**Batch Time Analysis of Transactional Data**

DESCRIPTION

Lenodo is a multinational e-commerce organization that sells products directly to consumers. The database administrator exports the data every night in a CSV file, but this export functionality is unused. Lenodo wants to use this data to uncover insights about the most-sold item and the countries where customers have bought this item.

You are a data analytics consultant, and you're asked to provide valuable insights and statistics across products, brands, categories, segments to the marketing, product, sales, and procurement teams and inform them about which product has the highest amount of sales and which product and its marketing needs the most improvement. These statistics will help to run effective digital marketing campaigns. The scope of this project is limited to data engineering and analysis.

**Objective:**

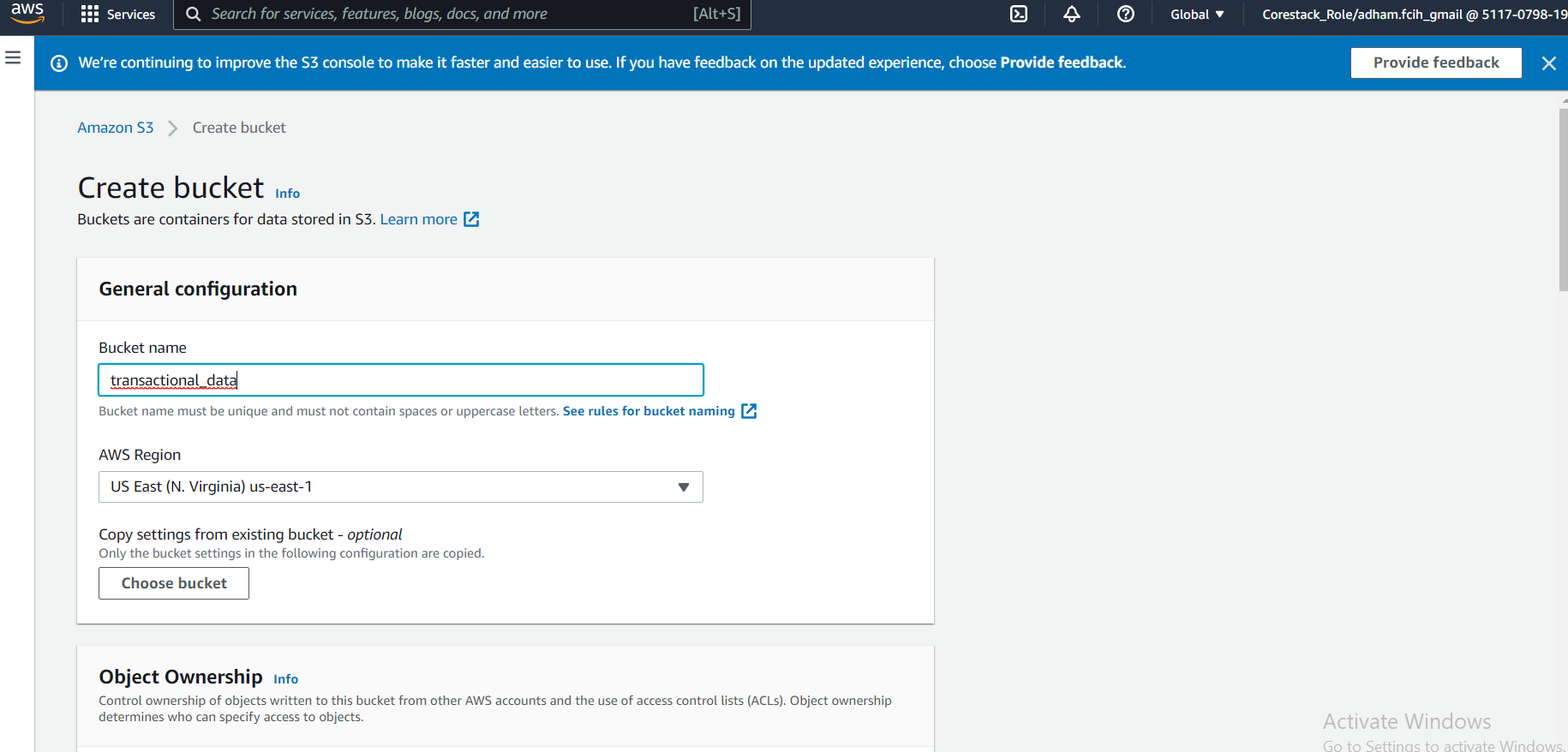
To use AWS Big Data stack for data engineering to analyze transactions, uncover patterns, and share actionable insights

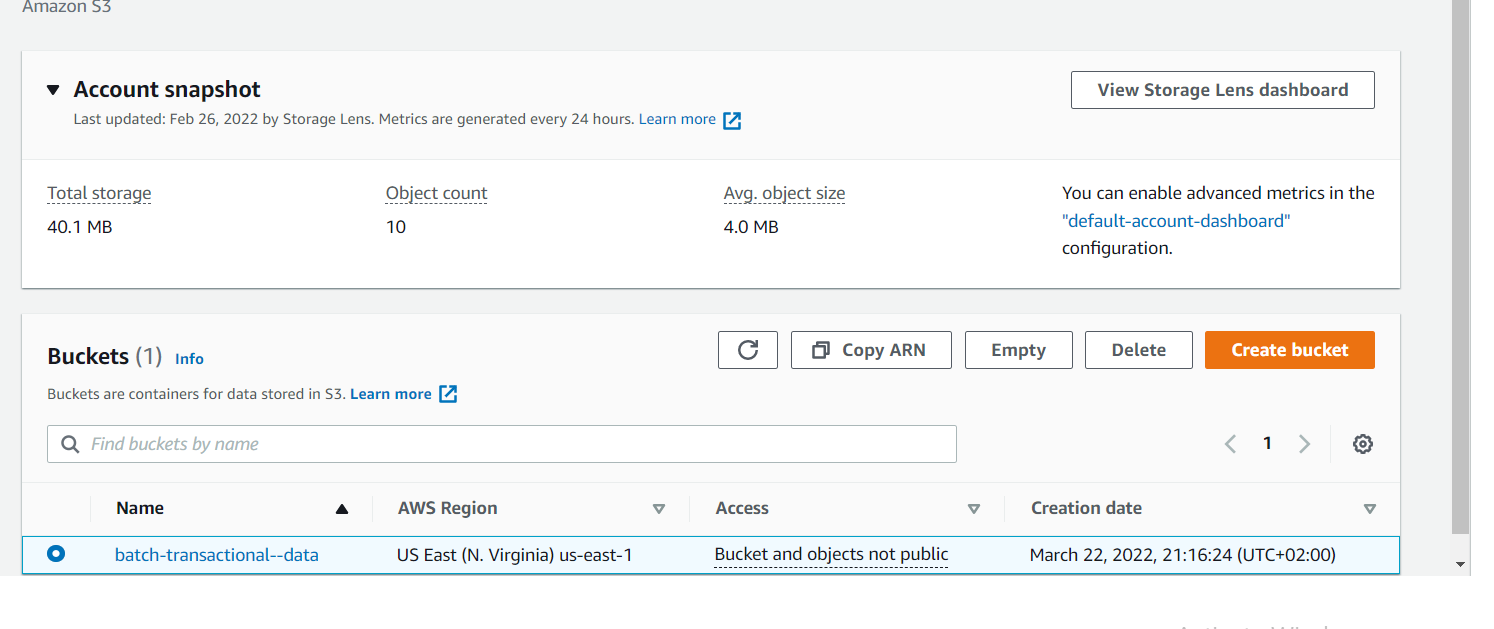
**Steps to perform:**

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)
2. Create a crawler to crawl the CSV data and generate a metadata catalog
3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries
4. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena
5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

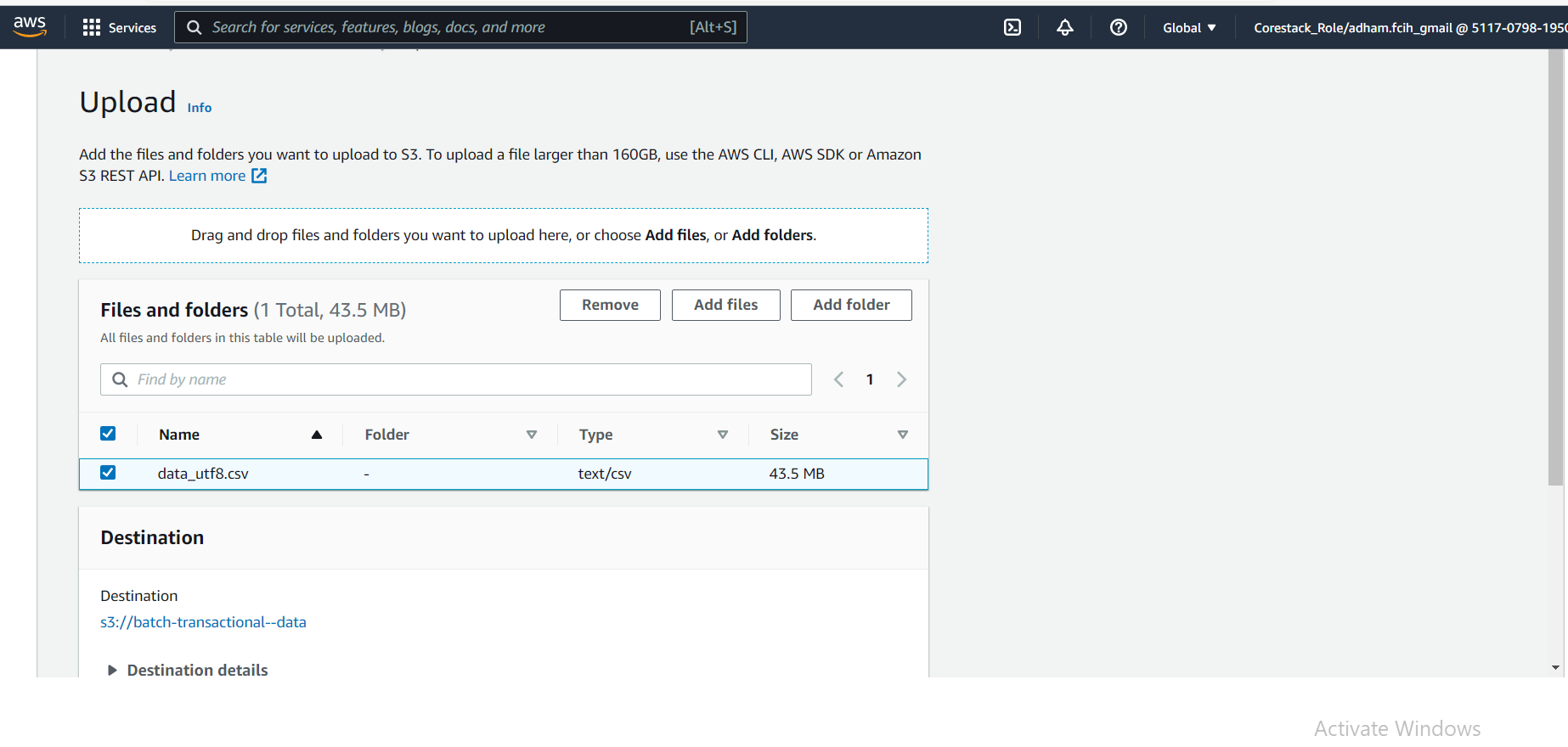
-------------------------------Solution -----------------------------

