02-05 23:19 /test_data
drwxr-xr-x - hdfs  hdfs        0 2023-02-05 23:26 /test_data2
drwxrwxrwx - hdfs  hdfs        0 2018-06-18 16:06 /tmp
drwxr-xr-x - hdfs  hdfs        0 2018-06-18 16:08 /user
[hdfs@sandbox-hdp /]$ hdfs dfs -copyfromLocal Test_video.rar /test_data2/

[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$ hdfs dfs -ls /test_data2
Found 1 items
-rw-r--r-- 1 hdfs hdfs 612342478 2023-02-05 23:28 /test_data2/Test_video.rar
[hdfs@sandbox-hdp /]$
```

d. Using fsck command

```
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$ hdfs dfs -ls /test_data2
Found 1 items
-rw-r--r-- 1 hdfs hdfs 612342478 2023-02-05 23:28 /test_data2/Test_video.rar
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$ hdfs fsck /test_data2/Test_video.rar
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&path=/test_data2/Test_video.rar
FSCK started by hdfs (auth:SIMPLE) from /172.16.0.2 for path /test_data2/Test_video.rar at Sun Feb 05 23:33:35 UTC 2023
Status: HEALTHY
Total size: 612342478 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 122468495 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Sun Feb 05 23:33:35 UTC 2023 in 23 milliseconds

The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
[hdfs@sandbox-hdp /]$
```

f. Check if the list is corrupted.

```
The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
[hdfs@sandbox-hdp /]$ hdfs fsck /test_data2/Test_video.rar -list-corruptfileblocks
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&listcorruptfileblocks=1&path=/test_data2/Test_video.rar
The filesystem under path '/test_data2/Test_video.rar' has 0 CORRUPT files
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
```

g. check the summary of the transfer with the help of -file and -blocks parameters

```
The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
[hdfs@sandbox-hdp /]$ hdfs fsck /test_data2/Test_video.rar -list-corruptfileblocks
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&listcorruptfileblocks=1&path=/test_data2/Test_video.rar
The filesystem under path '/test_data2/Test_video.rar' has 0 CORRUPT files
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
[hdfs@sandbox-hdp /]$
```

B. Use the following link and explore the commands for fsck and dfs . Choose five commands from fsk and five commands from dfs. make sure you explain the command along with screen shots)

DSF Commands

1. Append Command - Appends the content to the file which is present on HDFS. Append single source. or multiple sources from the local file system to the destination file system.

```
[root@sandbox-hdp ~]# sudo -i hdfs
[hdfs@sandbox-hdp ~]#
[hdfs@sandbox-hdp ~]# hdfs dfs -ls /
Found 13 items
drwxr-xr-x - yarn    hadoop      0 2023-02-06 00:31 /app-logs
drwxr-xr-x - hdfs    hdfs      0 2018-06-18 16:13 /apps
drwxr-xr-x - yarn    hadoop      0 2018-06-18 14:52 /ats
drwxr-xr-x - hdfs    hdfs      0 2018-06-18 14:52 /hdp
drwxr-xr-x - livy    hdfs      0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x - mapred  hdfs      0 2018-06-18 14:52 /mapred
drwxr-xr-x - mapred  hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs    hdfs      0 2018-06-18 15:55 /ranger
drwxr-xr-x - spark   hadoop      0 2023-02-06 01:15 /spark2-history
drwxr-xr-x - root    hdfs      0 2023-02-05 23:19 /test_data
drwxr-xr-x - hdfs    hdfs      0 2023-02-06 00:44 /test_data2
drwxr-xr-x - hdfs    hdfs      0 2018-06-18 16:06 /tmp
drwxr-xr-x - hdfs    hdfs      0 2023-02-06 00:31 /user
[hdfs@sandbox-hdp ~]# pwd
/home/hdfs
[hdfs@sandbox-hdp ~]# ls
dog.jfif map-reduce-sample.jar SalesJan2009.csv
[hdfs@sandbox-hdp ~]# hdfs dfs -appendToFile dog.jfif /test_data2/Test_video.rar
[hdfs@sandbox-hdp ~]# hdfs dfs -ls /test_data2/
Found 4 items
-rw-r--r-- 1 hdfs hdfs      65835 2023-02-06 00:20 /test_data2/SalesJan2009.csv
-rw-r--r-- 1 hdfs hdfs 612356983 2023-02-06 01:20 /test_data2/Test_video.rar
drwxr-xr-x - hdfs hdfs      0 2023-02-06 00:43 /test_data2/map_output
drwxr-xr-x - hdfs hdfs      0 2023-02-06 00:45 /test_data2/map_output2
[hdfs@sandbox-hdp ~]#
```

2. dfs du - Du command is used to How much file Occupied in the disk. The field is the base size of the file or directory before replication.

