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

mkdir -p ~/install/hdfsusecases

head -1000 ~/pigdata/NYSE\_daily > ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt

ls -lrt ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt

-rw-rw-r-- 1 hduser hduser 57446 Mar 7 17:31 /home/hduser/install/hdfsusecases/NYSE\_2020\_06\_20.txt

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

ls -lrt ~/install/hdfsusecases/NYSE\_2020\_06\_21.txt

-rw-rw-r-- 1 hduser hduser 55552 Mar 7 17:33 /home/hduser/install/hdfsusecases/NYSE\_2020\_06\_21.txt

1. Create a directory in Hadoop namely /tmp/hdfsusecases

\ hdfs dfs -mkdir -p /tmp/hdfsusecases

hdfs dfs -ls /tmp/hdfsusecases

1. 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

hdfs dfs -put ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt /tmp/hdfsusecases/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

hdfs dfs -put ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt /tmp/hdfsusecases/NYSE\_2020\_06.txt

hdfs dfs -ls /tmp/hdfsusecases/NYSE\_2020\_06.txt

-rw-r--r-- 1 hduser supergroup 57446 2022-03-07 17:36 /tmp/hdfsusecases/NYSE\_2020\_06.txt

[hduser@Inceptez ~]$ echo $?

0

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

hdfs dfs -appendToFile ~/install/hdfsusecases/NYSE\_2020\_06\_20.txt /tmp/hdfsusecases/NYSE\_2020\_06.txt

\hdfs dfs -ls /tmp/hdfsusecases/NYSE\_2020\_06.txt

-rw-r--r-- 1 hduser supergroup 114892 2022-03-07 17:38 /tmp/hdfsusecases/NYSE\_2020\_06.txt

1. Count the size of the file in HDFS /tmp/hdfsusecases/NYSE\_2020\_06.txt

hdfs dfs -count /tmp/hdfsusecases/NYSE\_2020\_06.txt

0 1 114892 /tmp/hdfsusecases/NYSE\_2020\_06.txt

1. 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)

hdfs dfs -cat /tmp/hdfsusecases/NYSE\_2020\_06.txt | wc -l

2000

1. Display only line 11 to 20 from the file in HDFS /tmp/hdfsusecases/NYSE\_2020\_06.txt

hdfs dfs -cat /tmp/hdfsusecases/NYSE\_2020\_06.txt | head -20 | tail

NYSE|CLI|2009-12-16|34.79|35.10|34.48|34.66|1007900|34.21

NYSE|CLI|2009-12-15|34.60|34.91|34.39|34.84|813200|34.39

NYSE|CLI|2009-12-14|34.21|34.90|33.86|34.82|987700|34.37

NYSE|CLI|2009-12-11|33.55|34.08|33.40|34.00|836500|33.56

NYSE|CLI|2009-12-10|33.61|33.80|33.09|33.26|1296300|32.83

NYSE|CLI|2009-12-09|33.25|33.71|33.01|33.25|863900|32.82

NYSE|CLI|2009-12-08|32.57|33.80|32.52|33.18|890000|32.75

NYSE|CLI|2009-12-07|33.63|33.77|32.78|33.00|961800|32.57

NYSE|CLI|2009-12-04|33.46|34.12|33.16|33.69|1412400|33.25

NYSE|CLI|2009-12-03|32.55|33.43|32.48|32.65|1170600|32.22

1. 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

hdfs dfs -cat /tmp/hdfsusecases/NYSE\_2020\_06.txt | head -20 | tail > ~/install/hdfsusecases/NYSE\_sampledata1.txt

ls -lrt ~/install/hdfsusecases/NYSE\_sampledata1.txt

-rw-rw-r-- 1 hduser hduser 574 Mar 7 17:42 /home/hduser/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

dfs dfs -get /tmp/hdfsusecases/NYSE\_2020\_06.txt NYSE\_2020\_06.txt

wc -l NYSE\_2020\_06.txt

2000 NYSE\_2020\_06.txt

tail -1999 NYSE\_2020\_06.txt > NYSE\_2020\_06.txt\_eidted

hdfs dfs -put -f NYSE\_2020\_06.txt\_eidted /tmp/hdfsusecases/NYSE\_2020\_06.txt

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

hdfs dfs -cp /tmp/hdfsusecases/NYSE\_2020\_06.txt /tmp/hdfsusecases/NYSE\_2020\_06\_bkp.txt

hdfs dfs -ls /tmp/hdfsusecases/NYSE\_2020\_06\_bkp.txt

-rw-r--r-- 1 hduser supergroup 114835 2022-03-07 17:45 /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 15 and replace the existing file in HDFS.

# hadoop fs -D dfs.block.size= 134217728 -put /home/hduser/install/hadoop-2.7.1.tar.gz /user/hduser/

1. 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/