1. Login to AWS Lab and Create S3 bucket to upload the csv used file



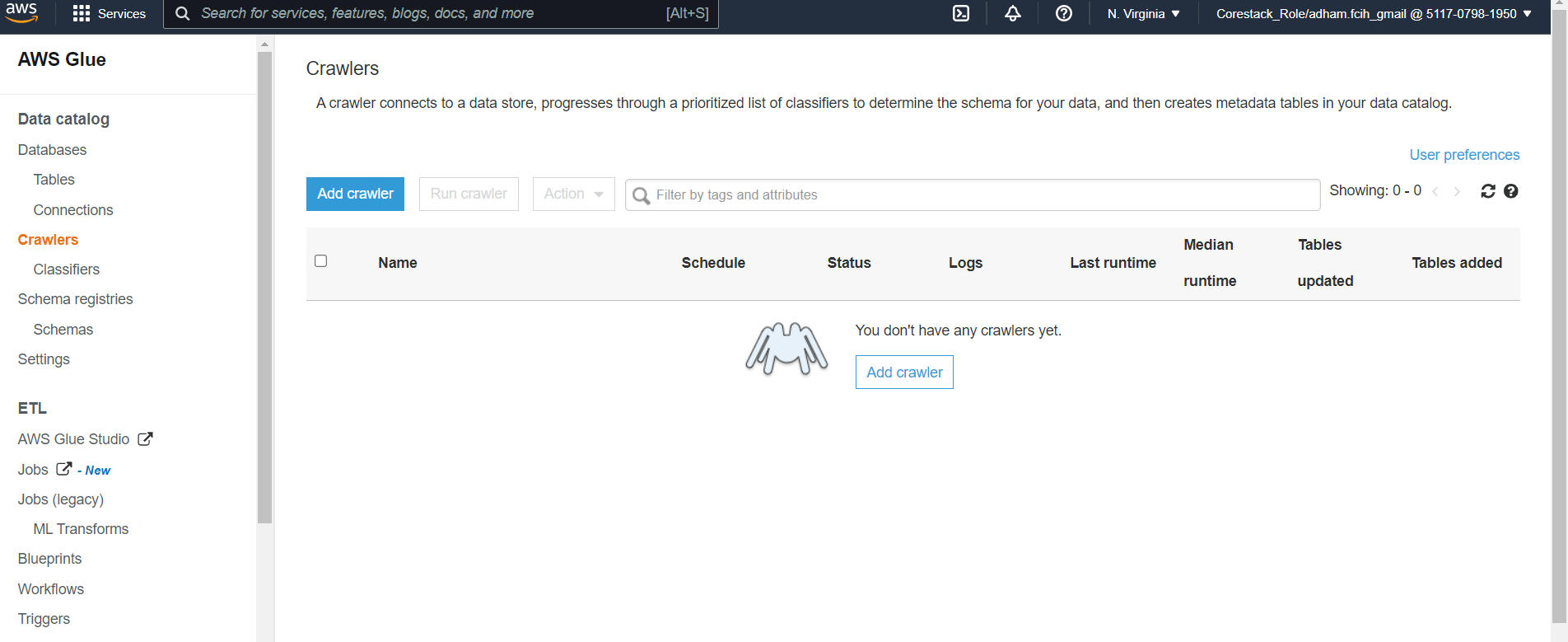


Now upload my transactions data file

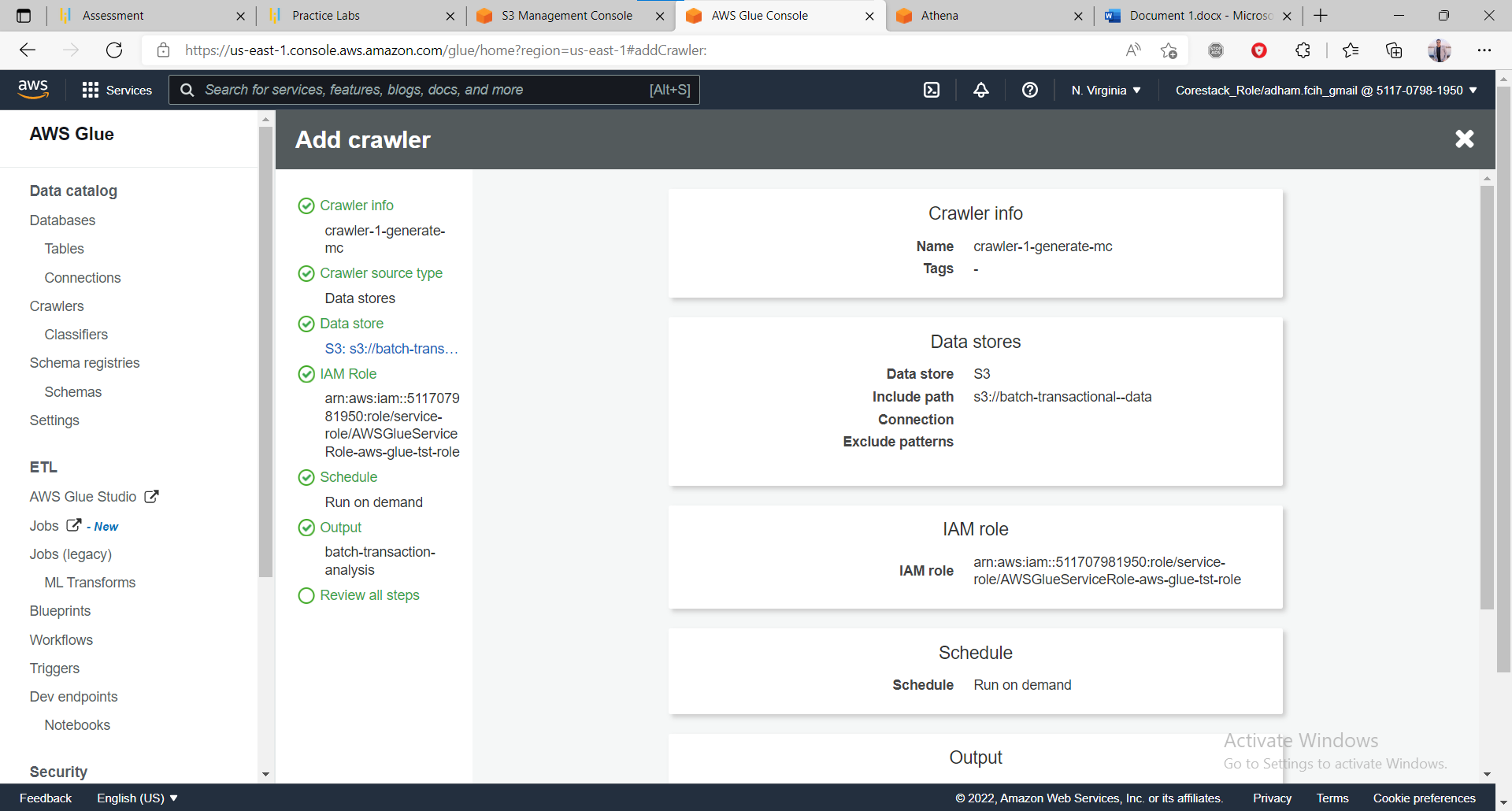


1. Create a crawler to crawl the CSV data and generate a metadata catalog

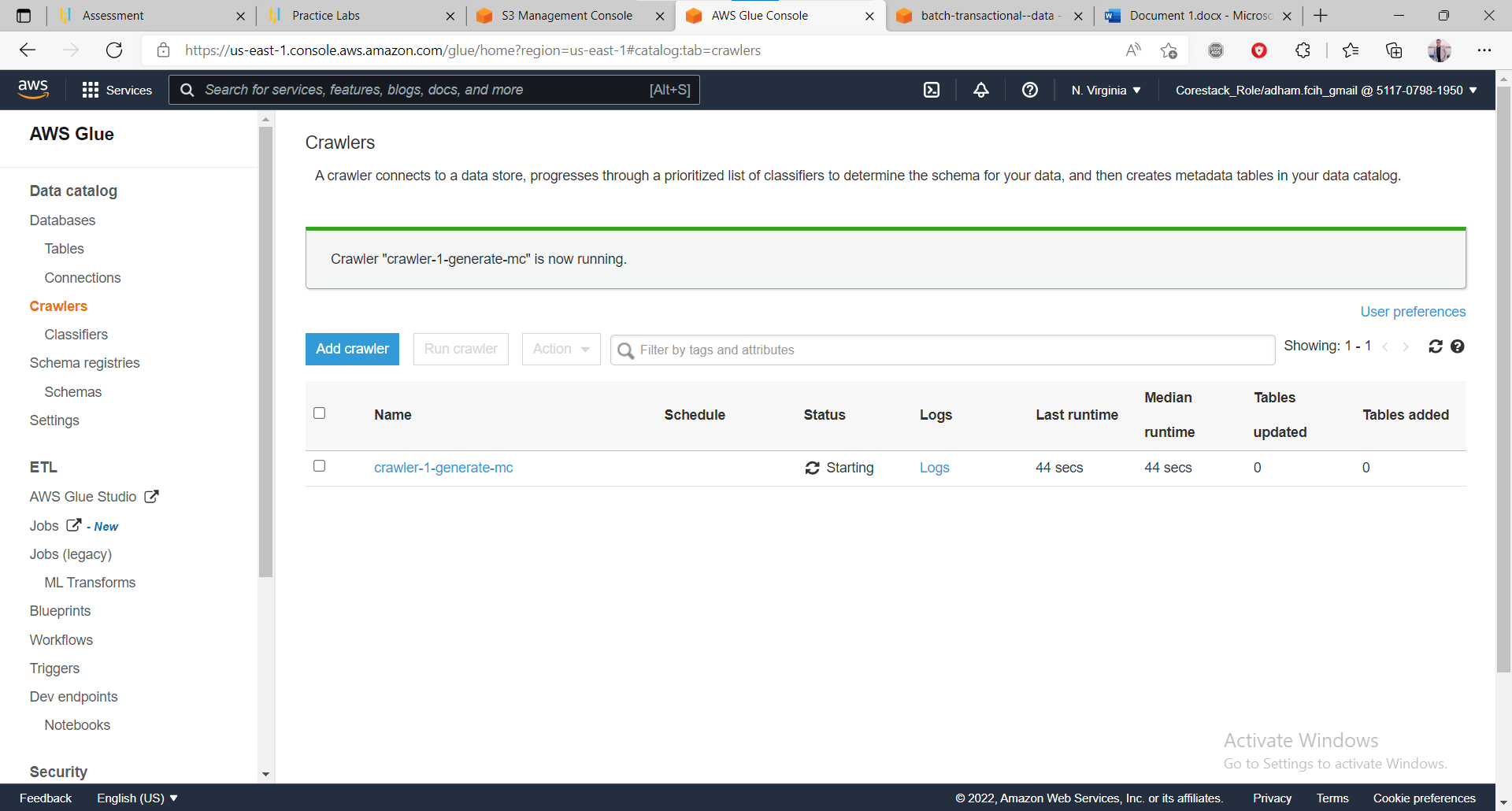
