

6409035109 กษิตศ อินทน

```
In [ ]: import boto3
import boto3
import boto3
#import pandas as pd
#from IPython.display import display, Markdown

s3 = boto3.client('s3')
s3_resource = boto3.resource('s3')

def create_bucket(bucket):
    import logging

    try:
        s3.create_bucket(Bucket=bucket)
    except boto3.exceptions.ClientError as e:
        logging.error(e)
        return 'Bucket ' + bucket + ' could not be created.'
    return 'Created or already exists ' + bucket + ' bucket.'

#create_bucket('nyctlc-cs653-5109')
#ถ้าไม่มี bucket ต้องใช้คำสั่งนี้
```

1. ในส่วนแรก จะเป็นการสร้าง bucket โดยการใช้ script โดยใช้ Boto 3 เพื่อสร้าง bucket โดยใน Set คำสั่งส่วนนี้จะทำการตรวจสอบหากไม่สามารถสร้าง bucket ได้หรือมี bucket ชื่อดังกล่าวอยู่แล้ว

```
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)
    print(f'File {to_key} saved to S3 bucket {to_bucket}')

copy_among_buckets(from_bucket='nyc-tlc', from_key='trip data/yellow_tripdata_2017-01.parquet',
                    to_bucket='nyctlc-cs653-5109', 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-cs653-5109', 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-cs653-5109', to_key='yellow_tripdata_2017-03.parquet')
```










2. จะเป็นการนำเข้าข้อมูลจาก open Data โดยในโจทยนี้จะทำการเลือกข้อมูลปี 2017 เดือน 1-3 รถสีเหลือง เมื่อนำเข้าสำเร็จจะขึ้นใน bucket ดังนี้


nyctlc-cs653-5109 [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](#)

  Copy S3 URI  Copy URL  Download  Open  Delete  Actions  Create folder  Upload

< 1 > 

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

6409035109 กษิตศ อินทนู

หลังจาก นำเข้าข้อมูลแล้วจะทำการ Query แสดงข้อมูลเบื้องต้น ดังนี้

```
# Set up S3 Select parameters
query = "select * from s3object s limit 10"
bucket = 'nyctlc-cs653-5109'
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)
```

หากไม่มีการผิดพลาดผลลัพธ์จะออกมาได้ดังนี้

```
[ec2-user@ip-172-31-5-201 ~]$ python3 hw2_5109.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,,
```

ข้อ 3.

เกิดปัญหาไม่สามารถ Query ได้เนื่องจาก Parquet ไม่รองรับ SQL ทำให้ไม่สามารถ Query ได้ ขออภัยด้วยครับ

6409035109 กษิตศ อินทนู

ข้อ 4

- จากการบ้านครั้งนี้ ทำให้ได้ explore การใช้ SDK อย่าง Boto3 ที่เป็นภาษา Python ซึ่งปกติไม่เคยใช้ Tools ตัวนี้ และได้ทบทวนการเขียน SQL
- สิ่งที่ไม่ชอบในการทำการบ้านครั้งนี้คือการที่ตัวเองยังขาดความเข้าใจในไฟล์ข้อมูลต่างๆ ส่งผลให้ไม่สามารถทำ task ให้สำเร็จได้
- จำเป็นต้องศึกษาข้อมูล Date Frame จาก lib ต่างๆเพิ่มเติมเพื่อที่จะนำมาใช้ในโครงการต่างๆในอนาคตให้สำเร็จได้