CSC 343: Big Data Programming and Management
Lab #2: File management in Docker and HDFS

For this assignment, follow the instructions below to make changes to the *gdancik/cloudera* container, save an updated image, with your username replacing *gdancik* in the image name, and push your updated image to your docker hub account. In addition, you should turn in your answers to the questions below. These answers must be submitted by hardcopy at the beginning of class on the due date.

Note: It is recommended that you periodically save your changes, by creating an updated image from the container you are modifying, and by pushing your updated image to Docker Hub.

This set of questions requires using the folders and files in the *logs.zip* file posted on the course page, http://gdancik.github.io.

To complete the steps below, create a container from the *gdancik/cloudera* image (or your own copy), which is running Hadoop as described in class.

1. Copy the file *log6-2018-12-31.txt* (in the folder *2018-12-31*) from your local machine to the /home/cloudera/ folder of the running container. What command did you use to copy this file?

2. In a single command, use the *docker cp* command to copy the *logs/* directory from your local machine to the directory */home/cloudera/*). This command will copy the *logs* directory, so that the container now has a *logs* directory that contains all of the log files. Note that when using the *docker cp* command, specifying the directory as *logs/* (with the slash) means to copy the folder, while specifying the directory as *logs* (without the slash) means to copy the *contents* of the folder (i.e., not the folder itself, as if you selected to copy *logs/**). What command did you use to copy these files?

3. In a single command, create a file named */home/cloudera/log_folders.txt* that lists all log folders in the *logs* directory. What is this command? Hint: how can you list all folders and files in a directory, and how can you redirect the output of a command to save the results in a file?

4. What is the command for copying the file */home/cloudera/log_folders.txt* to your Desktop on your local machine?

5. Create the directory */user/cloudera/logs* on HDFS, and specify the command used

6. Copy the file *log6-2018-12-31.txt* from your docker container to /user/cloudera/logs on HDFS. What is this command?

7. On HDFS, create the directories */user/cloudera/logs/2018-Dec* and */user/cloudera/logs/january*. In a single command, copy all log folders from December 2018 to */user/cloudera/logs/2018-Dec* and copy all January log files to */user/cloudera/logs/january*. What are the commands you used for copying these files?

8. Most Hadoop processing on HDFS data is done at the folder level. In other words, all files within a folder are processed. Therefore, if there are folders and files that we do not want to process, then we need to remove them. Files can removed with the command *hdfs dfs -rm*, while folders can be removed by adding the -r tag (which says to *recursively* remove files and folders). What is the command to remove the 2018-Dec log files corresponding to December 21 – December 25th?

9. Copy the 2018-Dec folder from HDFS to your container, and save this in the directory */home/cloudera/logs/2018-Dec-from-HDFS*.

10. For the remaining questions, you may use either *hdfs dfs* commands or HUE.

    a. Create the HDFS folder /user/cloudera/Holidays/NYE and add the logs from December 31st of each year (you should add each file to this folder).

    b. Modify the *log6-2018-12-31.txt* file stored on HDFS in /users/cloudera/logs to include a 2nd line which contains your name.