Open AES Glue and add crawler



Crawler configuration to read from my S3 folder where I put my transaction file

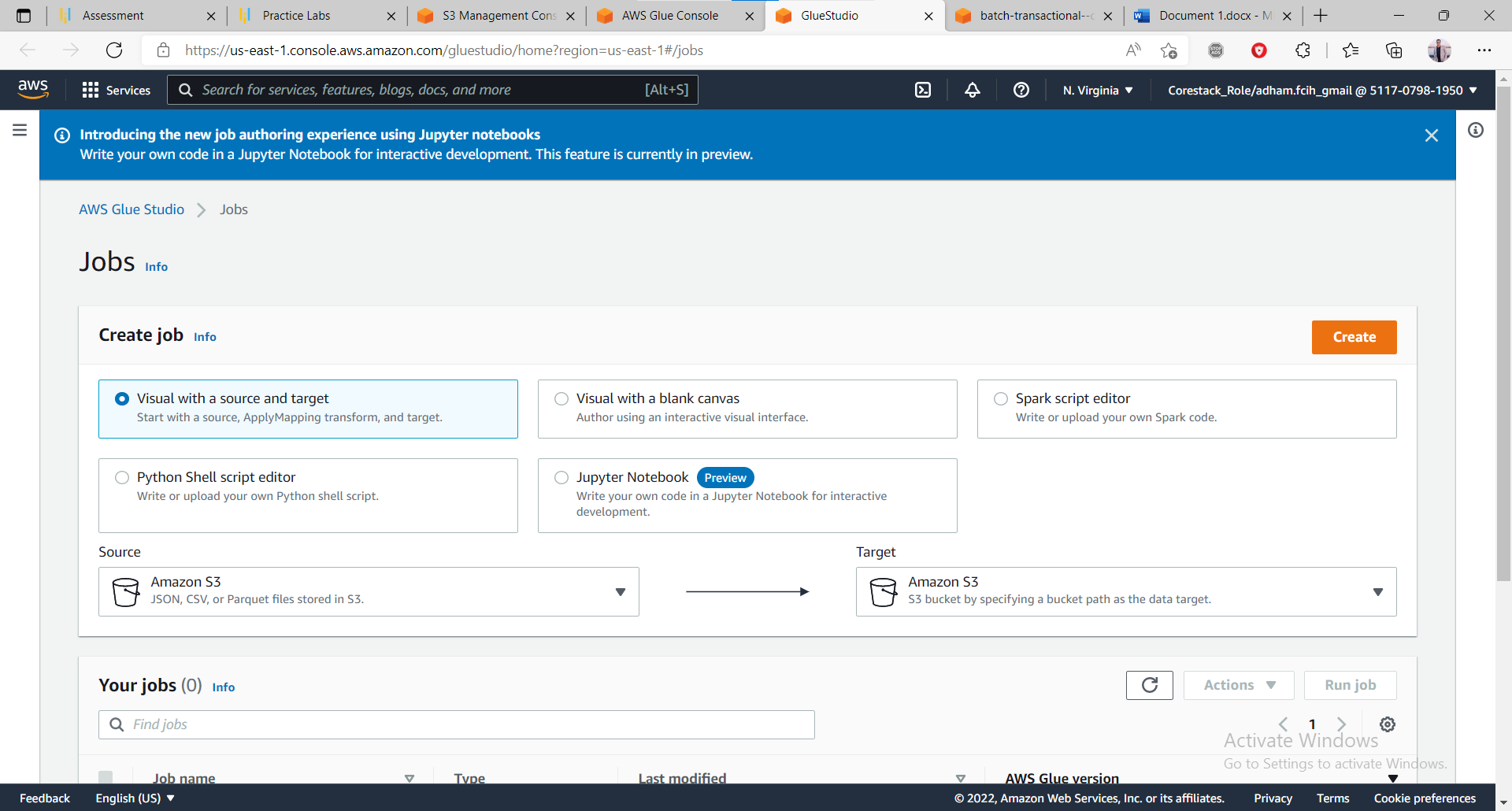


Now run crawler to get metadata from my file and put in Data catalog



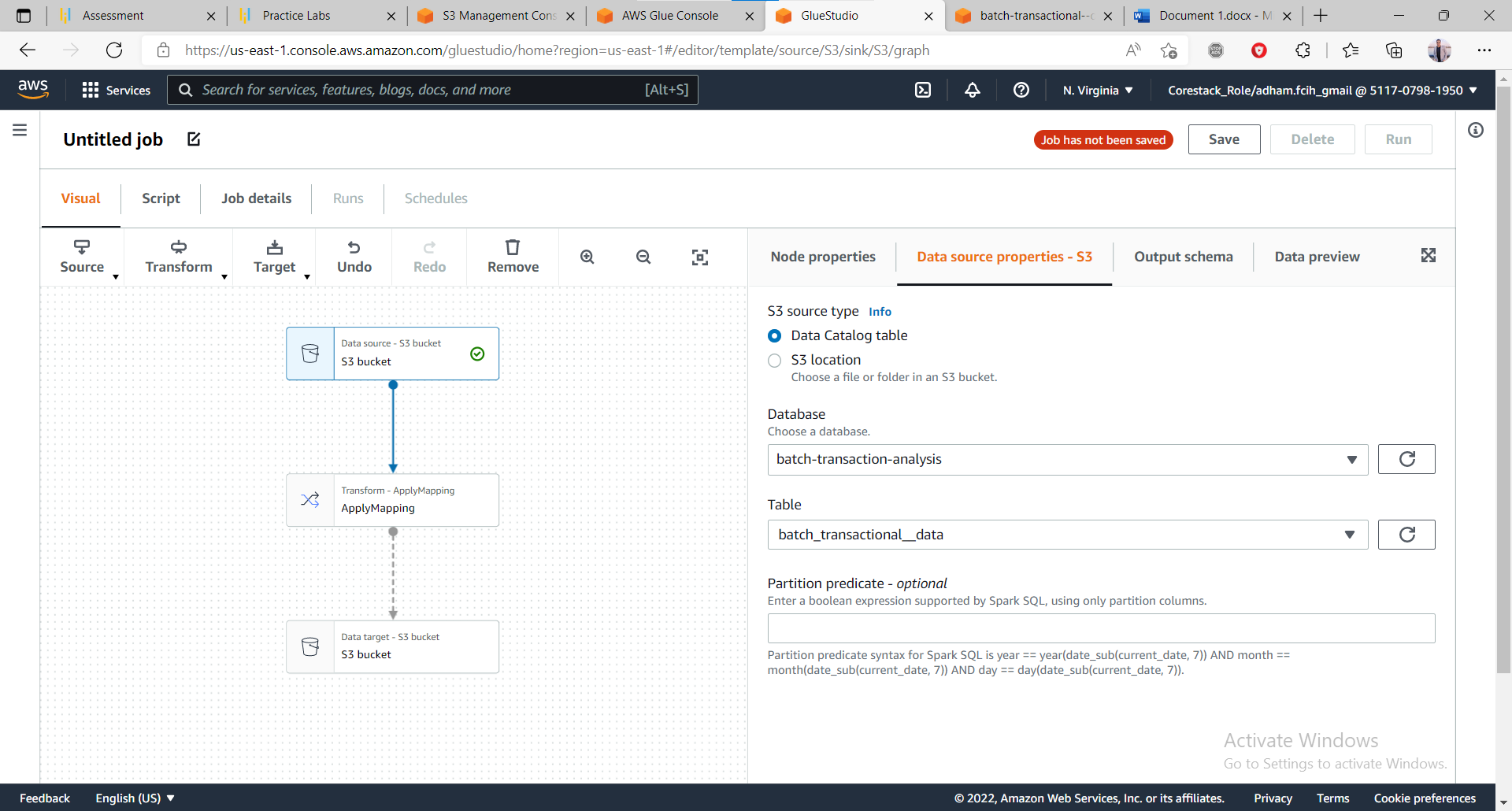
1. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