```
[hdfs@sandbox-hdp ~]#
[hdfs@sandbox-hdp ~]# hdfs dfs -du /test_data2/
65835      /test_data2/SalesJan2009.csv
612356983 /test_data2/Test_video.rar
43         /test_data2/map_output2
[hdfs@sandbox-hdp ~]#
[hdfs@sandbox-hdp ~]#
[hdfs@sandbox-hdp ~]#
```

3. dfs -ls - HDFS ls command is used to display the list of Files and Directories in HDFS, This ls command shows the files with permissions, user, group, and other details

```
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /
Found 12 items
drwxrwxrwx - yarn  hadoop      0 2018-06-18 15:18 /app-logs
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 16:13 /apps
drwxr-xr-x - yarn  hadoop      0 2018-06-18 14:52 /ats
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 14:52 /hdp
drwx----- - livy  hdfs      0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x - mapred hdfs      0 2018-06-18 14:52 /mapred
drwxrwxrwx - mapred hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 15:59 /ranger
drwxrwxrwx - spark hadoop      0 2023-02-05 23:11 /spark2-history
drwxr-xr-x - root  hdfs      0 2023-02-05 22:40 /test_data
drwxrwxrwx - hdfs  hdfs      0 2018-06-18 16:06 /tmp
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 16:08 /user
[hdfs@sandbox-hdp ~]$ hdfs dfs -copyFromLocal test_file.rar /test_data/
```

4. `dfs -mkdir` - HDFS `mkdir` command is used to create a directory in HDFS. By default, this directory would be owned by the user who is creating it.

```
mkdir: /test_data: file exists
[hdfs@sandbox-hdp /]$ hdfs dfs -mkdir /test_data2
[hdfs@sandbox-hdp /]$ hdfs dfs -ls /
Found 13 items
drwxrwxrwx - yarn  hadoop      0 2018-06-18 15:18 /app-logs
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 16:13 /apps
drwxr-xr-x - yarn  hadoop      0 2018-06-18 14:52 /ats
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 14:52 /hdp
drwx----- - livy  hdfs      0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x - mapred hdfs      0 2018-06-18 14:52 /mapred
drwxrwxrwx - mapred hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs  hdfs      0 2018-06-18 15:59 /ranger
drwxrwxrwx - spark hadoop      0 2023-02-05 23:26 /spark2-history
```

5. `dfs put` - Copy file/folder from local disk to HDFS. On `put` command specifies the `local-file-path` where you wanted to copy from and then `hdfs-file-path` where you wanted to copy to on hdfs.

```
[hdfs@sandbox-hdp ~]$
[hdfs@sandbox-hdp ~]$
[hdfs@sandbox-hdp ~]$ hdfs dfs -put dog.jfif /tmp
[hdfs@sandbox-hdp ~]$ hdfs dfs -ls /tmp
Found 4 items
-rw-r--r-- 1 hdfs hdfs 14505 2023-02-06 01:49 /tmp/dog.jfif
drwxrwxr-x - druid hadoop 0 2018-06-18 16:06 /tmp/druid-indexing
drwxr-xr-x - hdfs hdfs 0 2018-06-18 14:52 /tmp/entity-file-history
drwx-wx-wx - ambari-ga hdfs 0 2018-06-18 15:12 /tmp/hive
[hdfs@sandbox-hdp ~]$
```

6. `dfs rmr` - HDFS `rm` command deletes a file and a directory from HDFS recursively.

