Assignment - Week 2

1. Login to your Gateway node & open a terminal

2. write a command to know what's your home directory in gateway node

3. There is a third party service which will drop a file named orders.csv in the landing folder under your home directory.

Then you need to filter for all the orders where status is PENDING_PAYMENT & create a new file named orders_filtered.csv and put it to the staging folder.

Then take this file and put it to hdfs in landing folder in your hdfs

and do a couple of more things...

So to simulate this..

1. create two folders named landing and staging in your home directory.

2. copy the file present under /data/retail_db/orders folder to the landing folder in your home directory.

3. Apply the grep command to filter for all orders with PENDING_PAYMENT status.

4. create a new file named orders_filtered.csv under your staging folder with the filtered results.

5. create a folder hierarchy in your hdfs home named data/landing

6. copy this orders_filtered.csv file from your staging folder in local to data/landing folder in your hdfs.

7. Run a command to check number of records in orders_filtered.csv file under data/landing folder

8. Write a command to list the files in the data/landing folder of hdfs.

9. reframe this command so that you can see the file size in kb's

10. change the permission of this file
give read,write and execute to the owner
read and write to the group
read to others

11. create a new folder data/staging in your hdfs and move orders_filtered.csv from data/landing to data/staging

12. Now let's assume a spark program would have run on your staging folder to do some processing and let's say the processed results gives you just 2 lines as ouput
3617,2013-08-15 00:00:00.0,8889,PENDING_PAYMENT
68714,2013-09-06 00:00:00.0,8889,PENDING_PAYMENT

To simulate this, create a new file called orders_result.csv in the home directory of your local gateway node using vi editor and have the above 2 records..

13. move orders_result.csv from local to hdfs under a new directory called data/results (thing as if spark program has run and has created this file)

14. Now the processed results we want to bring back to local under a folder data/results in your local. so run a command to bring the file from hdfs to local.

15. rename the file orders_result.csv under data/results folder in your local to final_results.csv

16. Now we are done.. so delete all the directories that you have created in your local as well as hdfs.