Now navigate to jobs section and create new job

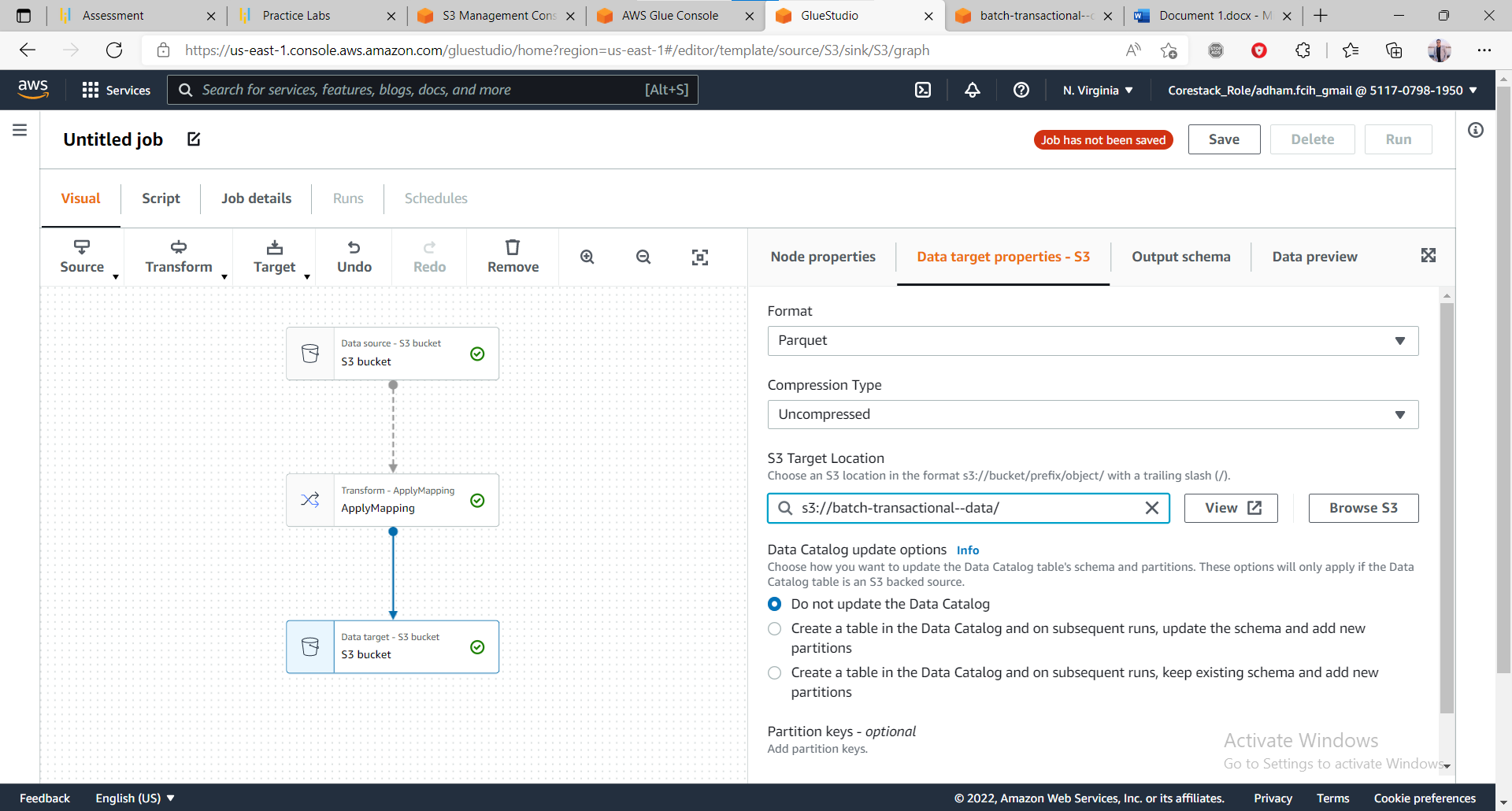


Configure job to get my file as source and make transformation mapping to schema

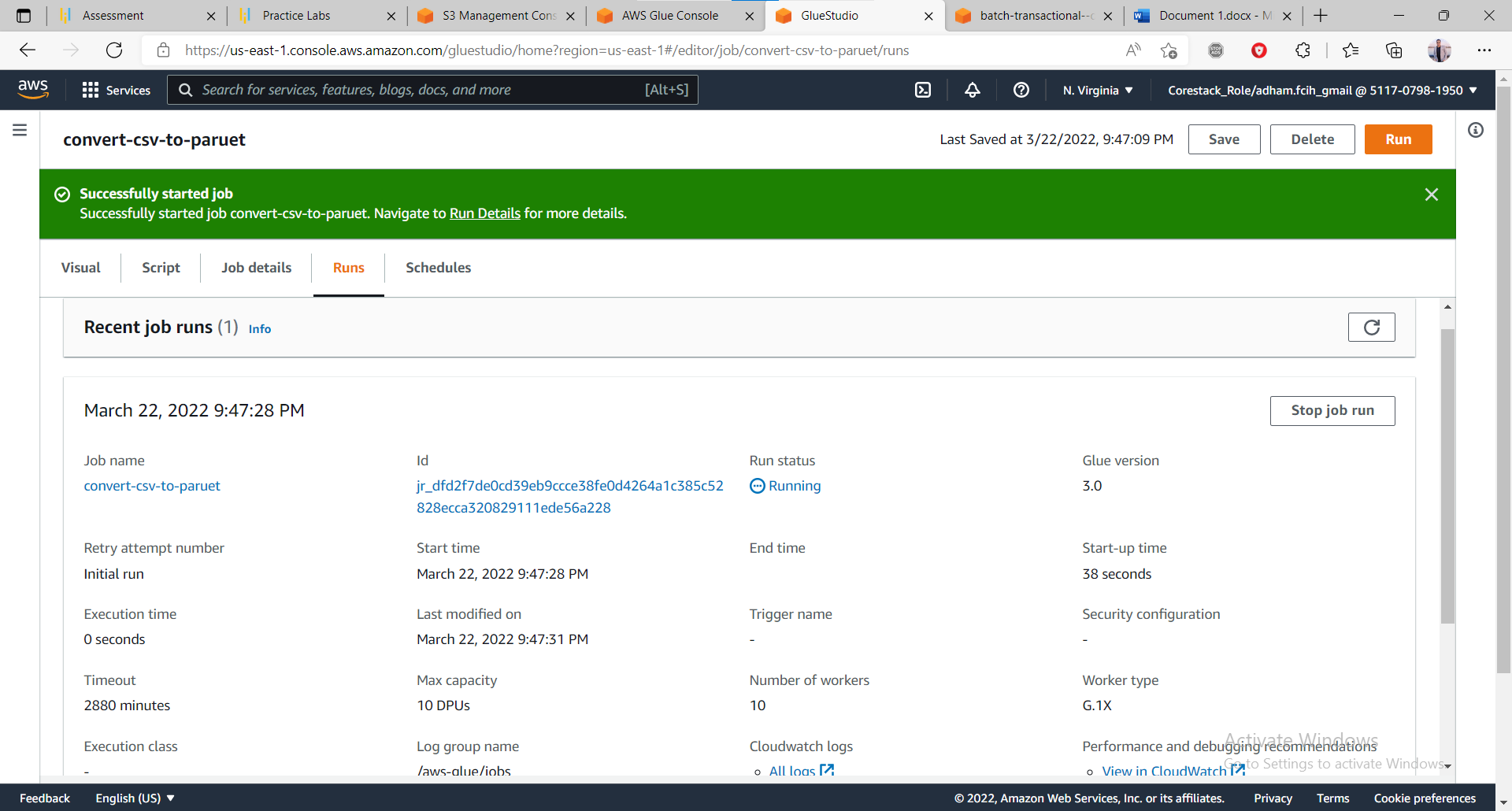
And configure to convert my file to parquet format by choose my destination to parquet



