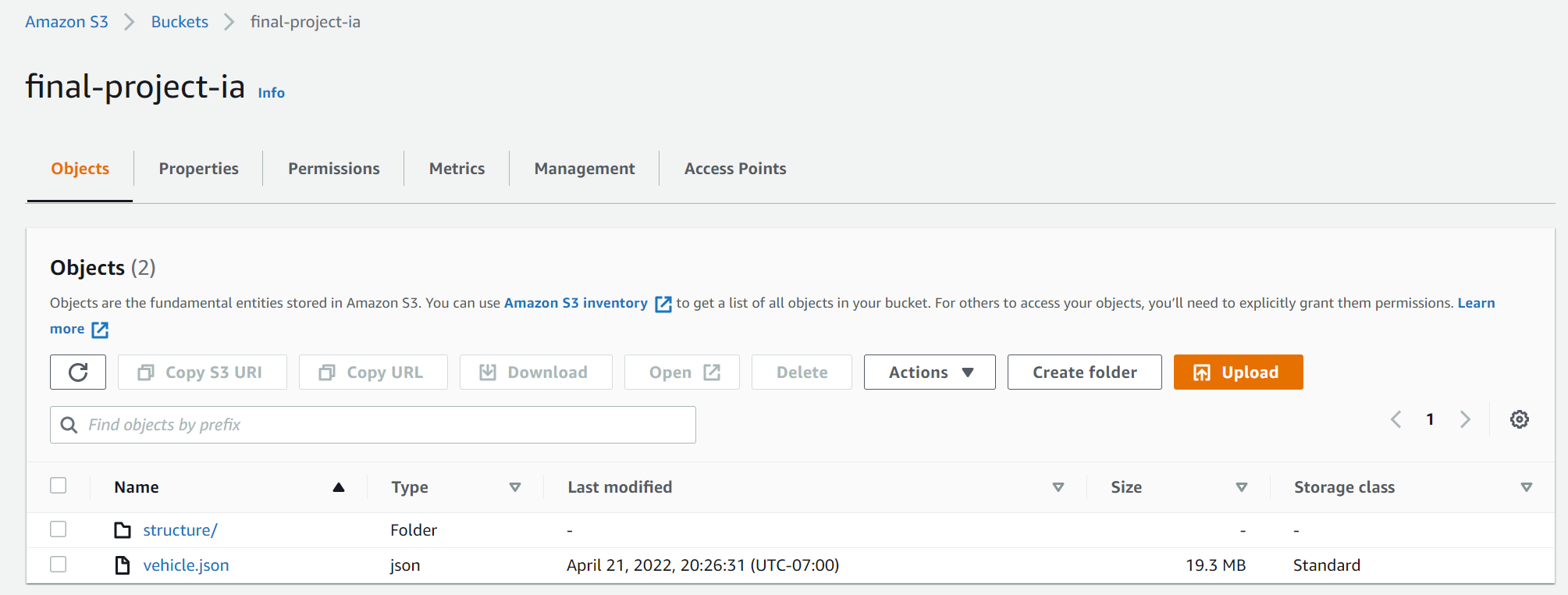
Download 3 structure data and 1 unstructured data from dmv government website

