Apache Hadoop

Working with Apache Hadoop HDFS

During this exercise, you become an Apache Hadoop/Analyst who will create physical objects in HDFS and perform some basic operations on them.

Please use your Linux account at the edge node (cdh00.cl.ii.pw.edu.pl)

Using hdfs command line interface

1.Login using ssh to edgenode cdh00.cl.ii.pw.edu.pl using your Linux account.

2. Go to exercise home directory, create a data folder and copy sample CSV file.

```
#you can replace ${USER} with your linux account name or use this variable in
your scripts

echo ${USER}


cd /data/local/datascience/home/
#create your home
mkdir ${USER}
cd ${USER}

#check if you are in your home dir (should be
/data/local/datascience/home/${USER})
pwd

#create a data dir
mkdir data

cp /data/local/datascience/data/measured_data.csv
/data/local/datascience/home/${USER}/data

#view first ten lines of the file
cat measured_data.csv | head
```

3a. Before you can access secured (by Kerberos) distributed file system you have to generate Kerberos ticket. When prompted provide your password and afterwards verify that ticket has been generated.

```
kinit

Password for <USERNAME>@CL.II.PW.EDU.PL:

klist
```

3b. List content of your home directory on HDFS (replace ${USER} with your account name at the edge node):

```
 hdfs dfs -ls /user/${USER}
```

4. Create a new directory in your home folder on HDFS. Do you know what “-p” is necessary in this operation? Check if the directory has been successfully created.

```
 hdfs dfs -mkdir -p /user/${USER}/external/measured_data

 hdfs dfs -ls -R  /user/${USER}/
```

5. Copy a CSV file from your home directory to HDFS

```
 cd /data/local/datascience/home/${USER}/data
 hdfs dfs -put measured_data.csv
 /user/${USER}/external/measured_data
```

6. Check if the file has been copied and then check it's size on HDFS.

7. Who is the owner of file?

```
 hdfs dfs -ls -R  /user/${USER}/external/measured_data

 hdfs dfs -du -h
```

```
/user/${USER}/external/measured_data/measured_data.csv
```

8. Get information about the “Health” of the file:

```
hdfs fsck /user/${USER}/external/measured_data/measured_data.csv

Connecting to namenode via
http://hnn.bkw-hdp.ch:50070/fsck?ugi=sar_wim&path=%2Fuser%2Fsar_wim%2Fext
ernal%2Fmeasured_data%2Fmeasured_data.csv
FSCK started by sar_wim (auth:SIMPLE) from /10.10.0.3 for path
/user/sar_wim/external/measured_data/measured_data.csv at Thu May 26
13:07:22 CEST 2016
.Status: HEALTHY
 Total size:    331035256 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      3 (avg. block size 110345085 B)
 Minimally replicated blocks:   3 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     3.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          4
 Number of racks:               1
FSCK ended at Thu May 26 13:07:22 CEST 2016 in 7 milliseconds
```

9. Check the content of the file:

```
hdfs dfs -cat /user/${USER}/external/measured_data/measured_data.csv | head

1|2015-05-18 19:24:00|236|0.09920929584628235|kW|s|D|Warsaw-Dereniowa
2|2015-05-18 19:24:00|237|0.8467751295230832|kW|s|D|Warsaw-Dereniowa
3|2015-05-18 19:24:00|238|0.4608953190788161|kW|s|D|Warsaw-Dereniowa
4|2015-05-18 19:24:00|239|0.029231154861803166|kW|s|D|Warsaw-Dereniowa
5|2015-05-18 19:24:00|240|0.7698375647136683|kW|s|D|Warsaw-Dereniowa
6|2015-05-18 19:24:00|241|0.7748969324130782|kW|s|D|Warsaw-Dereniowa
7|2015-05-18 19:24:00|242|0.20359460405586638|kW|s|D|Warsaw-Dereniowa
```

```
8|2015-05-18 19:24:00|243|0.452764286169582|kW|s|D|Warsaw-Dereniowa
9|2015-05-18 19:24:00|244|0.4235218062718823|kW|s|D|Warsaw-Dereniowa
10|2015-05-18 19:24:00|245|0.9795513878598727|kW|s|D|Warsaw-Dereniowa
```

10. Create a new directory and copy the file between two locations

```
hdfs dfs -mkdir /user/${USER}/external/temp

hdfs dfs -cp /user/${USER}/external/measured_data/measured_data.csv
/user/${USER}/external/temp

hdfs dfs -ls /user/${USER}/external/temp/measured_data.csv
```

11. Remove the file you copied and try to do the rollback using the Trash.

12. Try to repeat the same operation using -skipTrash parameter:

```
hdfs dfs -rm /user/${USER}/external/temp/measured_data.csv

17/05/18 20:17:14 INFO fs.TrashPolicyDefault: Moved:
'hdfs://cdh01.cl.ii.pw.edu.pl:8020/user/xmwiewio/external/temp/measured_d
ata.csv' to trash at:
hdfs://cdh01.cl.ii.pw.edu.pl:8020/user/xmwiewio/.Trash/Current/user/xmwie
wio/external/temp/measured_data.csv

####rollback
#check if the file exists int the Trash
hdfs dfs -ls  /user/${USER}/.Trash/Current/user/${USER}/external/temp
Found 1 items
-rw-r--r--   3 xmwiewio supergroup  505361782 2017-05-18 20:16
/user/xmwiewio/.Trash/Current/user/xmwiewio/external/temp/measured_data.c
sv

#mv
 hdfs  dfs -mv
/user/${USER}/.Trash/Current/user/${USER}/external/temp/measured_data.csv
/user/${USER}/external/temp/

#ls
hdfs dfs -ls /user/${USER}/external/temp/
```

```
#using skipTrash
hdfs dfs -rm -skipTrash /user/${USER}/external/temp/measured_data.csv
Deleted /user/sar_wim/external/temp/measured_data.csv

#remove the folder recursively
hdfs dfs -rm -r  /user/${USER}/external/temp/
```

13.Try to get some help on using “-mkdir” option:

```
hdfs dfs -help mkdir
-mkdir [-p] <path> ... :
  Create a directory in specified location.
  -p  Do not fail if the directory already exists
```

14. Download the CSV file back your linux home directory at the edgenode:

```
mkdir /data/local/datascience/home/${USER}/data/download
cd /data/local/datascience/home/${USER}/data/download
hdfs dfs -get /user/${USER}/external/measured_data/*.csv .
```

15. Change the permissions of the directory so that other users can read your file.

16. Check if it has been changed accordingly.

```
hdfs dfs -chmod -R 777
/user/${USER}/external/measured_data/

hdfs dfs -ls  /user/${USER}/external
```