PGP DE - Big Data on AWS

# Batch Time Analysis of Transactional Data Assessment

## 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

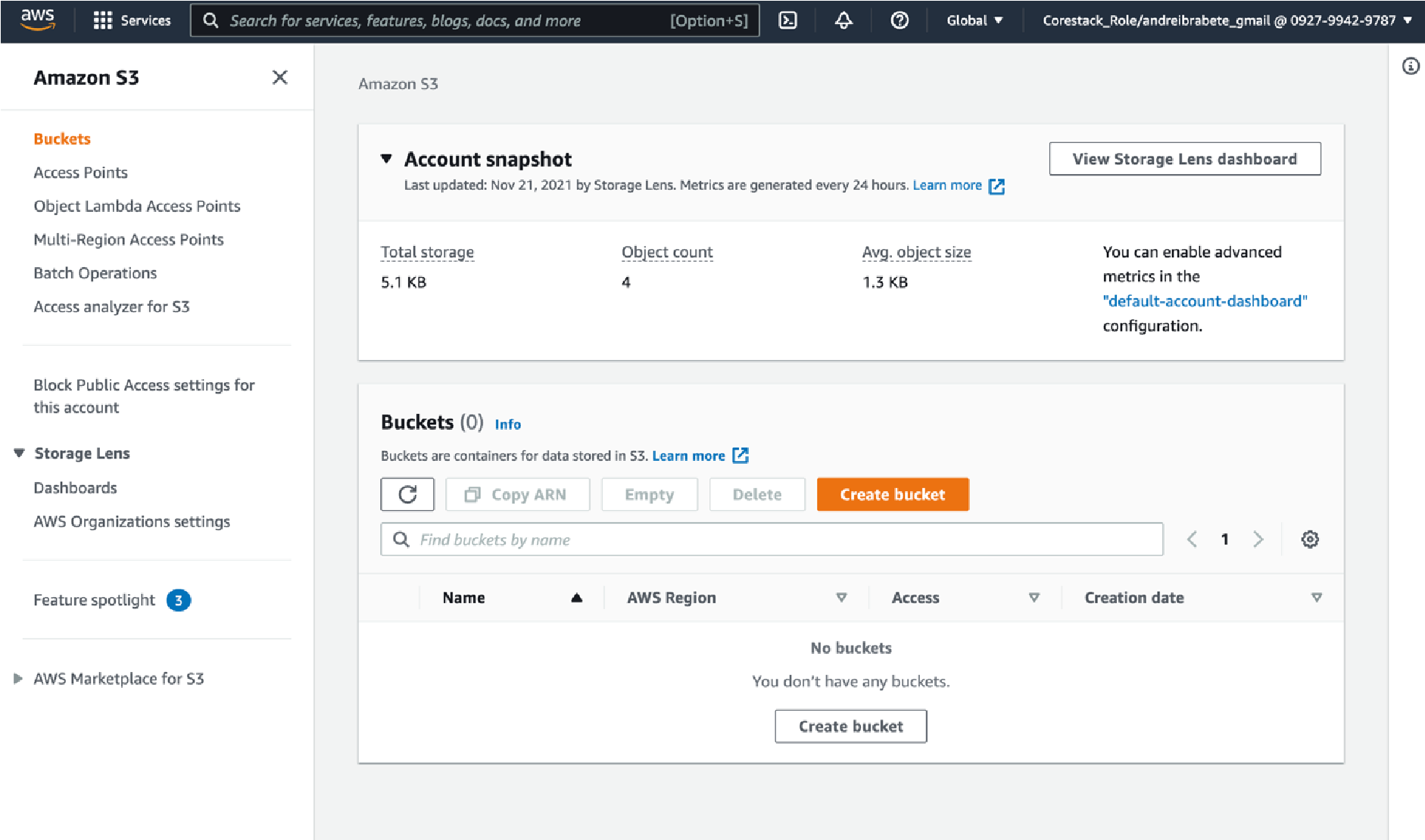
Steps to perform:

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

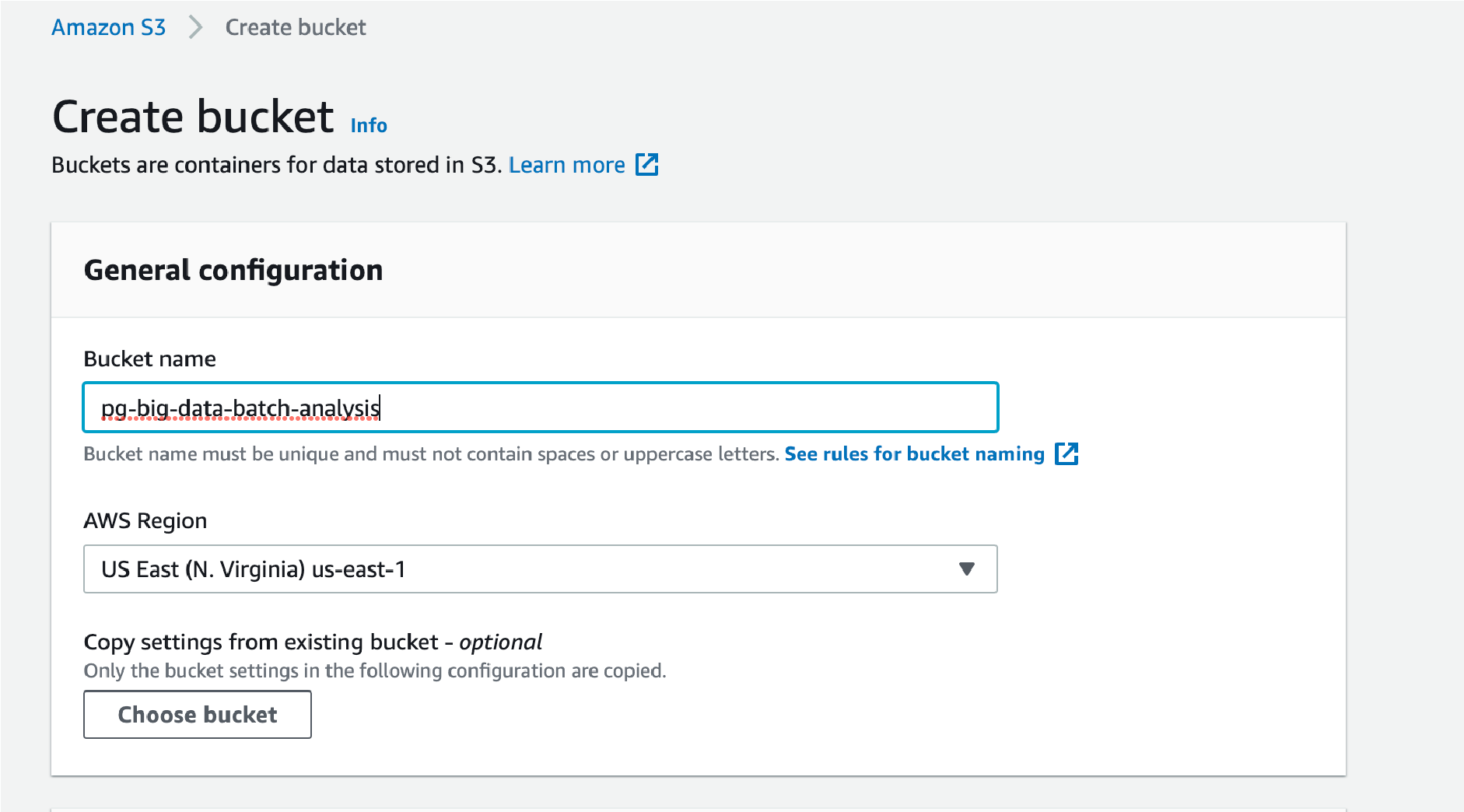
# Solution

## 1. 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)

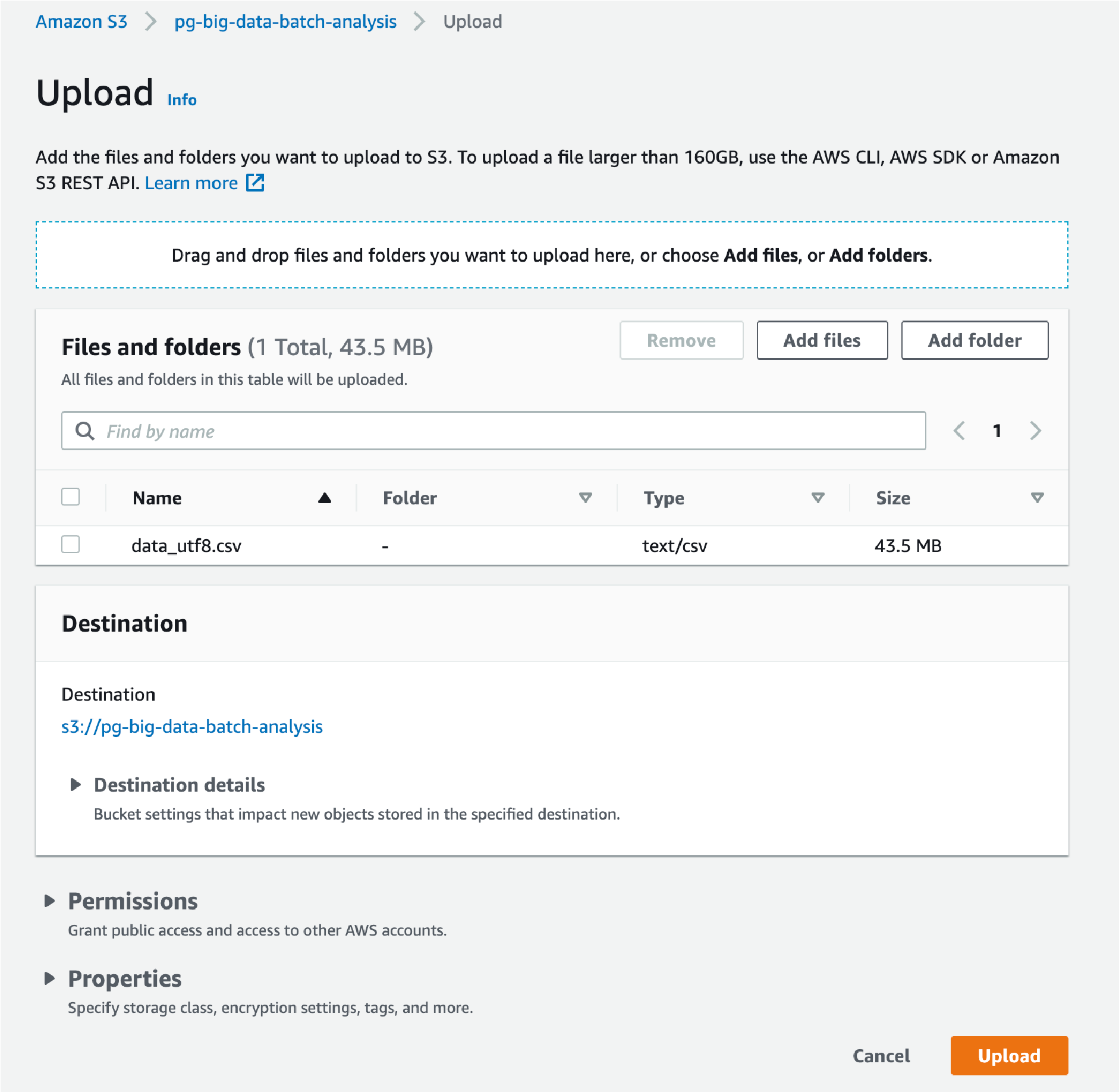
Sign into the AWS Management Console and open the AWS S3 console. The first step is to create a new S3 bucket to store the CSV file, we can achieve that through hitting the Create bucket button.