```
hdfs@sandbox-hdp:~$ login as: root
root@127.0.0.1's password:
Last login: Mon Feb  6 01:08:20 2023 from 172.18.0.3
[root@sandbox-hdp ~]# sudo -iu hdfs
[hdfs@sandbox-hdp ~]# hdfs dfs -rm /test_data2/map_output
rm: '/test_data2/map_output': Is a directory
[hdfs@sandbox-hdp ~]# hdfs dfs -rmr /test_data2/map_output
rmr: DEPRECATED: Please use 'rm -r' instead
23/02/06 01:44:57 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/test_data2/map_output' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/hdfs/.Trash/
[hdfs@sandbox-hdp ~]# hdfs dfs rm-r /test_data2/map_output
rm-r: Unknown command
Usage: hadoop fs [generic options]
[-appendToFile <localsrc> ... <dst>]
[-cat [-ignoreCrc] <src> ...]
[-checksum <src> ...]
[-chgrp [-R] GROUP PATH...]
[-chmod [-R] <MODE[,MODE]... [-OCTALMODE] PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
[-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-count [-n] [-h] [-v] [-t <storage type>]] [-u] <path> ...]
[-cp [-f] [-p] [-p[topax]] <src> ... <dst>]
[-createSnapshot <snapshotDir> [<snapshotName>]]
[-deleteSnapshot <snapshotDir> [<snapshotName>]]
[-df [-h] <path> ...]]
[-du [-h] [-h] <path> ...]
[-expunge]
[-find <path> ... <expression> ...]
[-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
[-getfacl [-R] <path>]
[-getfattr [-R] [-n name] [-d] [-e en] <path>]
[-getmerge [-nl] <src> <localdst>]
[-help [cmd ...]]
[-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-l] [-u] <path> ...]]
[-mkdir [-p] <path> ...]
[-moveFromLocal <localsrc> ... <dst>]
[-moveToLocal <src> <localdst>]
[-mv <src> ... <dst>]
[-put [-f] [-p] [-l] <localsrc> ... <dst>]
[-renameSnapshot <snapshotDir> <oldName> <newName>]
[-rm [-f] [-r] [-R] [-skipTrash] [-safely] <src> ...]
[-rmall [--ignore-fail-on-non-empty] <dir> ...]
[-setfacl [-R] [(b|k) (-m|-x <acl_spec>) <path>]] [--set <acl_spec> <path>]]
[-setfattr [-n name [-v value] | -x name] <path>]
[-setrep [-n] [-w] <rep> <path> ...]
[-stat [format] <path> ...]
```

FSCK Commands

1. fsck block - Print out block report.

```
[hdfs@sandbox-hdp ~]# hdfs fsck /test_data2/Test.Video.rar -block
fsck: illegal option '-block'
Usage: hdfs fsck <path> [-list-corruptfileblocks] [-move | -delete | -openforwrite] [-files [-blocks [-locations] [-racks]]] [-includeSnapshots] [-storagepolicies] [-blockId <blk_id>]
    <path> start checking from this path
    -move move corrupted files to /lost+found
    -delete delete corrupted files
    -files print out files being checked
    -openforwrite print out files opened for write
    -includeSnapshots include snapshot data if the given path indicates a snapshottable directory or there are snapshottable directories under it
    -list-corruptfileblocks print out list of missing blocks and files they belong to
    -blocks print out block report
    -locations print out locations for every block
    -racks print out network topology for data-node locations
    -storagepolicies print out storage policy summary for the blocks
    -blockId print out which file this blockId belongs to, locations (nodes, racks) of this block, and other diagnostics info (under replicated, corrupted or not, etc)
    -replicadetails print out each replica details

Please Note:
    1. By default fsck ignores files opened for write, use -openforwrite to report such files. They are usually tagged CORRUPT or HEALTHY depending on their block allocation status
    2. Option -includeSnapshots should not be used for comparing stats, should be used only for HEALTH check, as this may contain duplicates if the same file present in both original fs
    tree and inside snapshots.