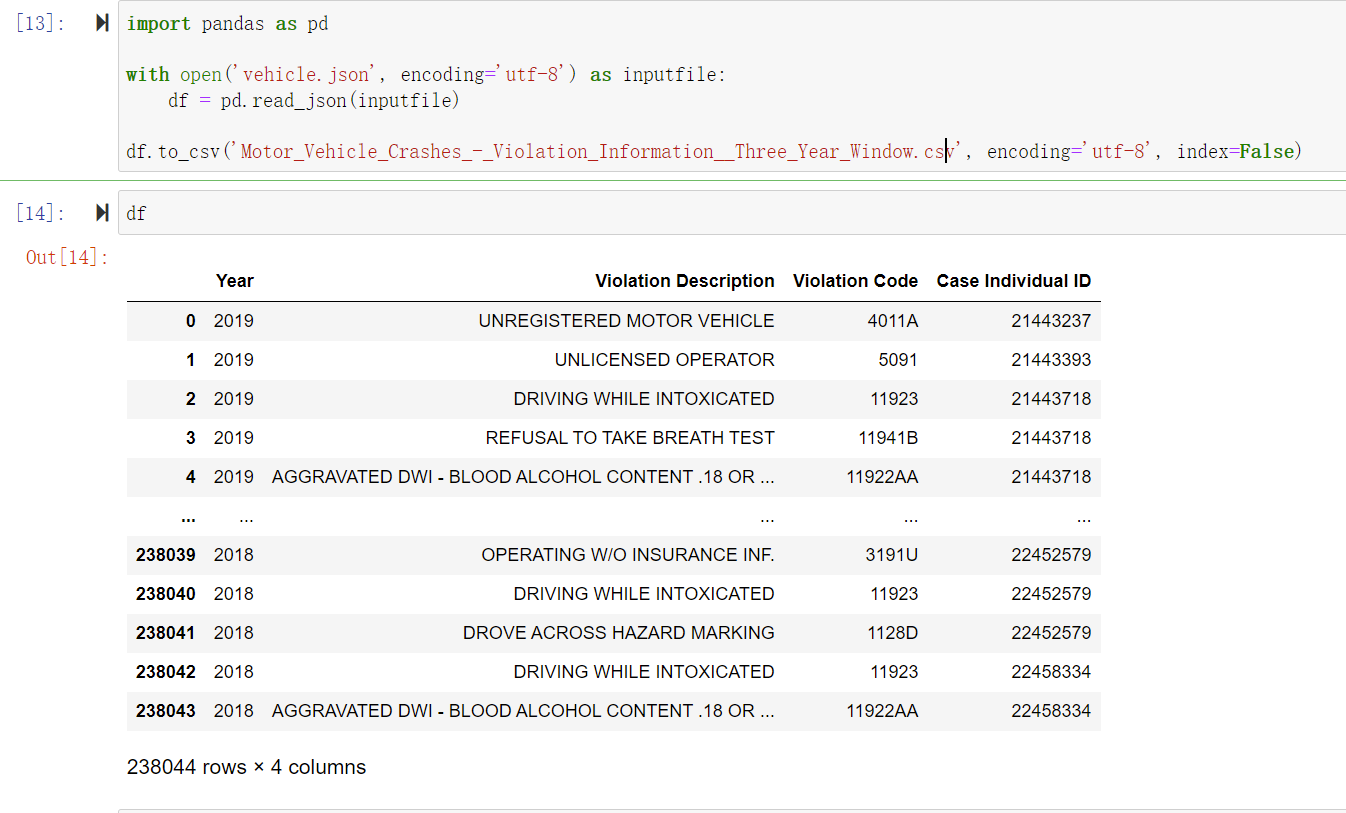
S3 Data:

Vehicle.json unstructure data

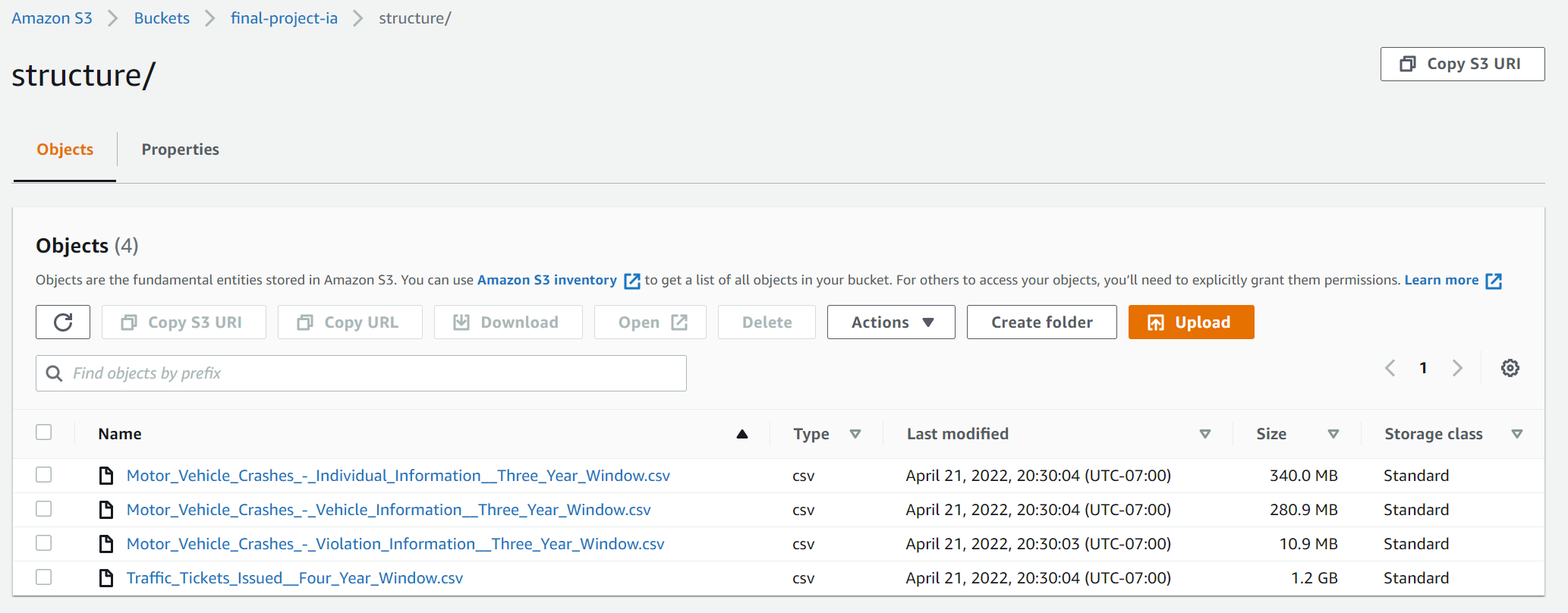




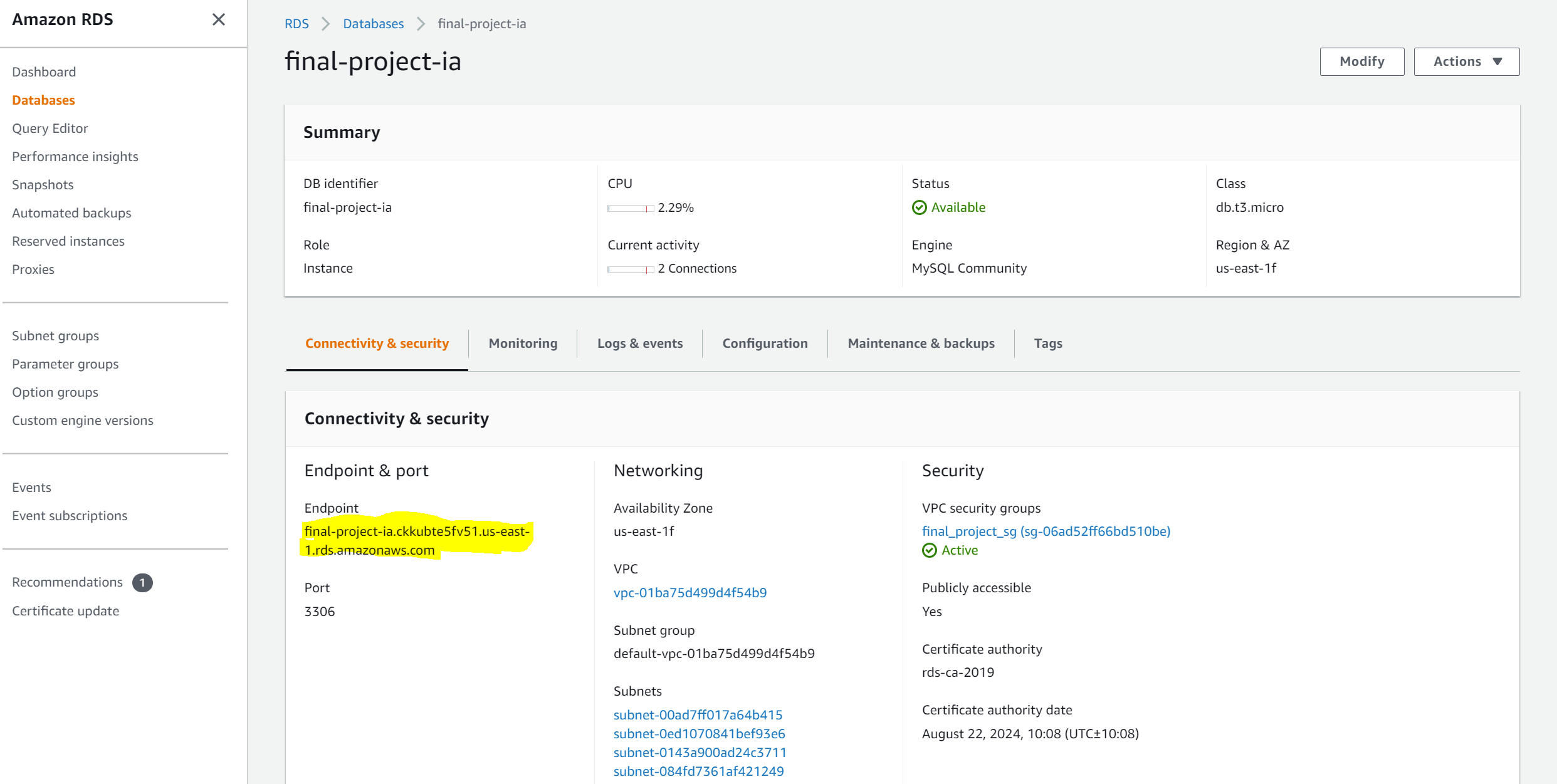
Unstructured data to structured data



Structure data



RDS:



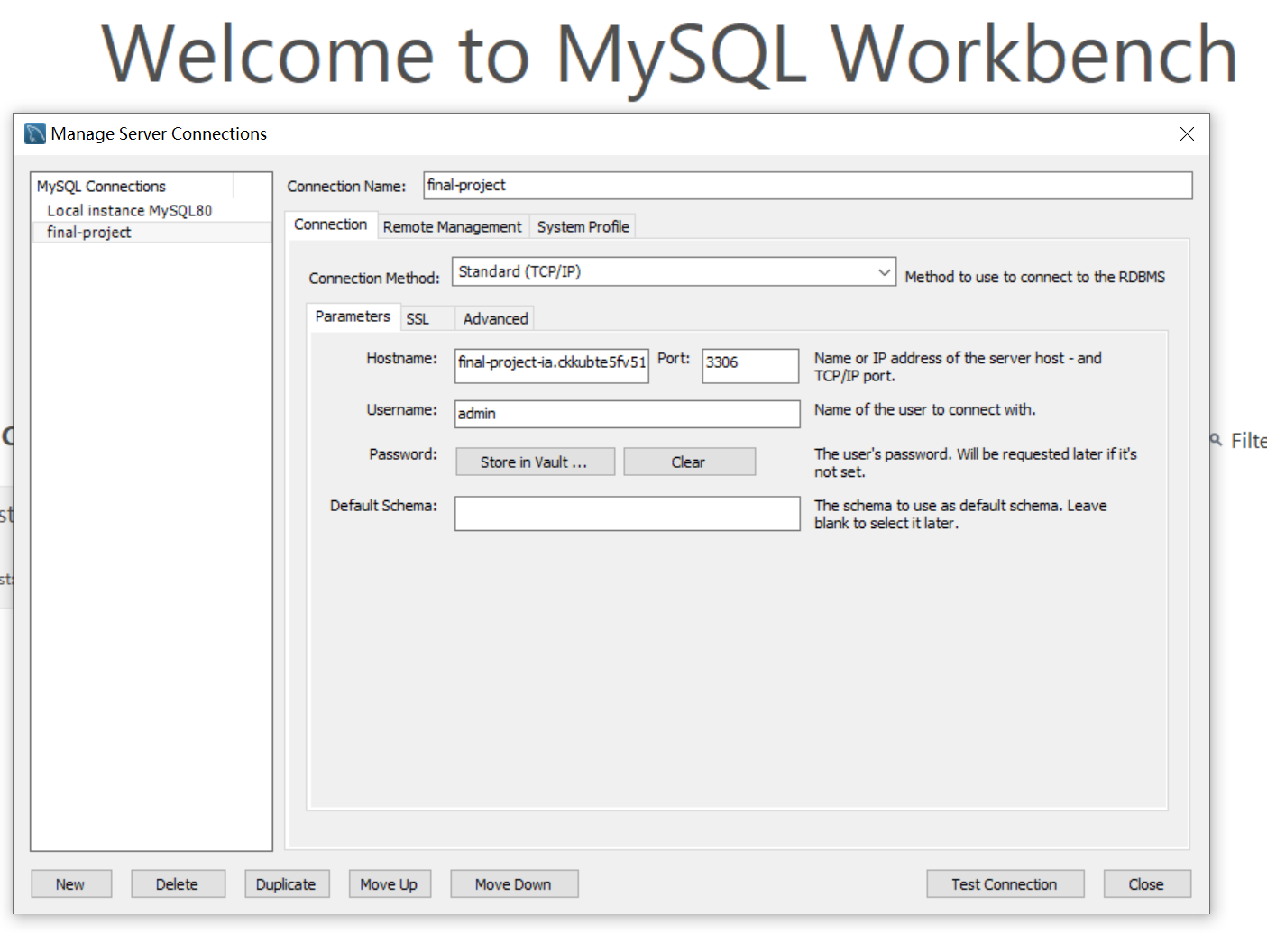
Endpoint: connect sql workbench

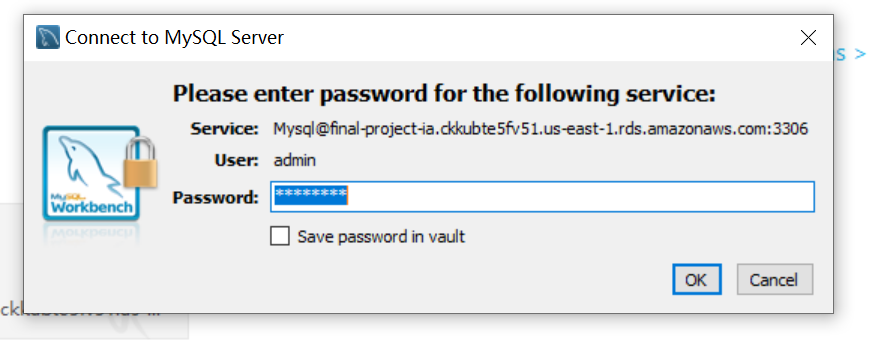
Endpoint: final-project-ia.ckkubte5fv51.us-east-1.rds.amazonaws.com

Username: admin

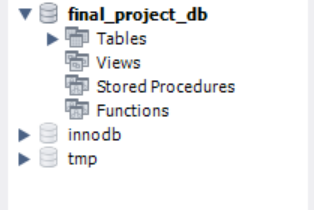
Password: 12345678

SQL Work Bench:





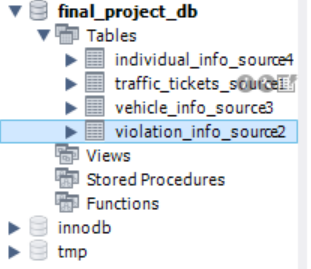
The dataset is named final\_project\_db which is empty



I created a ddl to create 4 empty tables as below



After running the sql ddl, there’re 4 empty tables under the dataset schema.

