Lab. AWS S3 Bucket 생성 및 데이터 저장

1. S3 Bucket 생성
   1)[서비스] > [스토리지] > S3
   2)[버킷 만들기] 버튼 클릭
   3)[버킷 만들기] 페이지에서, [버킷 이름] : {계정}-datalake-bucket
   4)[AWS 리전] : 아시아 태평양(서울) ap-northeast-2
   5)[버킷 만들기] 버튼 클릭

2. AWS CLI를 사용하여 Bucket List 출력하기
   1)Windows Command 창 또는 macOS Terminal에서
   2)AWS Access Key ID와 AWS Secret Access Key 입력
      $ aws configure
      AWS Access Key ID [****************GYVQ]:
      AWS Secret Access Key [****************jWIz]:
      Default region name [ap-northeast-2]:
      Default output format [json]:

   3)S3 Bucket List 출력
      $ aws s3 ls /
      2023-03-08 09:56:42 henry-datalake-bucket

   4)해당 Bucket 내용 출력
      $ aws s3 ls s3://{bucket name}
                  <--- Bucket 내에 어떤 Object도 없기 때문에 아무 것도 출력되지 않음.

3. Lab에서 사용할 Public DataSet 확인
   1)Google에서 "aws public datasets"로 검색
   2)검색 결과에서 [Open Data on AWS] 링크 클릭,
   https://aws.amazon.com/ko/opendata/?wwps-cards.sort-by=item.additionalFields.sortDate&wwps-cards.sort-order=desc
   3)페이지에서 [Find publicly available data on AWS] 버튼 클릭
   4)검색창에 "taxi" 입력하여 "New York City Taxi and Limousine Commission(TLC) Trip Record Data" 클릭
      -https://aws.amazon.com/marketplace/pp/prodview-okyonroqg5b2u?sr=0-1&ref_=beagle&applicationId=AWSMPContessa
   5)[New York City Taxi and Limousine Commission (TLC) Trip Record Data] 페이지에서, [Description] 탭에서 [Documentation]의 링크 클릭
   6)[TLC Trip Record Data] 페이지에서, [Data Dictionaries and MetaData] 섹션의 "Yellow Trips Data Dictionary" 클릭하여 문서의 내용 파악
   7)다시 [New York City Taxi and Limousine Commission (TLC) Trip Record Data] 페이지로 돌아와서, [Resources on AWS] 탭으로 이동
   8)[AWS CLI Access]의 값 확인
      aws s3 ls s3://nyc-tlc/
   9)Windows Command 창 또는 macOS의 Terminal에서,
      $ aws s3 ls s3://nyc-tlc/
            PRE csv_backup/
            PRE misc/
            PRE trip data/
   10)Object들 중에서 "trip data" 검색
      $ aws s3 ls s3://nyc-tlc/"trip data"/

   11)검색 결과 중 "2022-10" 필터하기
      -macOS
         $ aws s3 ls s3://nyc-tlc/"trip data"/ | grep 2022-10
      -Windows
         >aws s3 ls s3://nyc-tlc/"trip data"/ | findstr "2022-10"

      2022-12-20 06:42:16   12051434 fhv_tripdata_2022-10.parquet
      2022-12-20 06:42:12 495083481 fhvhv_tripdata_2022-10.parquet
      2022-12-20 06:42:14   1444642 green_tripdata_2022-10.parquet
      2022-12-20 06:42:12   57061938 yellow_tripdata_2022-10.parquet

4. "trip-data"의 데이터를 위에서 생성한 나의 Bucket으로 복사하기
   1)"trip-data"의 green_tripdata_2022-10.parquet를 위에서 생성한 나의 Bucket으로 복사하기
      $ aws s3 cp s3://nyc-tlc/"trip data"/green_tripdata_2022-10.parquet
      s3://henry-datalake-bucket/input/green_tripdata_2022-10.parquet
      copy: s3://nyc-tlc/trip data/green_tripdata_2022-10.parquet to
      s3://henry-datalake-bucket/input/green_tripdata_2022-10.parquet

   2)"trip-data"의 yellow_tripdata_2022-10.parquet를 위에서 생성한 나의 Bucket으로 복사하기
      $ aws s3 cp s3://nyc-tlc/"trip data"/yellow_tripdata_2022-10.parquet
      s3://henry-datalake-bucket/input/yellow_tripdata_2022-10.parquet
copy: s3://nyc-tlc/trip data/yellow_tripdata_2022-10.parquet to
s3://henry-datalake-bucket/input/yellow_tripdata_2022-10.parquet

   3)해당 파일들 복사되었는지 확인하기
      $ aws s3 ls s3://henry-datalake-bucket/input/
      2023-03-08 10:29:22   1444642 green_tripdata_2022-10.parquet
      2023-03-08 10:31:50   57061938 yellow_tripdata_2022-10.parquet

5. Local Machine에 CSV 파일 다운로드하여 Head 확인하기
   1)[New York City Taxi and Limousine Commission (TLC) Trip Record Data]의 CSV 파일 목록 확인
      $ aws s3 ls s3://nyc-tlc/csv_backup/

2)특정 CSV 파일 다운로드

$ aws s3 cp s3://nyc-tlc/csv_backup/yellow_tripdata_2020-04.csv .

3)CSV 파일 앞 부분 확인

-Windows

>more yellow_tripdata_2020-04.csv

-macOS

$ head yellow_tripdata_2020-04.csv

VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_flag,PULocationID,DOLocationID,payment_type,fare_amount,extra,mta_tax,tip_amount,tolls_amount,improvement_surcharge,total_amount,congestion_surcharge

1,2020-04-01 00:41:22,2020-04-01 01:01:53,1,1.20,1,N,41,24,2,5.5,0.5,0.5,0,0,0.3,6.8,0

1,2020-04-01 00:56:00,2020-04-01 01:09:25,1,3.40,1,N,95,197,1,12.5,0.5,0.5,2.75,0,0.3,16.55,0

1,2020-04-01 00:00:26,2020-04-01 00:09:25,1,2.80,1,N,237,137,1,10,3,0.5,1,0,0.3,14.8,2.5

1,2020-04-01 00:24:38,2020-04-01 00:34:38,0,2.60,1,N,68,142,1,10,3,0.5,1,0,0.3,14.8,2.5

2,2020-04-01 00:13:24,2020-04-01 00:18:26,1,1.44,1,Y,263,74,1,6.5,0.5,0.5,3,0,0.3,13.3,2.5

2,2020-04-01 00:24:36,2020-04-01 00:33:09,1,2.93,1,N,75,170,2,10.5,0.5,0.5,0,0,0.3,14.3,2.5