Assignment

Question 1

Open a Spark Notebook and create a spark base dataframe by reading all the files under /public/sms/users folder. Do check how many partitions are created in your dataframe.

Question 2

Do some basic analysis and find out the following

a. total number of records in the dataframe

b. how many users are from the state New York

c. which state has maximum number of postal codes

d. which city has the most number of users

e. how many users have email domain as bizjournals.com

f. how many users have 4 phone numbers mentioned

g. how many users do not have any phone number mentioned

Question 3

Write the data from the base dataframe as it is to the disk, but write in parquet format. Observe the number of files created, also the size of files.

Question 4

Package the code in question 1 & 2 and create a py file.

Invoke this py file using the spark submit command by grabbing 2 executors each of size 4 gb and 2 cpu cores.

Make sure you run it in client mode, so that you can see the result of your print statements.

Question 5

Create a pivot where states should be the rows and user_gender should be the columns.

Something like below, the aggregation is based on the number of records under each category where the phone number is not null.

	Male	Female
california	x	x
arizona	x	x
Wisconsin	x	x

Question 6

Package the code in question 5, and create a py file. Invoke this py file using the spark submit command by grabbing 4 executors each of size 2 gb and 1 cpu cores.

Also driver of size 2 gb and 1 cpu core. use the verbose option to see what configurations came into effect.

Make sure you run it in cluster mode, and save the results in your hdfs home directory in a folder named pivot_assignment_result

Question 7

- Read all the files under the folder /public/airlines_all/airlines and create a dataframe. (No need to infer the schema, nor you have to define it.)
- Try seeing how many initial partitions are there in your dataframe.
- Why do you see these many partitions, what is the logic?
- Now change the maxPartitionBytes to 140 mb
- Try creating the same dataframe again by loading all the files. (Again No need to infer the schema, nor you have to define it.)
- Try seeing how many initial partitions are there in this new dataframe.
- You will see a different number of partitions now, why?

Question 8

Clean up the resources and directories you created.