Lab 1. AWS S3 Bucket 생성 및 데이터 저장

1. S3 Bucket 생성

2. AWS CLI를 사용해 Bucket List 출력하기

1) Windows Command 창에서 AWS Access Key ID & AWS Secret Access Key 입력

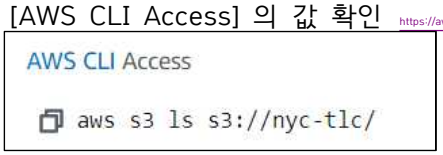
```
C:₩WINDOWS₩system32>aws configure
AWS Access Key ID [****************[redacted]]:
AWS Secret Access Key [****************[redacted]]:
Default region name [ap-northeast-2a]: ap-northeast-2
Default output format [JSON]: json
```

2) S3 Bucket List 출력

```
C:₩WINDOWS₩system32>aws s3 ls /
2023-03-13 15:44:44 mk-datalake-bucket
```

3. Lab에서 사용할 Public DataSet 확인

1) [Open Data on AWS]에서 [New York City Taxi and Limousine Commission (TLC) Trip Record Data]의 [AWS CLI Access] 의 값 확인 https://aws.amazon.com/marketplace/pp/prodview-okyonroqg5b2u?sr=0-1&ref_=beagle&applicationId=AWSMPContessa#usage



2) Windows Command에서

```
C:₩WINDOWS₩system32>aws s3 ls s3://nyc-tlc/
                           PRE csv_backup/
                           PRE misc/
                           PRE trip data/
```

3) Object들 중에서 "trip data" 검색

```
C:₩WINDOWS₩system32>aws s3 ls s3://nyc-tlc/"trip data"/
```

4) 검색 결과 중 "2022-11" 필터하기

```
C:₩WINDOWS₩system32>aws s3 ls s3://nyc-tlc/"trip data"/ | findstr "2022-11"
2022-11-14 22:49:35   11851834 fhv_tripdata_2022-07.parquet
2022-11-14 22:49:40   11826775 fhv_tripdata_2022-08.parquet
2023-01-26 01:43:42   11298968 fhv_tripdata_2022-11.parquet
2022-11-14 22:49:28  443730405 fhvhv_tripdata_2022-07.parquet
2022-11-14 22:49:28  436536084 fhvhv_tripdata_2022-08.parquet
2023-01-26 01:43:40  464298215 fhvhv_tripdata_2022-11.parquet
2022-11-14 22:49:31    1312353 green_tripdata_2022-07.parquet
2022-11-14 22:49:31    1346660 green_tripdata_2022-08.parquet
2023-01-26 01:43:41    1270324 green_tripdata_2022-11.parquet
2022-11-14 22:49:29   49367712 yellow_tripdata_2022-07.parquet
2022-11-14 22:49:29   49717159 yellow_tripdata_2022-08.parquet
2023-01-26 01:43:40   50106631 yellow_tripdata_2022-11.parquet
```

4. "trip-data"의 데이터를 위에서 생성한 나의 Bucket으로 복사하기

1) "trip-data"의 green_tripdata_2022-11.parquet을 위에서 생성한 나의 Bucket으로 복사하기

```
C:₩WINDOWS₩system32>aws s3 cp s3://nyc-tlc/"trip data"/green_tripdata_2022-11.parquet
s3://mk-datalake-bucket/input/green_tripdata_2022-11.parquet
copy: s3://nyc-tlc/trip data/green_tripdata_2022-11.parquet to s3://mk-datalake-bucket
/input/green_tripdata_2022-11.parquet
```

2) "trip-data"의 yellow_tripdata_2022-11.parquet을 위에서 생성한 나의 Bucket으로 복사하기

```
C:₩WINDOWS₩system32>aws s3 cp s3://nyc-tlc/"trip data"/yellow_tripdata_2022-11.parquet
 s3://mk-datalake-bucket/input/yellow_tripdata_2022-11.parquet
copy: s3://nyc-tlc/trip data/yellow_tripdata_2022-11.parquet to s3://mk-datalake-bucke
t/input/yellow_tripdata_2022-11.parquet
```

5. Local Machine에 CSV 파일 다운로드 하여 Head 확인하기

1) [New York City Taxi and Limousine Commission (TLC) Trip Record Data]의 CSV 파일 목록 확인

```
C:₩Users₩minky>aws s3 ls s3://nyc-tlc/csv_backup/
```

2) 특정 CSV 파일 다운로드

```
C:₩Users₩minky>aws s3 cp s3://nyc-tlc/csv_backup/yellow_tripdata_2022-02.csv .
```

3) CSV 파일 앞부분 확인

```
C:₩Users₩minky>more yellow_tripdata_2022-02.csv
```

```
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocati
onID,DOLocationID,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,cong
estion_surcharge
1,2022-02-01 00:06:58,2022-02-01 00:19:24,1.0,5.4,1.0,N,138,252,1,17.0,1.75,0.5,3.9,0.0,0.3,23.45,0.0
1,2022-02-01 00:38:22,2022-02-01 00:55:55,1.0,6.4,1.0,N,138,41,2,21.0,1.75,0.5,0.0,6.55,0.3,30.1,0.0
1,2022-02-01 00:03:20,2022-02-01 00:26:59,1.0,12.5,1.0,N,138,200,2,35.5,1.75,0.5,0.0,6.55,0.3,44.6,0.0
2,2022-02-01 00:08:00,2022-02-01 00:28:05,1.0,9.88,1.0,N,239,200,2,28.0,0.5,0.5,0.0,3.0,0.3,34.8,2.5
```