Generic options supported are
    -conf <configuration file> specify an application configuration file
    -D <property=value> use value for given property
    -fs <local|namenode:port> specify a namenode
    -jt <local|resourcemanager:port> specify a ResourceManager
    -files <comma separated list of files> specify comma separated files to be copied to the map reduce cluster
    -libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.
    -archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]

[hdfs@sandbox-hdp ~]#
```

2. fsck file - Print out files being checked.

```
(hdfs@sandbox-hdp ~)$ hdfs fsck /test_data2/Test_video.rar -files
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&files=1&path=42Ftest_data2%2FTest_video.rar
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /test_data2/Test_video.rar at Mon Feb 06 16:22:03 UTC 2023
/test_data2/Test_video.rar 612356993 bytes, 5 block(s): OK
Status: HEALTHY
Total size: 612356993 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 122471396 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Feb 06 16:22:03 UTC 2023 in 3 milliseconds

The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
(hdfs@sandbox-hdp ~)$
```

3. fsck location - Print out locations for every block.

```
(hdfs@sandbox-hdp ~)$ hdfs fsck /test_data2/Test_video.rar -locations
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&locations=1&path=42Ftest_data2%2FTest_video.rar
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /test_data2/Test_video.rar at Mon Feb 06 16:25:17 UTC 2023
Status: HEALTHY
Total size: 612356993 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 122471396 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Feb 06 16:25:17 UTC 2023 in 0 milliseconds

The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
(hdfs@sandbox-hdp ~)$
```

4. fsck racks - Print out network topology for data-node locations.

```
(hdfs@sandbox-hdp ~)$ hdfs fsck /test_data2/Test_video.rar -racks
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsck?ugi=hdfs&racks=1&path=42Ftest_data2%2FTest_video.rar
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /test_data2/Test_video.rar at Mon Feb 06 16:26:00 UTC 2023
Status: HEALTHY
Total size: 612356993 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 122471396 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Feb 06 16:26:00 UTC 2023 in 0 milliseconds

The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
(hdfs@sandbox-hdp ~)$
```

5. fsck openforwrite - Print out files opened for write.


