**Assignment info:**

Team XX has created a data pipeline based on marketing data. The data consists of complex (nested) data structures, and contains over XX columns. Every day, the raw data is written to the cluster, based on a functional loaddate (ses\_dts). The data averages around 500 mb per day. For two years now, data has been written to the cluster in this way.  
  
Team XX created a folder structure that looks like this on the cluster:

$ hdfs dfs -ls /data/sample\_data/raw  
  
output:

Found 730 items  
/data/dis/sample\_data/raw/ ses\_dts=20181001  
 /data/dis/sample\_data/raw/ses\_dts=20181002  
/data/dis/sample\_data/raw/ses\_dts=20181003  
 /data/dis/sample\_data/raw/ses\_dts=20181004  
etc.  
etc.

The content of the folders look like this:  
  
hdfs dfs -ls /data/dis/sample\_data/raw/ses\_dts=20181001

Found 500 items  
/data/dis/sample\_data/raw/ ses\_dts=20181001/part-00000.parquet  
/data/dis/sample\_data/raw/ ses\_dts=20181001/part-00001. parquet  
/data/dis/sample\_data/raw/ ses\_dts=20181001/part-00002. parquet  
etc.  
etc.

As you can see, every parquet file averages around 1mb. Problem is that the small files are decreasing performance on the and the issue should be solved *for the data that is already on the cluster.*   
  
**Assignment:**Solve the small file issue *for the data that is already on the cluster.* Please take into account that,   
- (1) Your script will be run by YY and YY has no knowledge or interest in PySpark, and just want to run the scripts provided by you. The easier you make this job for YY, the happier YY will get.  
- (2) The cluster you are currently working on is a small cluster, so it is not possible to read in 2 years of marketing data at once without overwhelming the cluster. Instead, you need to read the data day by day, and solve the issue in that way. If you fail to read day-by-day, production will go down, and YY will get very angry.  
-(3) YY would like to run the script on times when the cluster is unoccupied. At the least, YY needs to be able to run the script(s) for just one day of data. However, more flexibility (running per week, month, or year) will make YY happier. YY does not enjoy running a single script 730 times.   
  
Please provide with your solution in the form of scripts. In addition, also explain how you came to this solution. For this assignment you can use both Spark Python (including Spark shell) and bash.