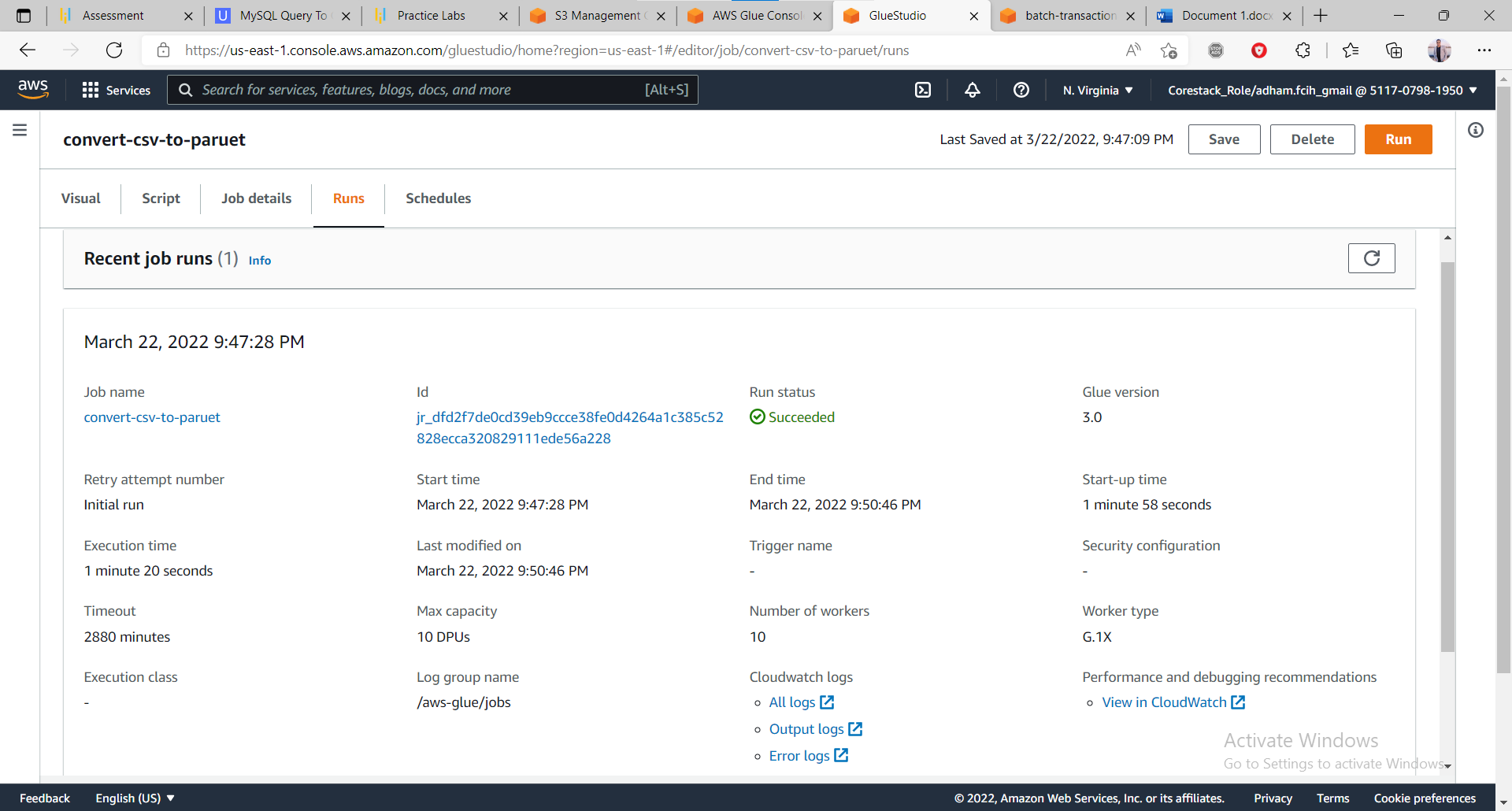
Make my configuration and save job



Job running now

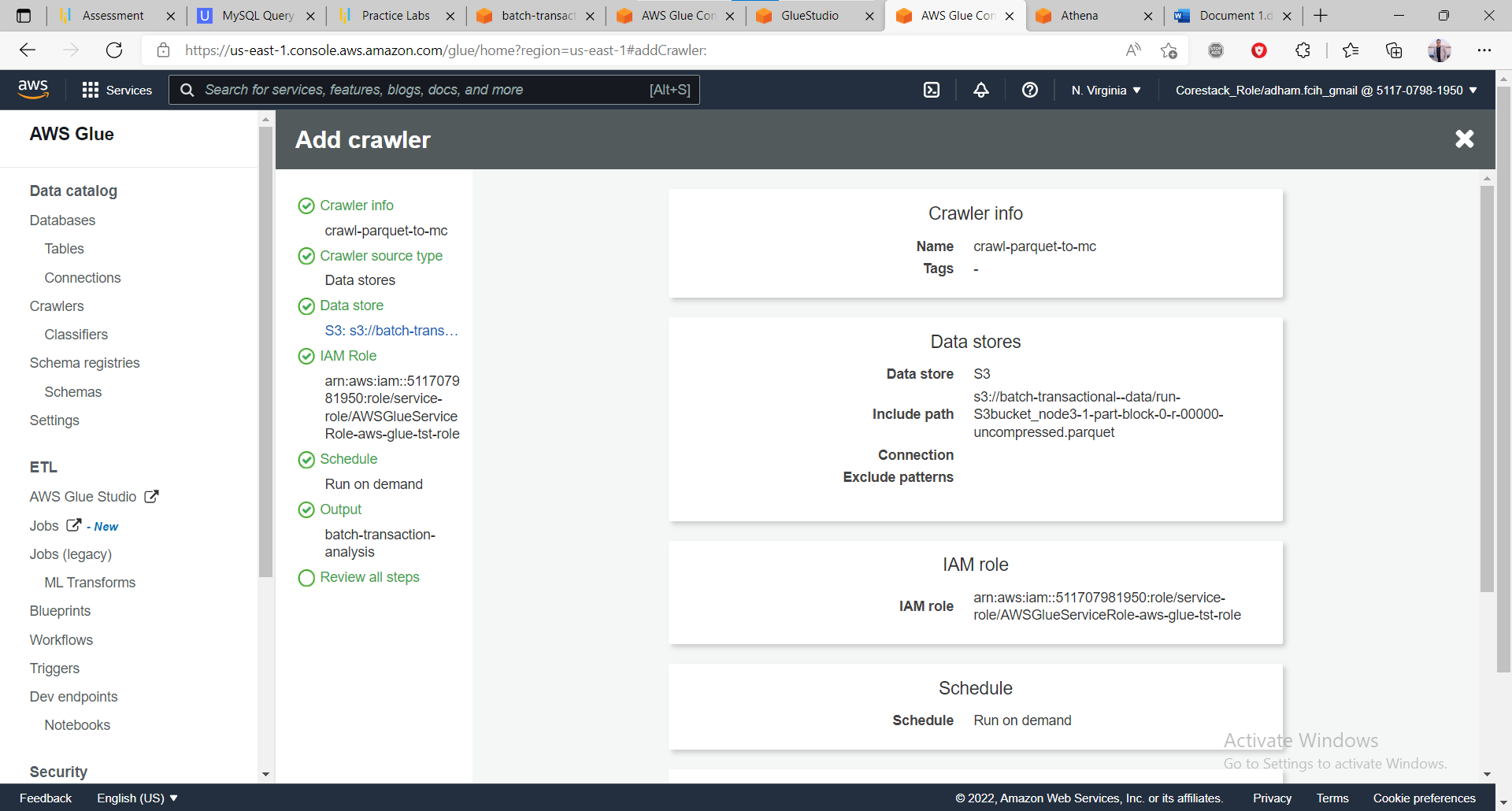


Job runs successfully

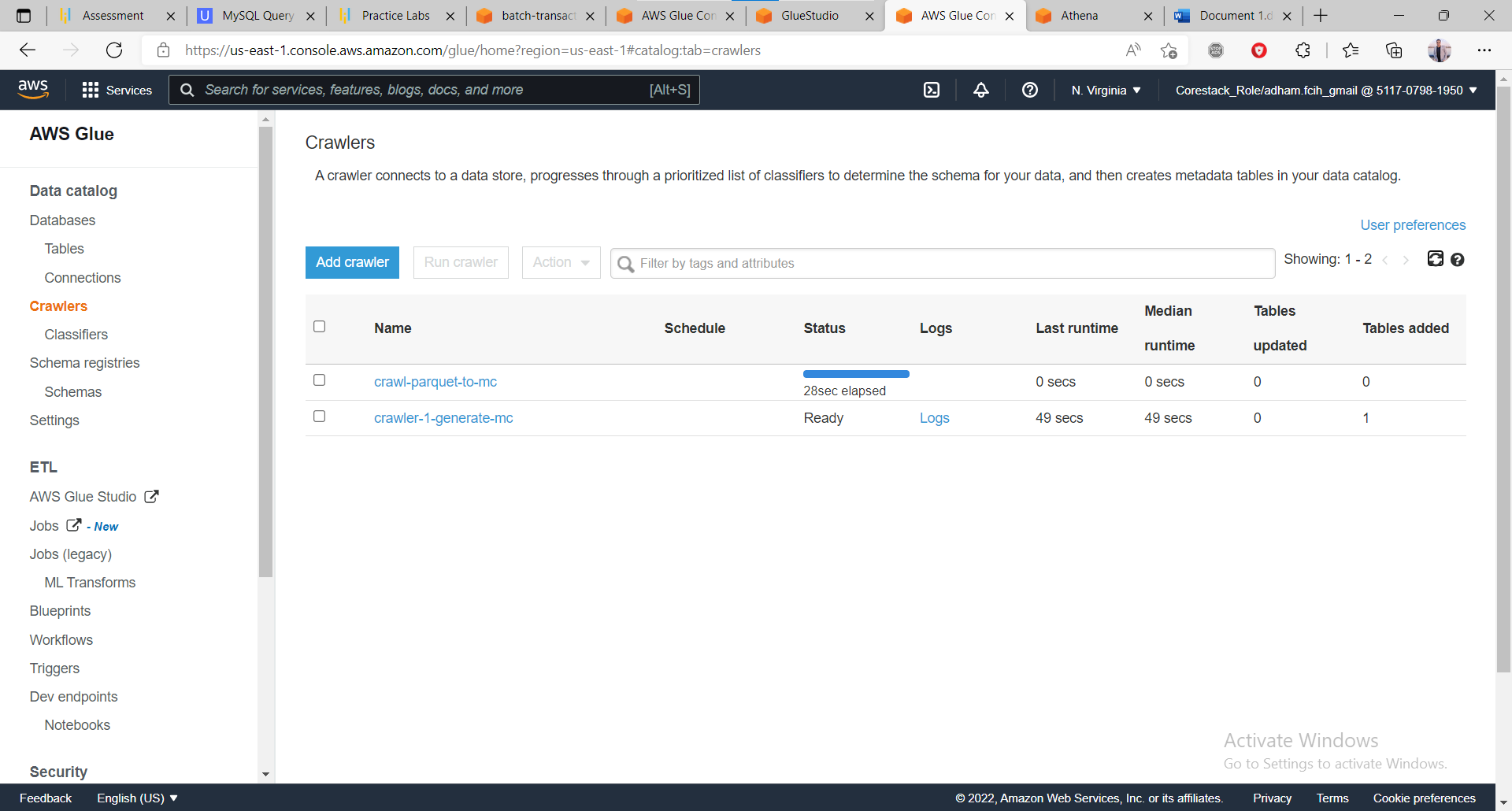


1. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena

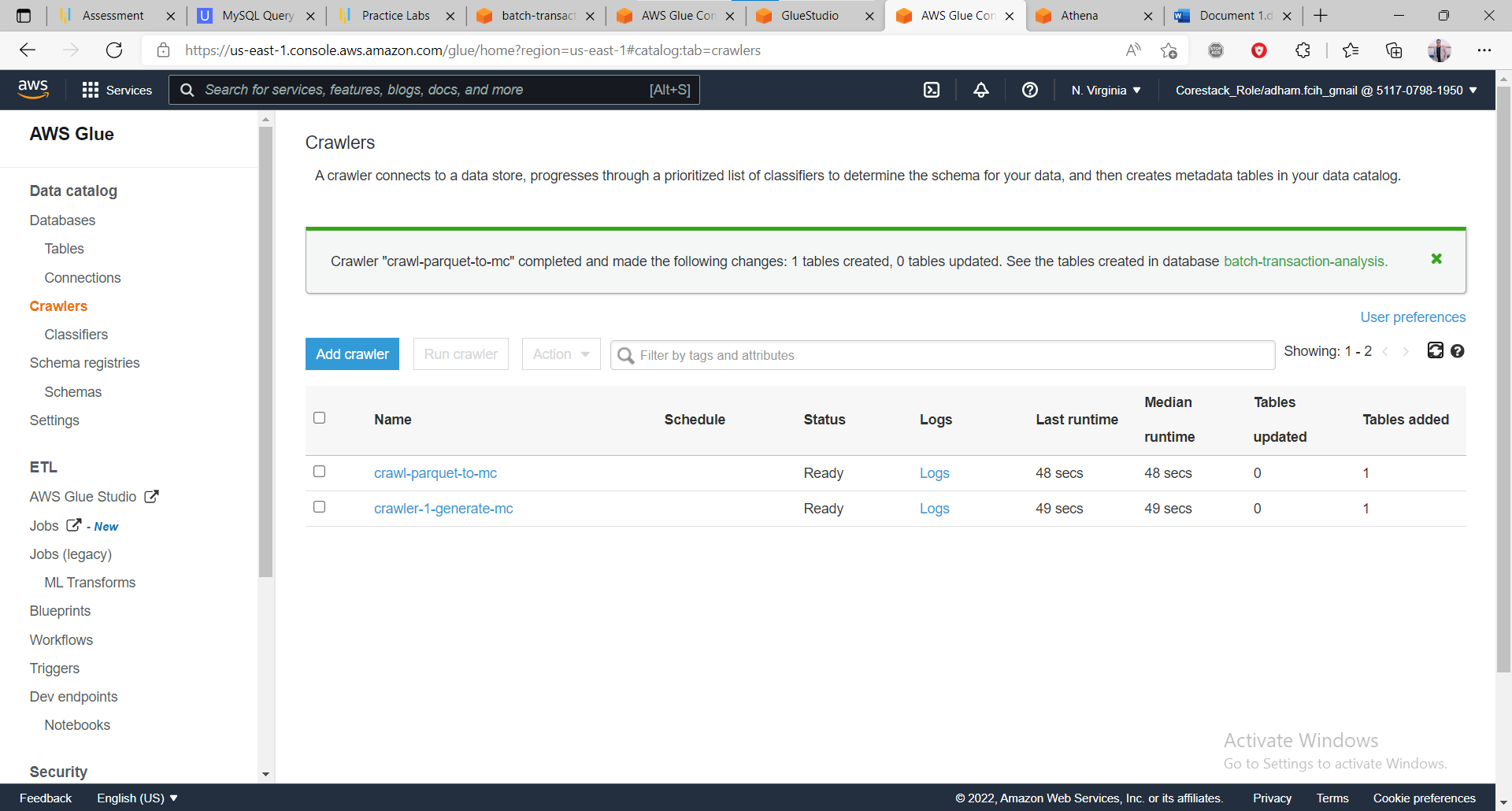
Navigate to glue and add new crawler and configure as previous one