```

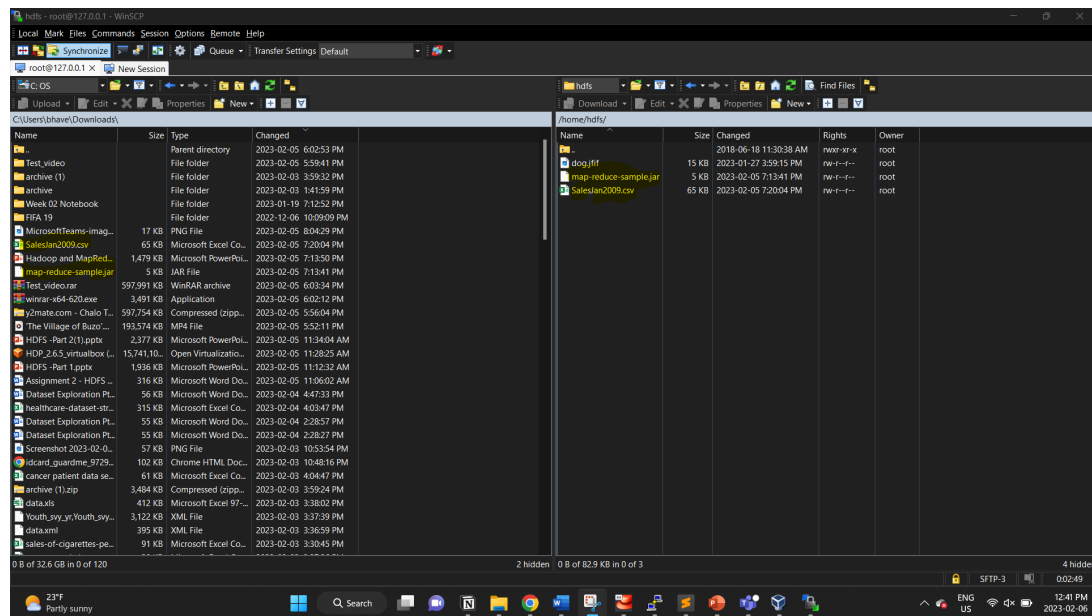
[hdfs@sandbox-hdp ~]$ hdfs fsck /test_data2/Test_video.rar openforwrite
Connecting to namenode via http://sandbox-hdp.hortonworks.com:50070/fsc?ugi=hdfs;openforwrite=1;path=/test_data2/Test_video.rar
FSCK started by hdfs (auth:SIMPLE) from /172.18.0.2 for path /test_data2/Test_video.rar at Mon Feb 06 16:35:21 UTC 2023
Status: HEALTHY
Total size: 612356983 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 5 (avg. block size 122471396 B)
Minimally replicated blocks: 5 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Feb 06 16:35:21 UTC 2023 in 1 milliseconds

The filesystem under path '/test_data2/Test_video.rar' is HEALTHY
[hdfs@sandbox-hdp ~]$ openforwrite

```

Part 2 - Run Map Reduce Jar File

1. Use WinSCP File Transfer to Local System.
 - a. Download the jar file and the csv file from the blackboard link.
 - b. Transfer the file from your windows to local machine via winscp to home/hdfs folder check for files with ls command.



2. Login to putty and run the following commands

- a. login into hdfs - `sudo -iu hdfs`
- b. Check the contents of the folder test_data - `hdfs dfs -ls /test_data/`
- c. Move the csv file to test_data folder - `hdfs dfs -copyFromLocal SalesJan2009.csv /test_data/`

```

hdfs@sandbox-hdp-~$
login as: root
root@127.0.0.1's password:
Last login: Sun Feb  5 23:11:28 2023 from 172.18.0.3
[root@sandbox-hdp ~]#
[root@sandbox-hdp ~]# pwd
/root
[root@sandbox-hdp ~]# cd /home/hdfs/
[root@sandbox-hdp hdfs]# ls
[root@sandbox-hdp hdfs]# ls
map-reduce-sample.jar SalesJan2009.csv
[root@sandbox-hdp hdfs]# ls -l
total 76
-rw-r--r-- 1 root root 4624 Feb  6 00:13 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Feb  6 00:20 SalesJan2009.csv
[root@sandbox-hdp hdfs]# sudo -iu hdfs
[hdfs@sandbox-hdp ~]# ls -l
total 76
-rw-r--r-- 1 root root 4624 Feb  6 00:13 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Feb  6 00:20 SalesJan2009.csv
[hdfs@sandbox-hdp ~]# hdfs dfs -ls /test_data2/
Found 1 items
-rw-r--r-- 1 hdfs hdfs 612342478 2023-02-05 23:28 /test_data2/test_video.rar
[hdfs@sandbox-hdp ~]# ls
map-reduce-sample.jar SalesJan2009.csv
[hdfs@sandbox-hdp ~]# ls -l
total 76
-rw-r--r-- 1 root root 4624 Feb  6 00:13 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Feb  6 00:20 SalesJan2009.csv
[hdfs@sandbox-hdp ~]# hadoop jar map-reduce-sample.jar /test_data2/c
[hdfs@sandbox-hdp ~]# hdfs dfs -copyFromLocal SalesJan2009.csv /test_data2/
[hdfs@sandbox-hdp ~]# ls -l
total 76
-rw-r--r-- 1 root root 4624 Feb  6 00:13 map-reduce-sample.jar
-rw-r--r-- 1 root root 65835 Feb  6 00:20 SalesJan2009.csv
[hdfs@sandbox-hdp ~]# ls -l /test_data2
ls: cannot access /test_data2: No such file or directory
[hdfs@sandbox-hdp ~]# ls -l /test_data2/
ls: cannot access /test_data2/: No such file or directory
[hdfs@sandbox-hdp ~]# hdfs dfs -ls /test_data2/
Found 2 items
-rw-r--r-- 1 hdfs hdfs 65835 2023-02-06 00:28 /test_data2/SalesJan2009.csv
-rw-r--r-- 1 hdfs hdfs 612342478 2023-02-05 23:28 /test_data2/test_video.rar
[hdfs@sandbox-hdp ~]# hadoop jar map-reduce-sample.jar /test_data2/SalesJan2009.csv /test_data2/map_output
23/02/06 00:31:29 INFO client.RMPProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
23/02/06 00:31:30 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
23/02/06 00:31:30 INFO client.RMPProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
23/02/06 00:31:30 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.18.0.2:10200
23/02/06 00:31:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

```

- d. Check the contents of the folder test_data - `hdfs dfs -ls /test_data/`

3. Run the following command to get the output - `hadoop jar map-reduce-sample.jar /test_data2/SalesJan2009.csv /test_data/map_output`