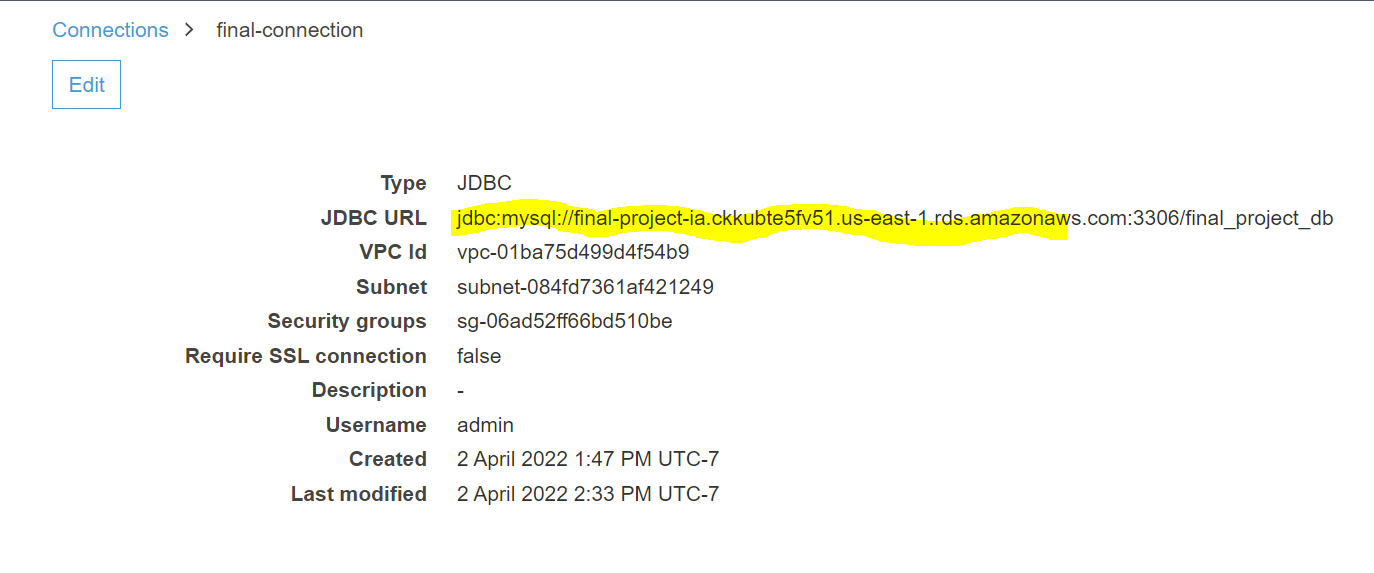


Glue:

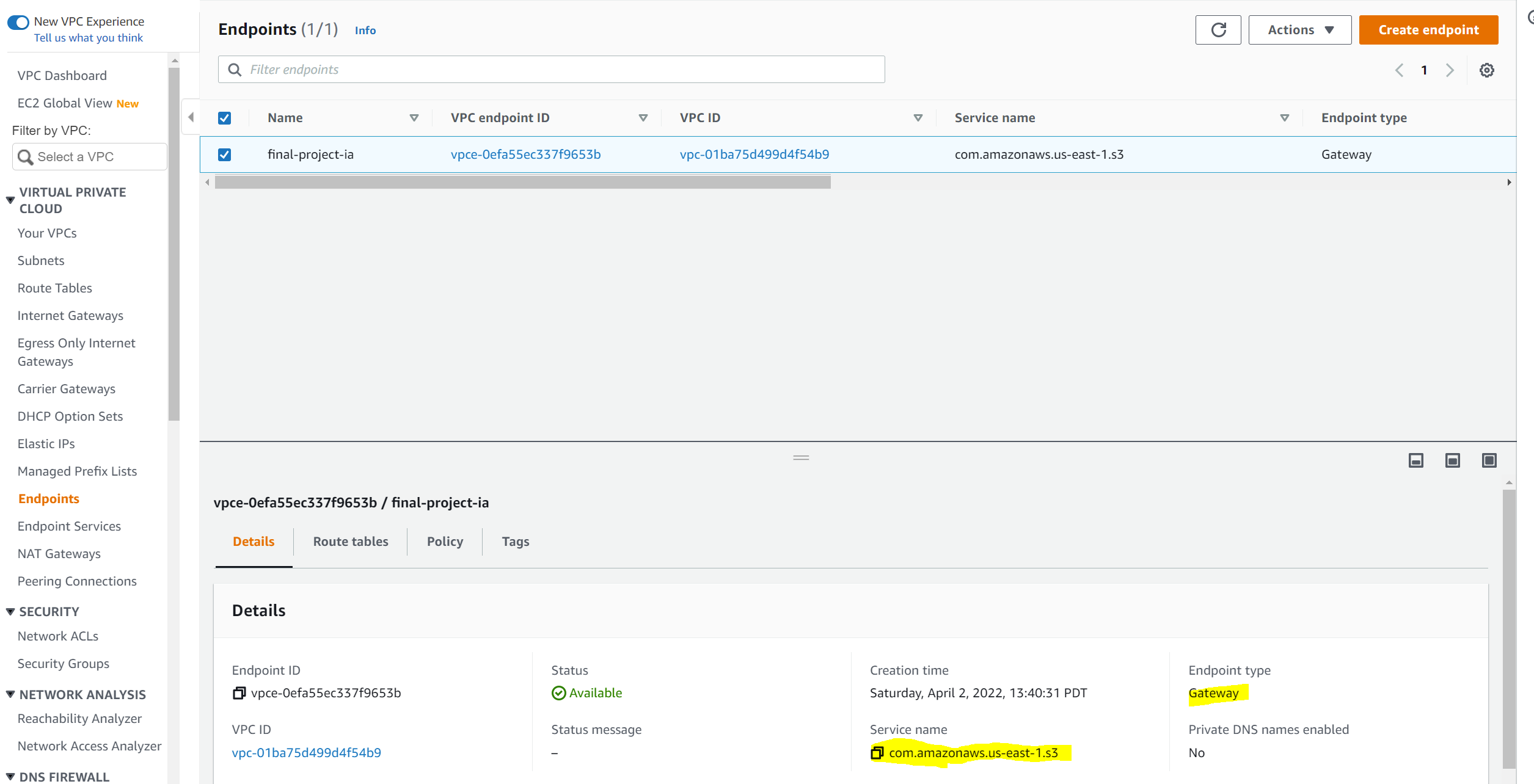
Create Database



Create Connection to SQL

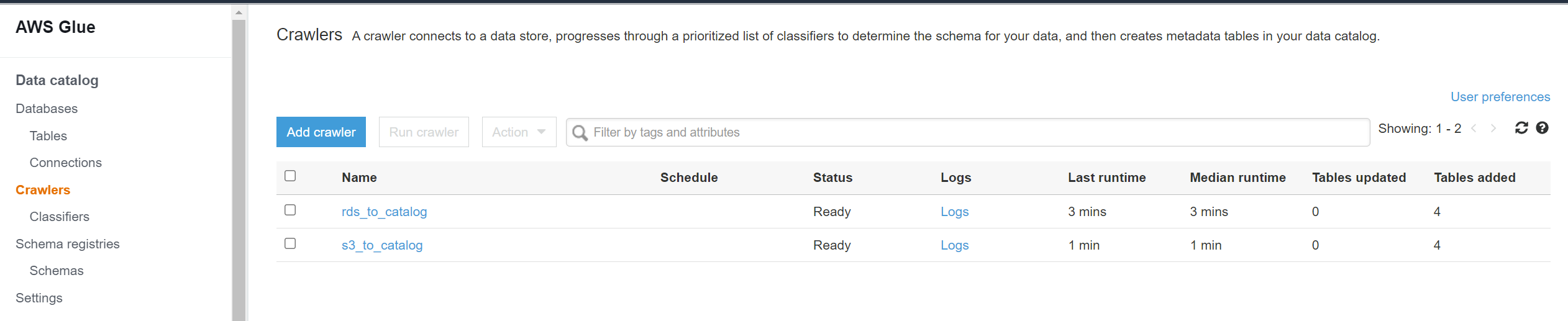


Configuration: VPC service. create endpoint with gateway to connect S3 service, so Glue can do the crawler from s3.

