Lab1. AWS Glue의 Data Catalog 및 Crawler 사용하기
-New York Text Data 사용하는 지난 Lab에서 계속 진행

1. AWS Glue 방문
   1)[서비스] > [분석] > [AWS Glue]

   2)AWS Glue Database 생성
      -왼쪽 메뉴 [Data Catalog] > [Databases] 이동
      -[Add database] click
      -[Create a database] 페이지에서,
         --Name : {계정}-glue-db
         --[Create database] click

   3)[Data Catalog] > [Crawlers] 이동
      -[Create crawler] click
      -[Set crawler properties] 페이지에서,
         --Name : {계정}-crawler
         --[Next] click
      -[Choose data sources and classifiers] 페이지에서
         --[Is your data already mapped to Glue tables?] : Not yet
         --[Data sources] > [Add a data source] click
         --[Add data source] 창에서
            ---[Data source] : S3
            ---[Location of S3 data] : In this account
            ---[S3 path] :
               ----[Browse S3] click
               ----[Choose S3 path] 팝업창에서, "{계정}-datalake-bucket > output > green_yellow_2022-10" 선택 > [Choose] click
            ---[Subsequent crawler runs] : Crawl all sub-folders 선택
            ---[Add an S3 data source] click
         --[Next] click
      -[Configure security settings] 페이지에서
         --[IAM role]
            ---[Create new IAM role] click
               ---[Create new IAM role] 팝업창에서
                  ----[Enter new IAM role] : AWSGlueServiceRole-{계정}-crawler-role
                  ----[Create] click
         --[Next] click
      -[Set output and scheduling] 페이지에서
         --[Target database] : 목록에서 위에서 생성한 {계정}-glue-db 선택
         --[Table name prefix] : {계정}-
         --[Crawler schedule] > [Frequency] : On demand
         --[Next] click
      -[Create crawler] click
      -[Crawlers] 페이지에서 방금 생성한 crawler의 [State]가 Ready임을 확인할 수 있다.

2. S3 데이터 Crawling하기
   1)위에서 생성한 "{계정}-crawler"의 [State]가 Ready임을 확인하고, 체크한다.
   2)메뉴 중 [Run] 클릭
   3)그럼, [State]가 Running중으로 바뀌고, 끝나면 Stopping -> Ready로 바뀌게 된다.
   4)모두 끝나면 [Last run]이 "Succeeded"으로 변경되고, [Table changes from last run]에 "1 created"로 변경된다.

3. Table Data 확인하기
   1)[Data Catalog] > [Databases] > [Tables] 방문
   2)Tables 목록에 "{계정}-" 즉 Table prefix로 시작하는 테이블을 확인할 수 있다.
   3)[Classification]은 csv 이다.
   4)Table data를 확인하기 위해 [View data]의 [Table data] 링크를 클릭한다.
   5)[View data] 팝업창에서, [Proceed]를 클릭한다.
   6)[Amazon Athena] 창이 열리면서, [테이블]에 생성된 테이블({계정}-green_yellow_2022-10)이 확인되고 +를 클릭하여 확장하면 각 칼럼의 이름이 보인다.
   7)테이블 이름 옆 ":" > [쿼리 실행] > [테이블 미리 보기]
   8)테이블 이름 옆 ":" > [쿼리 실행] > [테이블 DDL 생성]을 클릭하여 DDL 구문 확인
      -OUTPUTFORMAT을 보면 현재 hadoop의 hive를 사용하는 것을 확인할 수 있다.
   9)[Partitions] 탭을 확인하면 파티션이 없다는 것을 확인할 수 있다.

4. 파티션 생성하기
   1)현재 테이블에 생성된 Data는 Partion이 없다. 왜냐하면 단일 데이터이기 때문이다.
   2){계정}-datalake-bucket의 output 폴더의 데이터를 모두 지운다.
   3)green taxi와 yellow taxi의 데이터를 2022-01 ~ 2022-11까지 Bucket에 복사한다.
      (DataLake) :DataLake $ aws s3 cp s3://nyc-tlc/"trip data"/green_tripdata_2022-01.parquet
      s3://{계정}-datalake-bucket/input/green_tripdata_2022-01.parquet
      ...
      ...
      (DataLake) :DataLake $ aws s3 cp s3://nyc-tlc/"trip data"/yellow_tripdata_2022-11.parquet
      s3://{계정}-datalake-bucket/input/yellow_tripdata_2022-11.parquet

   4)pyspark-demo4.py 실행
      {계정}-datalake-bucket/output/ym=2022-01
      ...
      ...
      {계정}-datalake-bucket/output/ym=2022-11

5. AWS Glue 다시 Crawling
   1)[Data Catalog] > [Crawlers] > {계정}-crawler 선택 후 [Action] > [Edit crawler]
   2)Step 2: Choose data sources and classifiers 섹션에서 [Edit] click
      -[Data sources] 섹션에서 S3 Type 라이오버튼 선택 > [Edit] click
      -[Edit data source] 창에서
         --[S3 path] : s3://{계정}-datalake-bucket/output 으로 수정
         --[Update S3 data source] click
      -계속 [Next] click
      -[Step 5 Review and update]에서 [Update] click

   3)Crawler로 돌아가서 목록에서 {계정}-crawler를 선택 후 [Run] click
   4)[State]가 Running중으로 바뀌고, 끝나면 Stopping -> Ready로 바뀌게 된다.
   5)모두 끝나면 [Last run]이 "Succeeded"으로 변경되고, [Table changes from last run]에 "1 created"로 변경된다.

6. Table Data 확인하기
   1)[Data Catalog] > [Databases] > [Tables] 방문
   2)Tables 목록에 "{계정}-output" 테이블을 확인할 수 있다.
   3)[Classification]은 csv 이다.
   4)Table data를 확인하기 위해 [View data]의 "Table data] 링크를 클릭한다.
   5)[View data] 팝업창에서, [Proceed]를 클릭한다.
   6)[Amazon Athena] 창이 열리면, [데이터베이스]는 "{계정}-glue-db" 선택하고, [테이블]에 {계정}-output이 있음을 확인되고 +를 클릭하여 확장하면 각 칼럼의 이름이 보인다.
   7)"{계정}-output" 테이블 이름옆에 "파티션됨"이 있음을 확인한다.
   8)테이블 이름 옆 ":" > [쿼리 실행] > [테이블 미리 보기]
      -"ym" 칼럼 확인
   9)다시 AWS Glue로 돌아와서 생성된 테이블({계정}-output) 상세페이지로 이동
   10)[Partitions] 탭을 확인하면 "ym" 칼럼으로 파티션이 생성되어 있음을 확인할 수 있다.