**continuous merging small files into HDFS using streaming API and cron**

step 1 : create a tmp directory

hadoop fs -mkdir tmp

step 2 : move all the small files to the tmp directory at a point of time

hadoop fs -mv input/\*.txt tmp

step 3 -merge the small files with the help of hadoop-streaming jar

hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \

-Dmapred.reduce.tasks=1 \

-input "/user/abc/input" \

-output "/user/abc/output" \

-mapper cat \

-reducer cat

step 4- move the output to the input folder

hadoop fs -mv output/part-00000 input/large\_file.txt

step 5 - remove output

hadoop fs -rm -R output/

step 6 - remove all the files from tmp

hadoop fs -rm tmp/\*.txt

Create a shell script from step 2 till step 6 and schedule it to run at regular intervals to merge the smaller files at regular intervals (may be for every minute based on your need)

**Steps to schedule a cron job for merging small files**

step 1: create a shell script **/home/abc/mergejob.sh** with the help of above steps (2 to 6)

**important note:** you need to specify the absolute path of hadoop in the script to be understood by cron

#!/bin/bash

/home/abc/hadoop-2.6.0/bin/hadoop fs -mv input/\*.txt tmp

wait

/home/abc/hadoop-2.6.0/bin/hadoop jar /home/abc/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \

-Dmapred.reduce.tasks=1 \

-input "/user/abc/input" \

-output "/user/abc/output" \

-mapper cat \

-reducer cat

wait

/home/abc/hadoop-2.6.0/bin/hadoop fs -mv output/part-00000 input/large\_file.txt

wait

/home/abc/hadoop-2.6.0/bin/hadoop fs -rm -R output/

wait

/home/abc/hadoop-2.6.0/bin/hadoop fs -rm tmp/\*.txt

step 2: schedule the script using cron to run every minute using cron expression

a) edit crontab by choosing an editor

>crontab -e

b) add the following line at the end and exit from the editor

\* \* \* \* \* /bin/bash /home/abc/mergejob.sh > /dev/null 2>&1

The merge job will be scheduled to run for every minute.

Hope this was helpful.