

นายณณวุฒิ จัตรชินวุฒิ [Mr.Yannawut Chatchinwut] 6409035174

hw2_5174.py

Description : ไฟล์ script แบ่งเป็น 2 ส่วนครับคือ setup section และ query section โดย

setup section จะเป็นการสร้าง s3 bucket และนำเข้าไฟล์ข้อมูลที่ต้องการจาก nyc-tlc มาไว้ที่ bucket ของเราซึ่ง

create_bucket << ใช้สร้าง bucket

```
#Section I : Setup
#Importing required Python dependencies.
import boto3
import botocore
import pandas as pd
from IPython.display import display, Markdown

#Setting up variables for S3 client and S3 resource
s3 = boto3.client('s3')
s3_resource = boto3.resource('s3')

#Define create_bucket function
def create_bucket(bucket):
    import logging

    try:
        s3.create_bucket(Bucket=bucket)
    except botocore.exceptions.ClientError as e:
        logging.error(e)
        return 'Bucket ' + bucket + ' could not be created.'
    return 'Created or already exists ' + bucket + ' bucket.'

#Call create_bucket function
create_bucket('nyctlc-cs653-5174')

#Importing data
def copy_among_buckets(from_bucket, from_key, to_bucket, to_key):
    s3_resource.meta.client.copy({'Bucket': from_bucket, 'Key': from_key},
                                to_bucket, to_key)

#Call copy_among_buckets function for cp yellow taxi rides Jan - Mar 2017
copy_among_buckets(from_bucket='nyc-tlc', from_key='trip data/yellow_tripdata_2017-01.parquet',
                    to_bucket='nyctlc-cs653-5174', to_key='yellow_tripdata_2017-01.parquet')
copy_among_buckets(from_bucket='nyc-tlc', from_key='trip data/yellow_tripdata_2017-02.parquet',
                    to_bucket='nyctlc-cs653-5174', to_key='yellow_tripdata_2017-02.parquet')
copy_among_buckets(from_bucket='nyc-tlc', from_key='trip data/yellow_tripdata_2017-03.parquet',
                    to_bucket='nyctlc-cs653-5174', to_key='yellow_tripdata_2017-03.parquet')
```

Amazon S3 > Buckets > nyctlc-cs653-5174

nyctlc-cs653-5174 [Info](#)

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	yellow_tripdata_2017-01.parquet	parquet	March 24, 2023, 22:38:28 (UTC+07:00)	128.6 MB	Standard
<input type="checkbox"/>	yellow_tripdata_2017-02.parquet	parquet	March 24, 2023, 22:38:29 (UTC+07:00)	121.7 MB	Standard
<input type="checkbox"/>	yellow_tripdata_2017-03.parquet	parquet	March 24, 2023, 22:38:30 (UTC+07:00)	137.9 MB	Standard

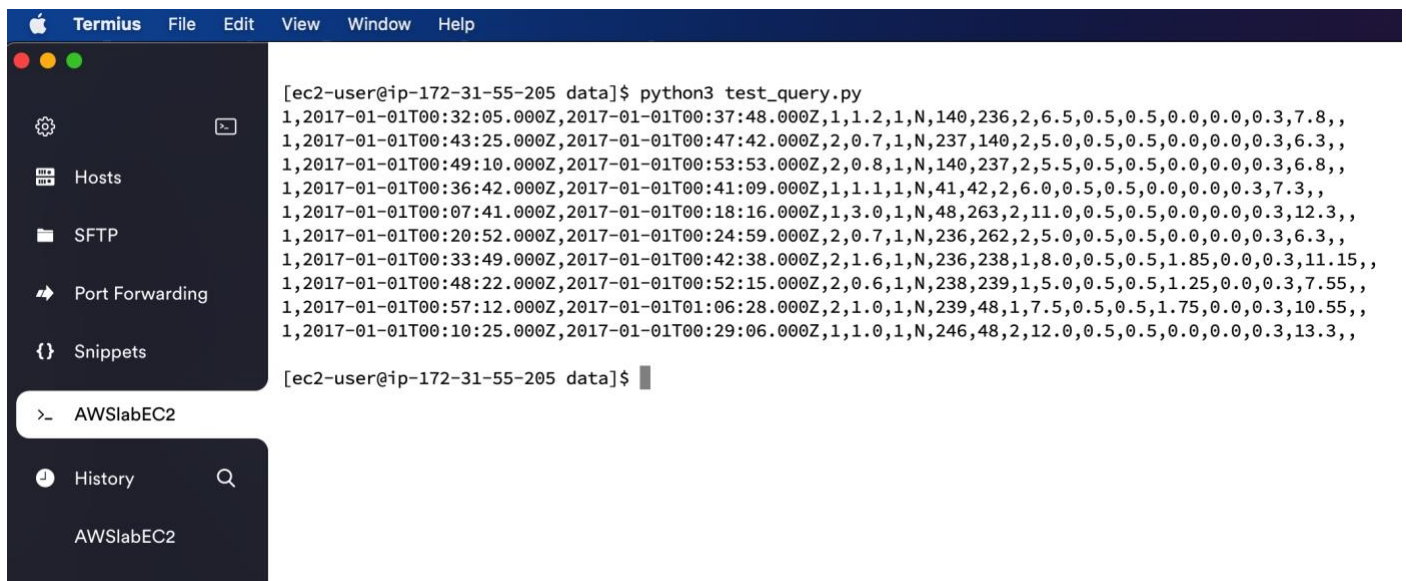
query section จะเป็นการ query ข้อมูลจาก bucket

```
#Section II : Query
# Set up S3 Select parameters
query ="select * from s3object s limit 10"
bucket = 'nyctlc-cs653-5174'
key = 'yellow_tripdata_2017-01.parquet'
expression_type = 'SQL'
input_serialization = {'Parquet': {}}
output_serialization = {'CSV': {}}

# Execute S3 Select query
response = s3.select_object_content(
    Bucket=bucket,
    Key=key,
    Expression=query,
    ExpressionType=expression_type,
    InputSerialization=input_serialization,
    OutputSerialization=output_serialization,
)

# Iterate through the response and print each line
for event in response['Payload']:
    if 'Records' in event:
        records = event['Records']['Payload'].decode('utf-8')
        print(records)
```