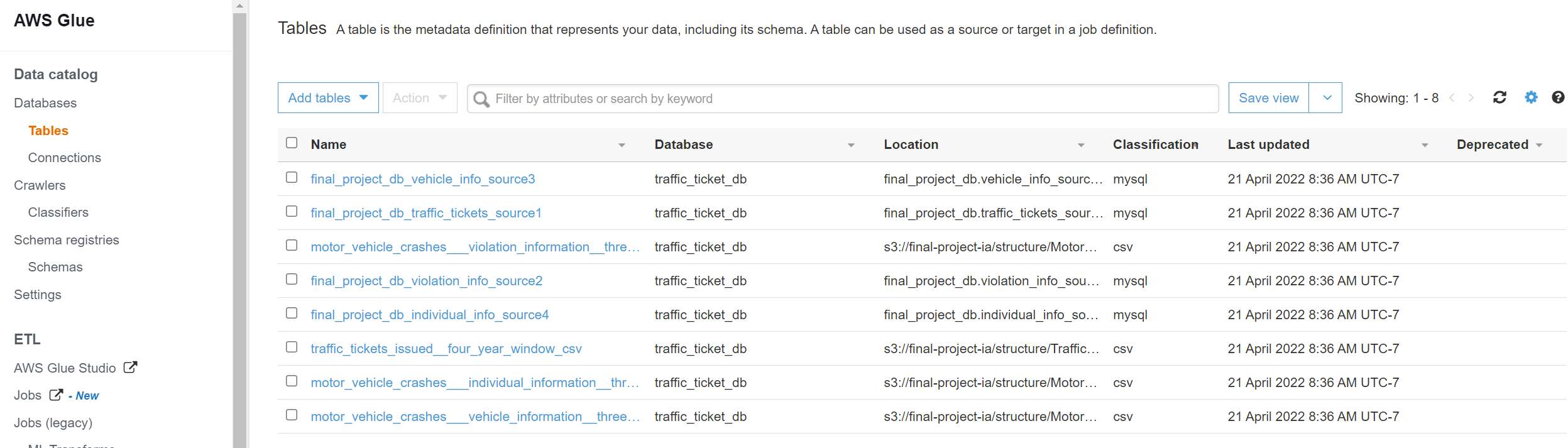


Use the Glue Crawler to copy the tables from sql workbench to Glue data catalog

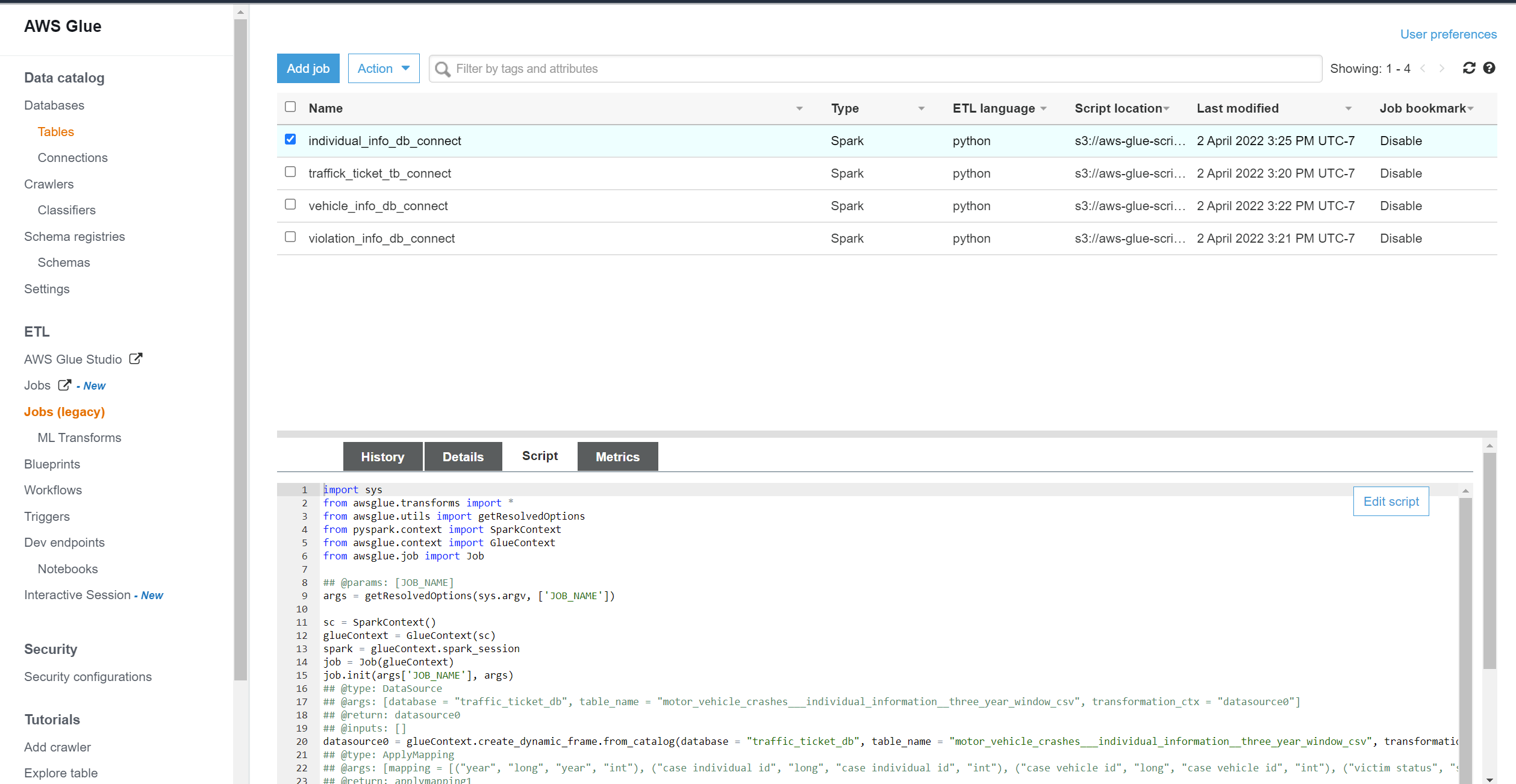
Use the Glue Crawler to copy the tables from S3 bucket/structure/ to Glue data catalog