Crawler running

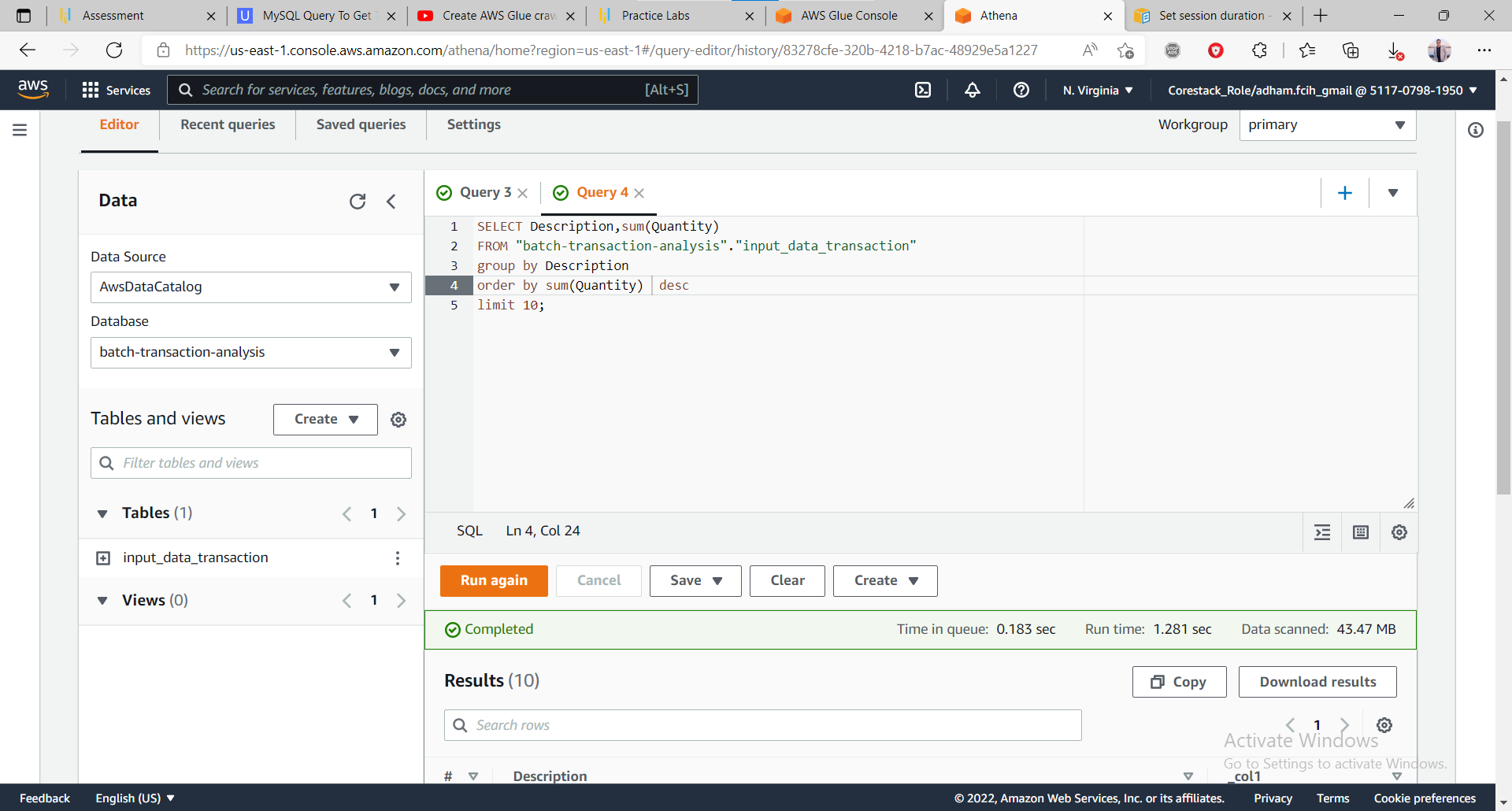


We have two crawlers

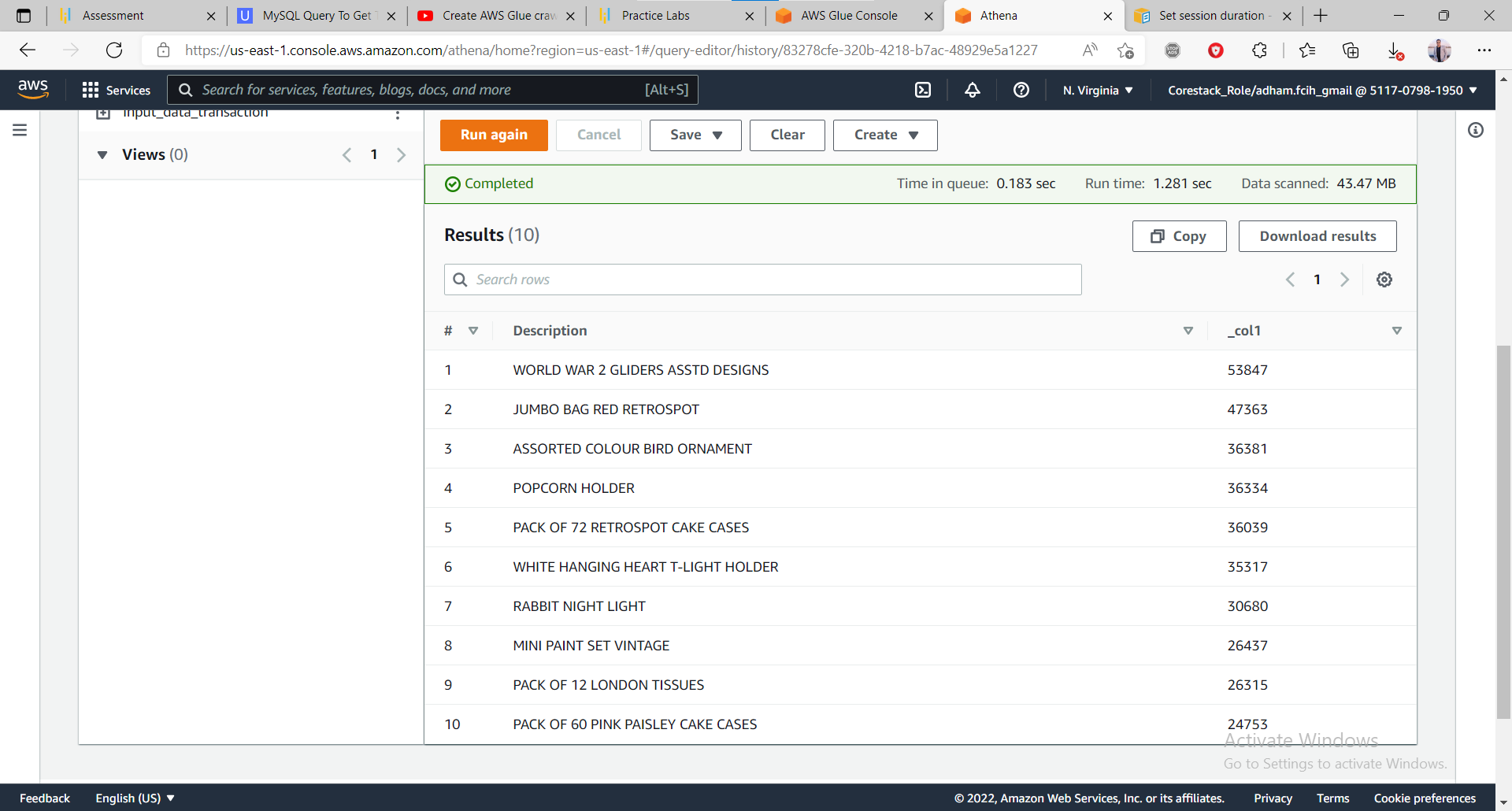


1. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

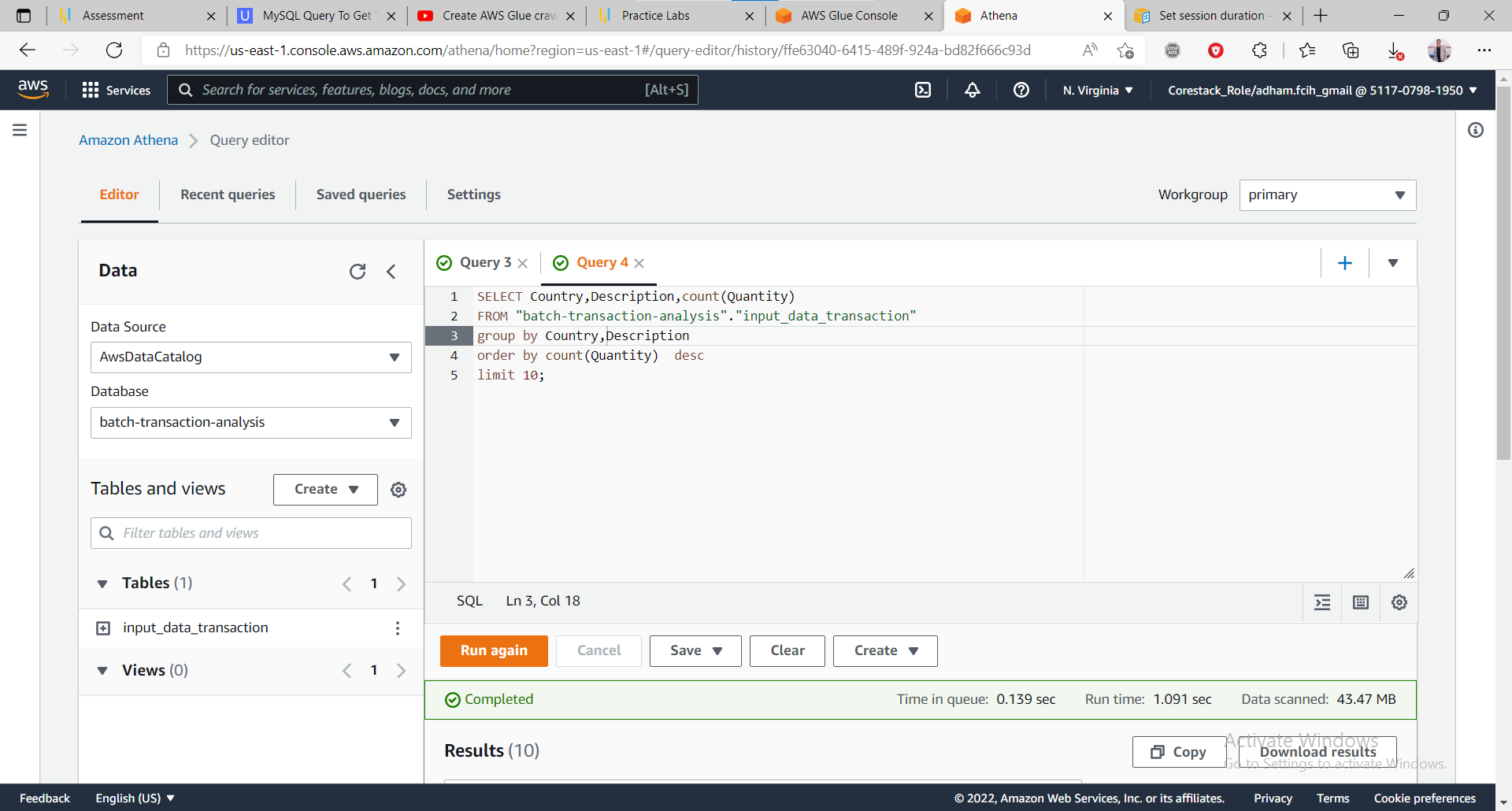
Now analyses my data by query my table in Athena and getting top 10 item by sold quantity