ทดลอง query



```
[ec2-user@ip-172-31-55-205 data]$ python3 test_query.py
1,2017-01-01T00:32:05.000Z,2017-01-01T00:37:48.000Z,1,1.2,1,N,140,236,2,6.5,0.5,0.5,0.0,0.0,0.3,7.8,,
1,2017-01-01T00:43:25.000Z,2017-01-01T00:47:42.000Z,2,0.7,1,N,237,140,2,5.0,0.5,0.5,0.0,0.0,0.3,6.3,,
1,2017-01-01T00:49:10.000Z,2017-01-01T00:53:53.000Z,2,0.8,1,N,140,237,2,5.5,0.5,0.5,0.0,0.0,0.3,6.8,,
1,2017-01-01T00:36:42.000Z,2017-01-01T00:41:09.000Z,1,1.1,1,N,41,42,2,6.0,0.5,0.5,0.0,0.0,0.3,7.3,,
1,2017-01-01T00:07:41.000Z,2017-01-01T00:18:16.000Z,1,3.0,1,N,48,263,2,11.0,0.5,0.5,0.0,0.0,0.3,12.3,,
1,2017-01-01T00:20:52.000Z,2017-01-01T00:24:59.000Z,2,0.7,1,N,236,262,2,5.0,0.5,0.5,0.0,0.0,0.3,6.3,,
1,2017-01-01T00:33:49.000Z,2017-01-01T00:42:38.000Z,2,1.6,1,N,236,238,1,8.0,0.5,0.5,1.85,0.0,0.3,11.15,,
1,2017-01-01T00:48:22.000Z,2017-01-01T00:52:15.000Z,2,0.6,1,N,238,239,1,5.0,0.5,0.5,1.25,0.0,0.3,7.55,,
1,2017-01-01T00:57:12.000Z,2017-01-01T01:06:28.000Z,2,1.0,1,N,239,48,1,7.5,0.5,0.5,1.75,0.0,0.3,10.55,,
1,2017-01-01T00:10:25.000Z,2017-01-01T00:29:06.000Z,1,1.0,1,N,246,48,2,12.0,0.5,0.5,0.0,0.0,0.3,13.3,,
[ec2-user@ip-172-31-55-205 data]$
```

หมายเหตุ ไม่สามารถเขียน query statement ตอบข้อ a-c ได้ครับ เนื่องจากติด group by สะท้อนการเรียนรู้ จากการบ้านในครั้งนี้

- ได้ทดลองใช้ python script สร้าง s3 bucket แทนการใช้ gui
- ได้ทดลองเขียน function ต่าง ๆ
- ชอบที่ได้ลองเขียน code ครับ แต่ไม่ชอบตรงที่ต้อง query กับไฟล์ parquet ครับ และ lib ของ python บางตัวเช่น pyarrow ก็มีปัญหาคอนไร์กใช้บนเครื่อง EC2 ครับ
- อยากศึกษาเรื่องการจัดการไฟล์ parquet เพิ่ม หรือ tool อื่นที่ใช้ในการ manage parquet ไฟล์เพิ่มครับ
- https://github.com/James-Chatchinwut/cs653_homework/blob/main/hw2/hw2_5174.py