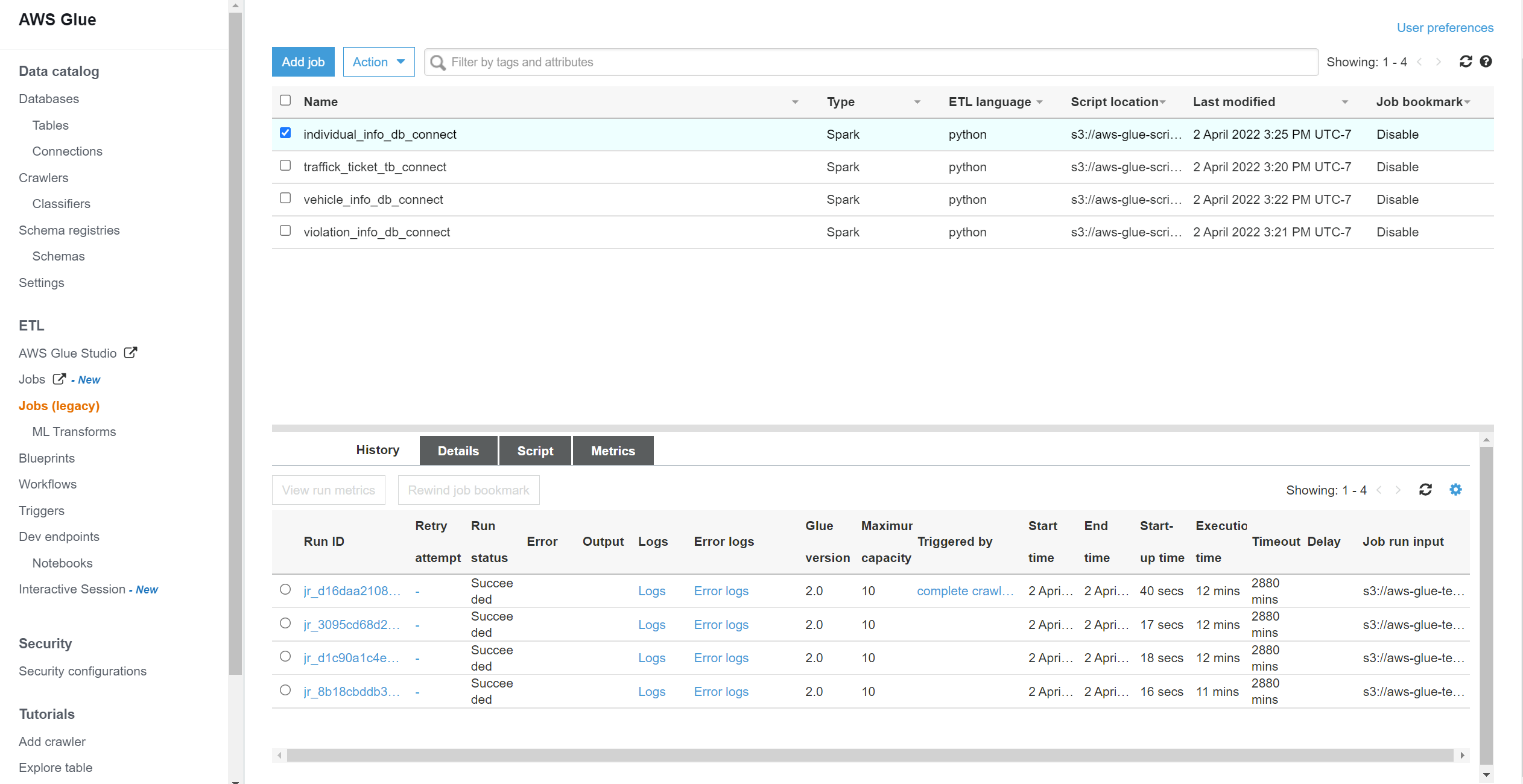
Data Catalog tables (4 from sql, 4 from s3 bucket, totally 8 tables):



Create Glue jobs: script is automatically generated from connection



After running all jobs, it’s successful



All tables from s3 are copied to sql tables. This is the result of one of the tables

