Date

Load aws open data

1. Goal

In this assignment, we will try to load aws open data set to snowflake data base.

Aws open data can be found in below location,
https://registry.opendata.aws/nyc-tlc-trip-records-pds/

Browse to this location, and try to understand the data.

2. Preparation

For this exercise, you have to load , green taxi trip data. green_tripdata*

In your aws console list all the files in the location, s3://nyc-tlc/

Write aws s3 list command, to list only green taxi files.

Because you don't have any idea about the file which you are trying to load, it's always good practice to copy a single file from s3 to local and check the data and number of columns.

Write command to copy file , green_tripdata_2013-08.csv to your local system.

Open the file in excel and observe the data.

How many columns you have ?

How many date columns you can see?

3. BEFORE YOU LOAD.

Now you should be having better understanding of data you are going to load.

So go ahead and create table in **snowflake database** to capture this data in table.

```
CREATE OR REPLACE TRANSIENT TABLE GREEN_TRIP_DATA
(
VendorID    NUMBER          NOT NULL,
lpep_pickup_datetime        TIMESTAMP_NTZ       ,
lpep_dropoff_datetime       TIMESTAMP_NTZ       ,
store_and_fwd_flag VARCHAR(1)   ,
RatecodeID          NUMBER          ,
PULocationID        NUMBER          ,
DOLocationID        NUMBER          ,
passenger_count     NUMBER          ,
trip_distance       FLOAT  ,
fare_amount         FLOAT  ,
extra       FLOAT  ,
mta_tax     FLOAT  ,
tip_amount FLOAT  ,
tolls_amount        FLOAT  ,
ehail_fee   FLOAT  ,
improvement_surcharge       FLOAT  ,
total_amount        FLOAT  NOT NULL,
payment_type        NUMBER          ,
trip_type   NUMBER,
FILE_NAME VARCHAR
)
```

Creating table is first step. But you have to create few more objects to facilitate loading data to snowflake.

1. *It’s always good practice to create file format object, to parse the file you are loading. In this current scenario, you are trying to load csv file.*

   Go ahead and create CSV file format by name, ***aws_csv_format***

   Create file format object. Mention command below,

Hint : You should be aware that, data has ***header column*** *. While creating file format you should be skipping the header column.*

2. *Create stage object pointing to,* ***aws*** *open data s3 location*

   .

   Create stage object by name , *aws_s3_open_data*. Mention command below,

Hint : you don’t need to create integration object as you are trying to copy ***open data****. You can use URL,* ***'s3://nyc-tlc/trip\ data/'***

Remember, you should be attaching file format object to stage object.

4. Loading data.

I hope, by now, you should be having your TABLE, FILE FORMAT OBJECT AND STAGE OBJECT ready.

ANALYSIS

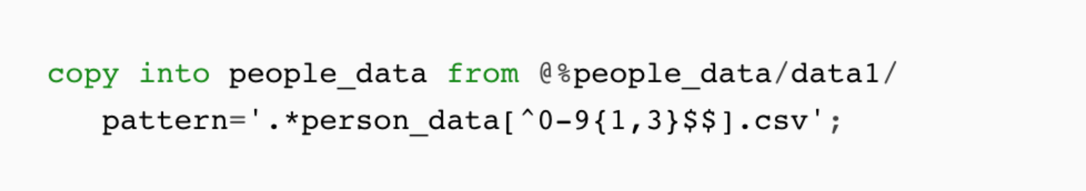
Before we load the data, let's try to do some analysis.

You can create a view on top of s3 data and query or you can query it directly. Your choice!!!!

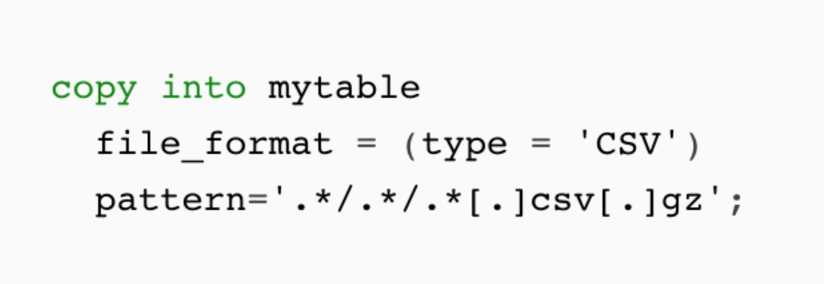
1. *How many distinct green taxi trip data files are there ?*
   Write the command below,

Hint : Use, ***metadata$filename*** build in column name. You can also use, pattern property. Refer link below,

https://docs.snowflake.com/en/user-guide/data-load-considerations-load.html



```
copy into people_data from @%people_data/data1/
   pattern='.*person_data[^0-9{1,3}$$].csv';
```



```
copy into mytable
  file_format = (type = 'CSV')
  pattern='.*/.*/.*[.]csv[.]gz';
```

How many distinct green taxi files are there ?

2. *Check how many records each**, green_trip*** file has. Which file has less number of records?*

    Group by file name and count number of records in each file.

    Write your query below,

3. *Check total number of records you are going to load.*

    *Write your query below for **green_trip****

How many records ?

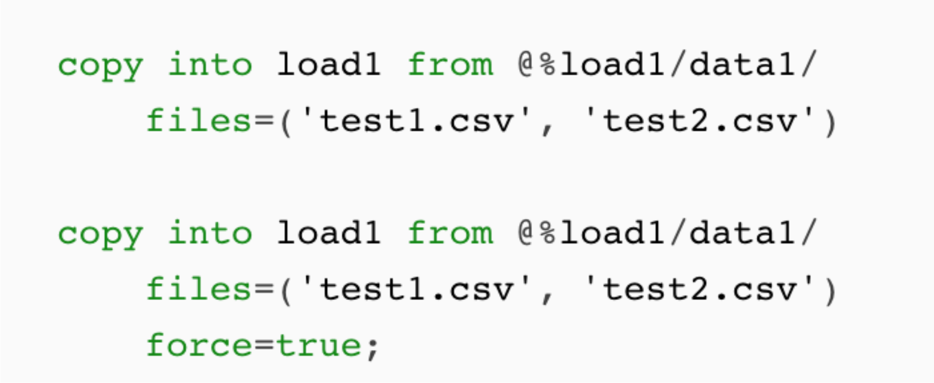
START LOAD

It’s not a good practice to load all files at once to snowflake database.

You should first try loading single file to the table.

Refer this link below,

https://docs.snowflake.com/en/sql-reference/sql/copy-into-table.html



```
copy into load1 from @%load1/data1/
    files=('test1.csv', 'test2.csv')

copy into load1 from @%load1/data1/
    files=('test1.csv', 'test2.csv')
    force=true;
```

Try to copy file , ***green_tripdata_2015-05.csv*** to the table.

Write your copy command below,

Hint : Use ON_ERROR =’CONTINUE’ to reject bad records.

Try to check if there is any rejected records.

Write command to capture rejected records,

If you are sure that, there is not many rejects, then copy all Green_trip* files to table.

Write your copy command below,

Hint : You should only copy Green_trip* files to table. Use pattern option while copying.

Pattern should be something like, '.*green_tripdata.*csv'