```
at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
(hdfs@sandbox-hdp ~)$
(hdfs@sandbox-hdp ~)$ hadoop jar map-reduce-sample.jar /test_data2/SalesJan2009.csv /test_data2/map_output5
23/02/06 00:43:59 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.16.0.2:8032
23/02/06 00:43:59 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.16.0.2:10200
23/02/06 00:43:59 INFO client.RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.16.0.2:8032
23/02/06 00:43:59 INFO client.AHSProxy: Connecting to Application History server at sandbox-hdp.hortonworks.com/172.16.0.2:10200
23/02/06 00:44:01 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/02/06 00:44:03 INFO mapred.FileInputFormat: Total input paths to process : 1
23/02/06 00:44:07 INFO mapreduce.JobSubmitter: number of splits:2
23/02/06 00:44:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675636832385_0002
23/02/06 00:44:09 INFO impl.YarnClientImpl: Submitted application application_1675636832385_0002
23/02/06 00:44:10 INFO mapreduce.Job: The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1675636832385_0002/
23/02/06 00:44:10 INFO mapreduce.Job: Running job: job_1675636832385_0002
23/02/06 00:44:30 INFO mapreduce.Job: Job job_1675636832385_0002 running in uber mode : false
23/02/06 00:44:30 INFO mapreduce.Job: map 0% reduce 0%
23/02/06 00:44:46 INFO mapreduce.Job: map 50% reduce 0%
23/02/06 00:44:47 INFO mapreduce.Job: map 100% reduce 0%
23/02/06 00:45:04 INFO mapreduce.Job: map 100% reduce 100%
23/02/06 00:45:09 INFO mapreduce.Job: Job job_1675636832385_0002 completed successfully
23/02/06 00:45:09 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=12824
  FILE: Number of bytes written=485658
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=98901
  HDFS: Number of bytes written=43
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
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)=26591
  Total time spent by all reduces in occupied slots (ms)=15751
  Total time spent by all map tasks (ms)=26591
  Total time spent by all reduce tasks (ms)=15751
  Total vcore-milliseconds taken by all map tasks=26591
  Total vcore-milliseconds taken by all reduce tasks=15751
  Total megabyte-milliseconds taken by all map tasks=6647750
  Total megabyte-milliseconds taken by all reduce tasks=3937750
```