HDFS Additional Usecases

**Note:** This use cases covers some additional commands and extreme usecases, please ensure to try all of them.

1. Create a new directory in linux namely ~/install/hdfsusecases and create a new file inside the above directory namely ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt copying the first 1000 lines from an existing file

~/pigdata/NYSE\_daily

1. Create another new file inside the above directory namely ~/install/hdfsusecases/NYSE\_2020\_06\_21.txt copying the line from 1001 to 2000 from an existing file ~/pigdata/NYSE\_daily
2. Create a directory in Hadoop namely /tmp/hdfsusecases
3. Check whether the above directory is created in HDFS or not using the below command (**Note:** We use –test –d option to check whether the given path is a directory or not)

# hadoop fs -test -d /tmp/hdfsusecases

1. Check what is the status code of the above command using, if it shows 0 then directory is created, if shows non zero then the directory is not created then check the step 3 again.

# echo $?

1. Copy file generated only in step 1 (~/install/hdfsusecases/NYSE\_2020\_06\_20.txt) from linux to hdfs directory

/tmp/hdfsusecases in the name of NYSE\_2020\_06.txt

1. Like step 4 and 5, check whether the above file (/tmp/hdfsusecases/NYSE\_2020\_06.txt) is created or not in HDFS, using -f option and check for the status code using $? and create a zero byte file in HDFS directory

/tmp/hdfsusecases in the name of \_SUCCESS

1. Append the file generated in step 2 in linux (~/install/hdfsusecases/NYSE\_2020\_06\_20.txt) with the file generated in step 6 in the hdfs directory /tmp/hdfsusecases/NYSE\_2020\_06.txt
2. Count the size of the file in HDFS /tmp/hdfsusecases/NYSE\_2020\_06.txt
3. Count the number of rows are there in the /tmp/hdfsusecases/NYSE\_2020\_06.txt (Which should show the total count of the files created in step1 and 2)
4. Display only line 11 to 20 from the file in HDFS /tmp/hdfsusecases/NYSE\_2020\_06.txt
5. Store line 11 to 20 from the file in HDFS /tmp/hdfsusecases/NYSE\_2020\_06.txt into linux file namely

~/install/hdfsusecases/NYSE\_sampledata1.txt

1. Delete the line number 1 from the HDFS file /tmp/hdfsusecases/NYSE\_2020\_06.txt , for example if the above file contains 100 rows, after deletion it should have only 99 rows in HDFS

**Note:** we can’t do this directly because of the WORM property of HDFS data, think about the possible work around and try to achive the result

1. Copy the above file /tmp/hdfsusecases/NYSE\_2020\_06.txt in the name of

/tmp/hdfsusecases/NYSE\_2020\_06\_bkp.txt

1. Set the blocksize 64MB while writing the file in HDFS, check in the UI how many blocks are generated

# hadoop fs -D dfs.block.size=67108864 -put /home/hduser/install/hadoop-2.7.1.tar.gz /user/hduser/

1. Set the blocksize 128MB (134217728) for the same file generated in step 16 and replace the existing file in HDFS.
2. Set the replication to 3 while writing the file in HDFS

# hadoop fs -D dfs.replication=3 -put /home/hduser/install/hadoop-2.7.1.tar.gz /user/hduser/

1. **Important Command** DistCp (distributed copy) is a tool used for copying data between one Hadoop cluster to another cluster or with in the same cluster using mappers. (*Interview Question – how do you copy data from production Hadoop cluster to Dev Hadoop cluster*)

# hadoop distcp hdfs://localhost:54310/user/hduser/hadoop-2.7.1.tar.gz hdfs://localhost:54310/user/hduser/hadoop/