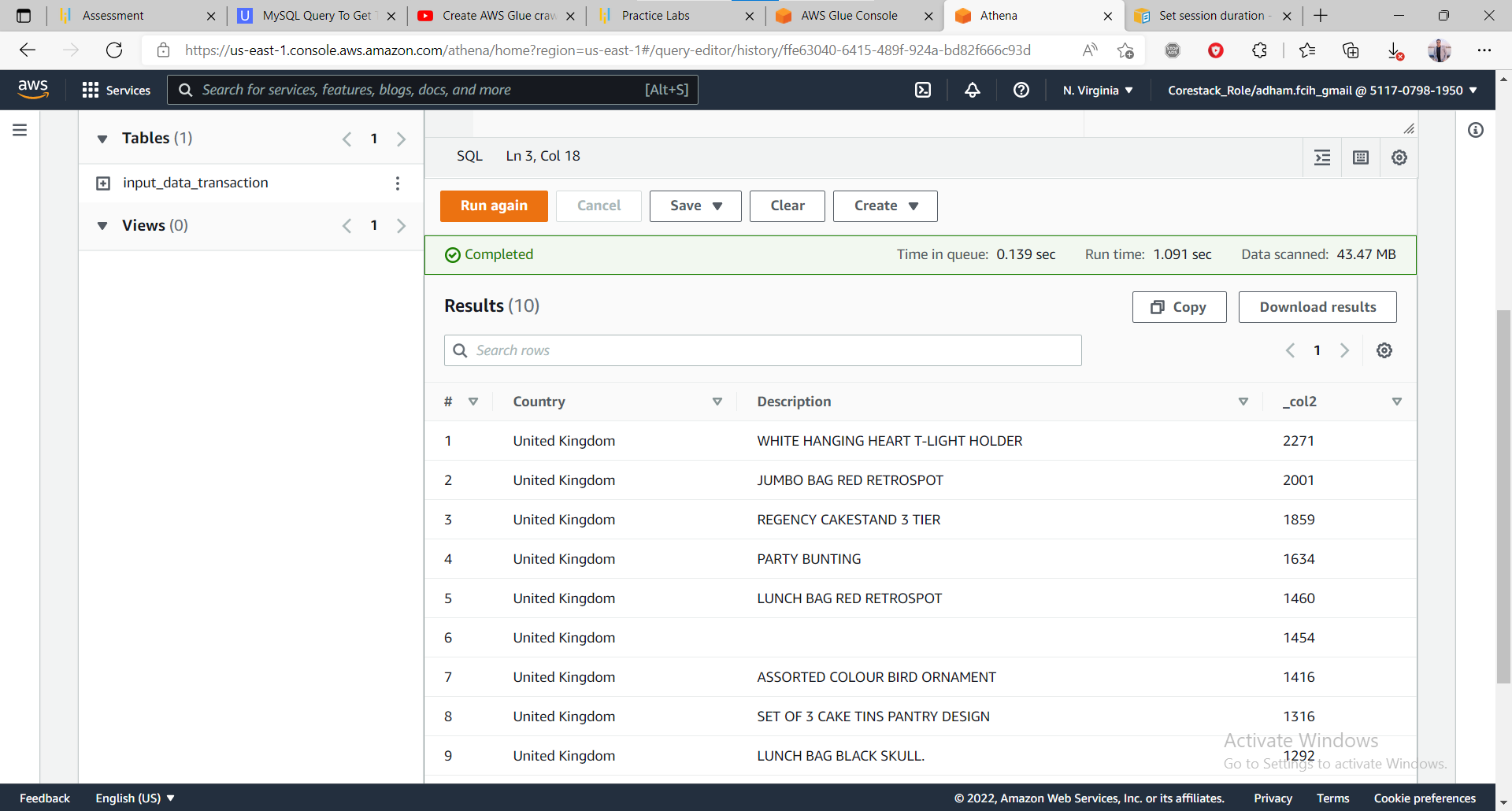


Result of query

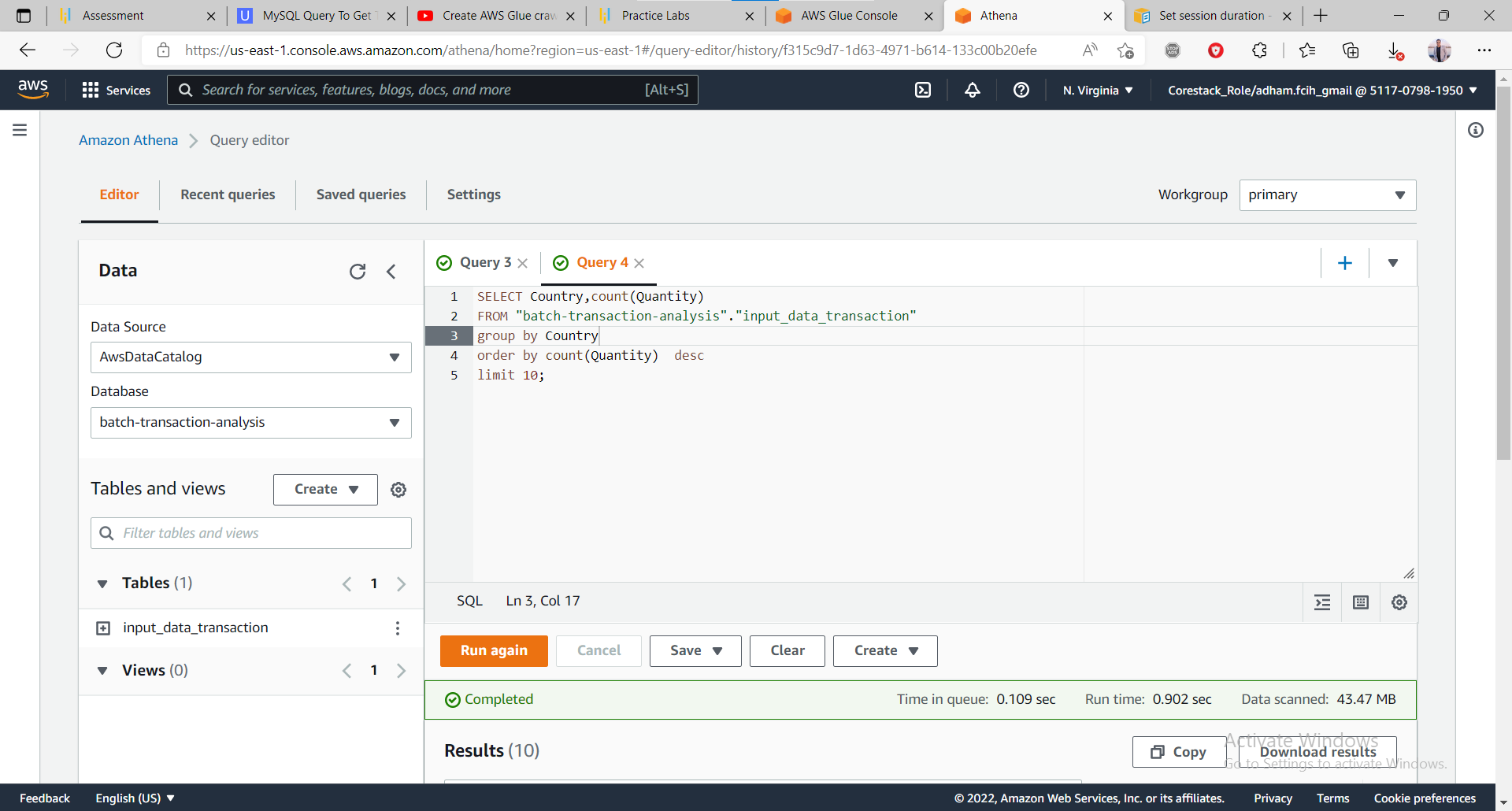


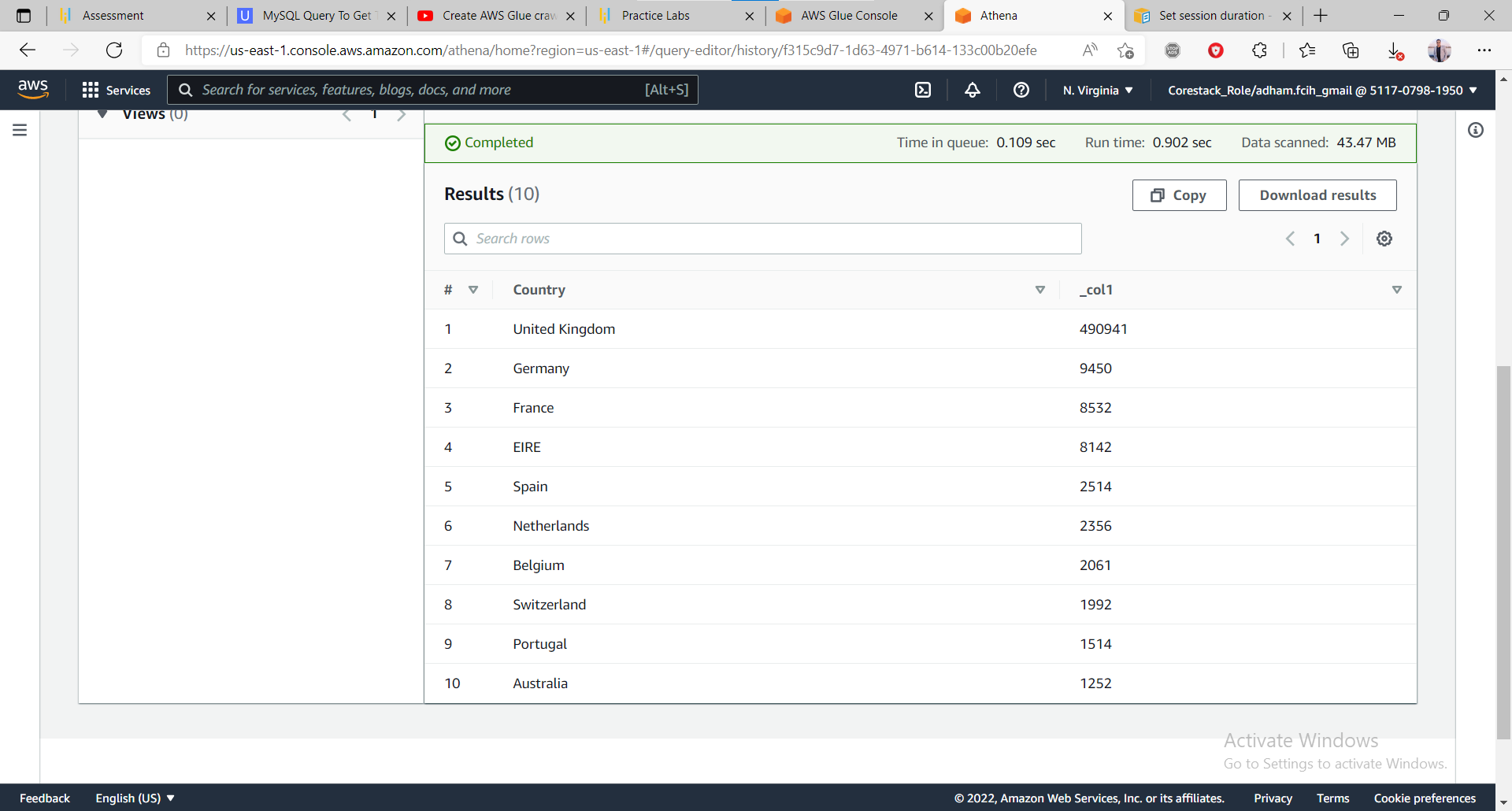
Now getting count of sold quantity over top 10 country and items sold

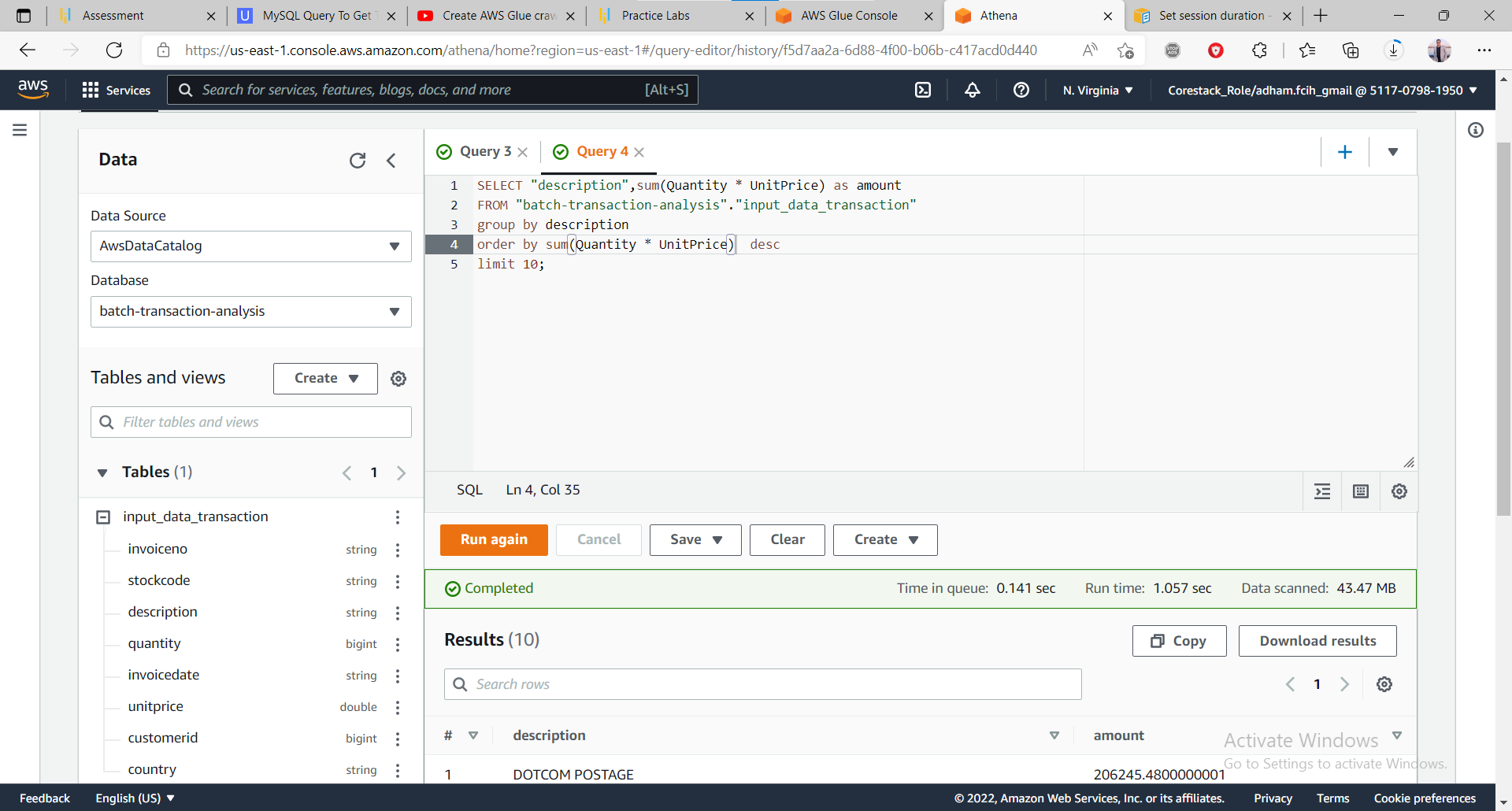


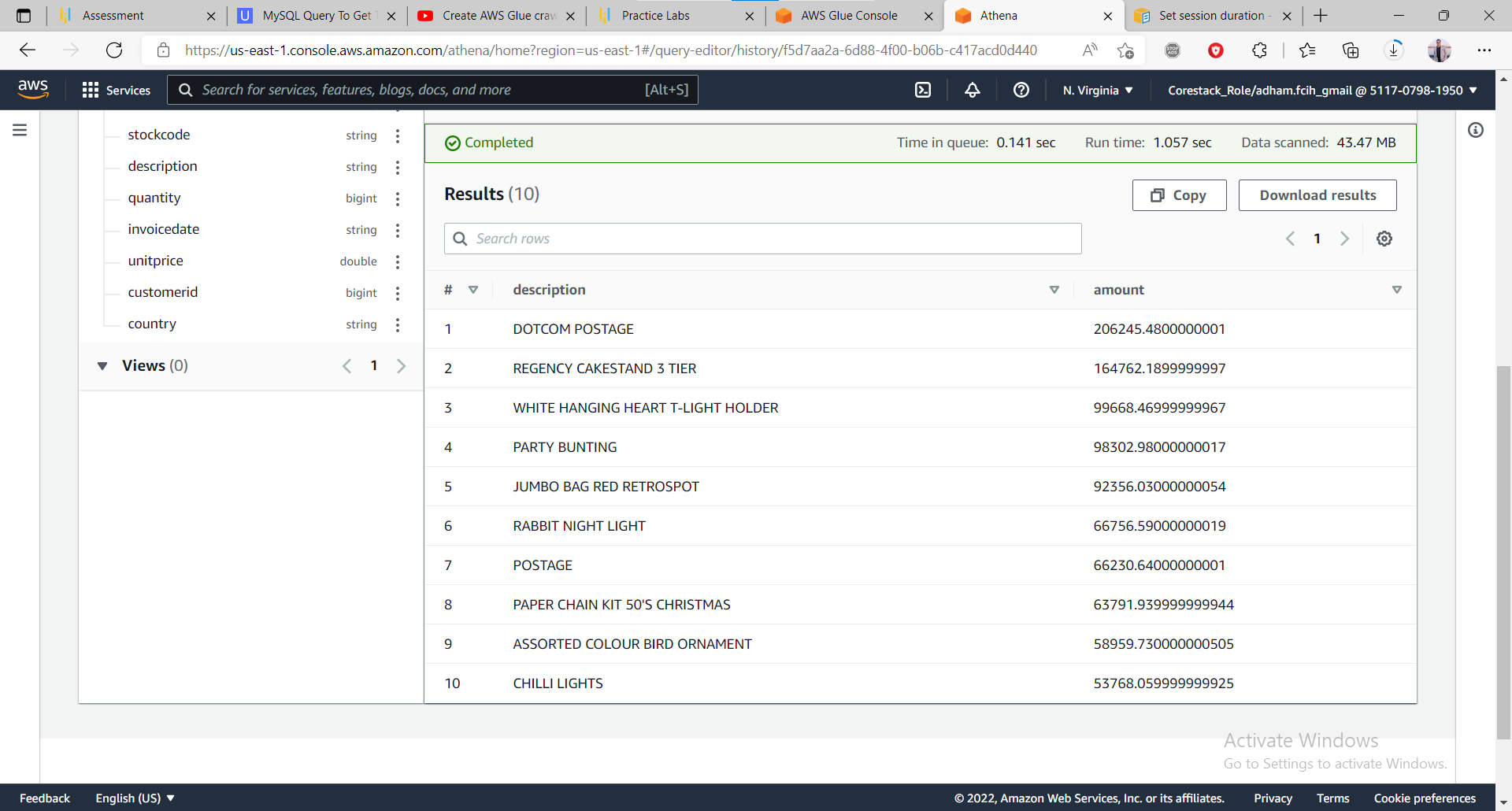


We see that United Kingdom best selling area

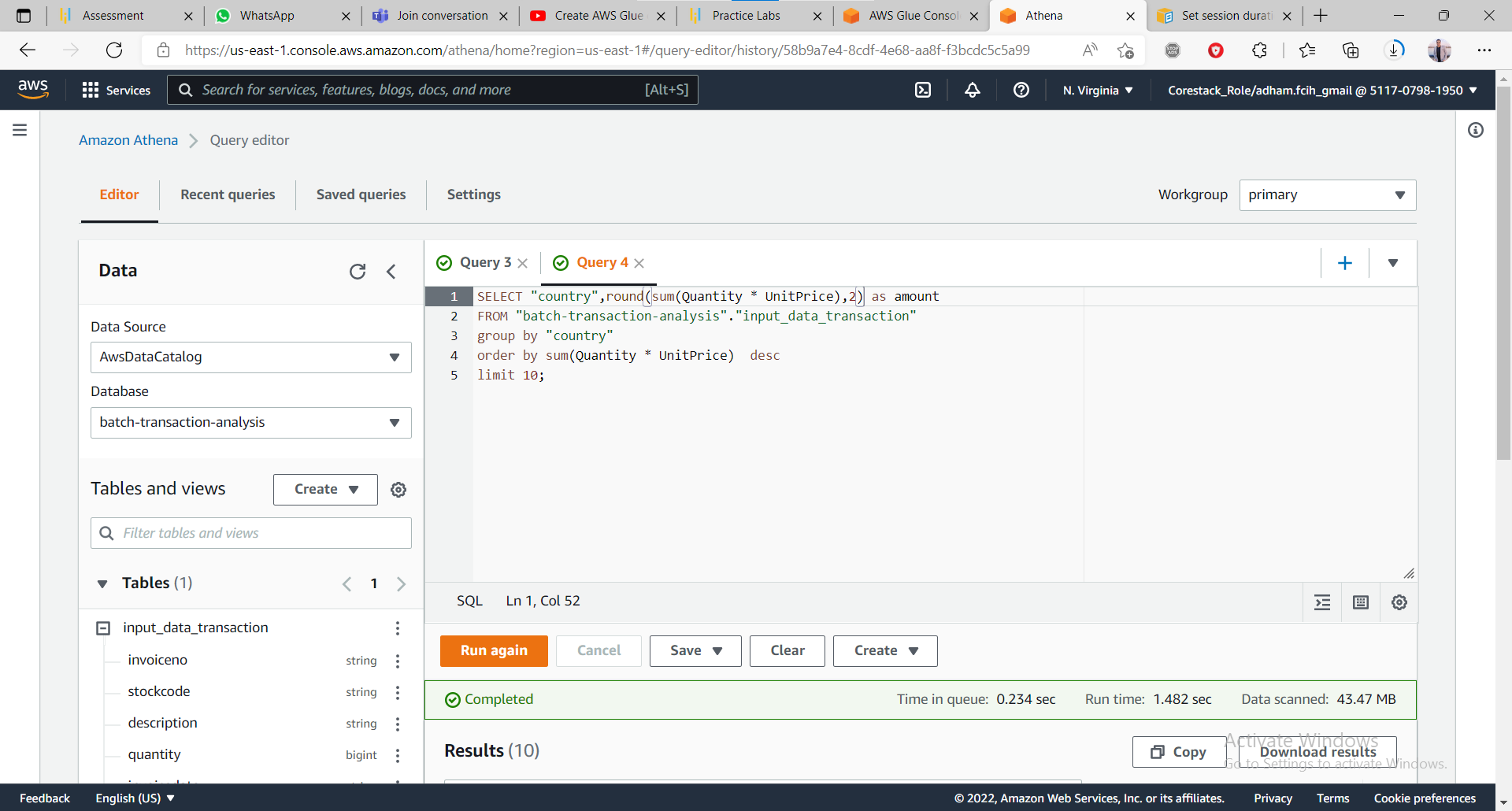








Here we get data based on amount by multiply quantity with unit price and get top 10 counties best selling based on amount



Result of query