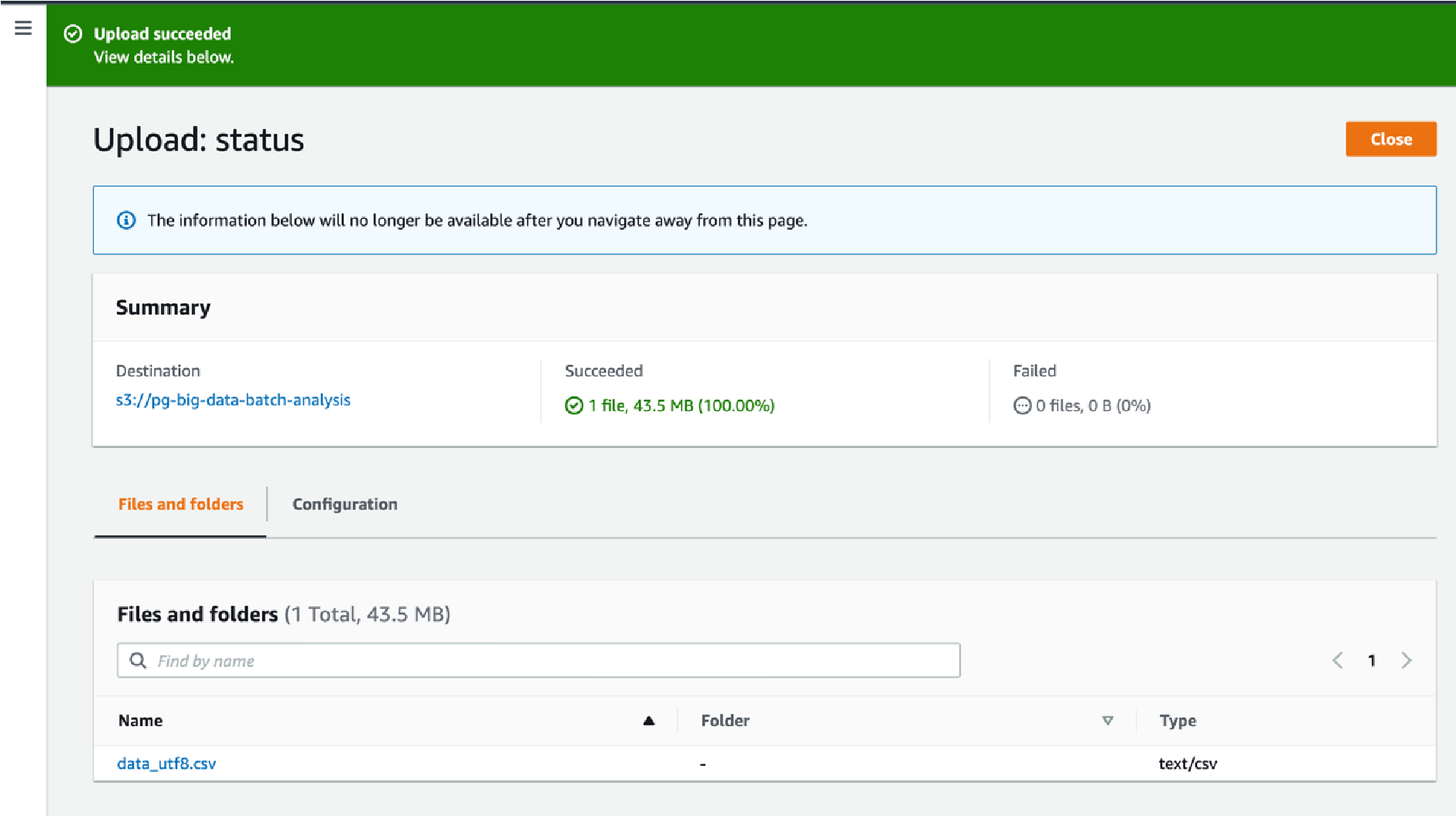


When we create a bucket, we need to give it a name:



After we have created the bucket successfully, we can now upload our CSV file in there:

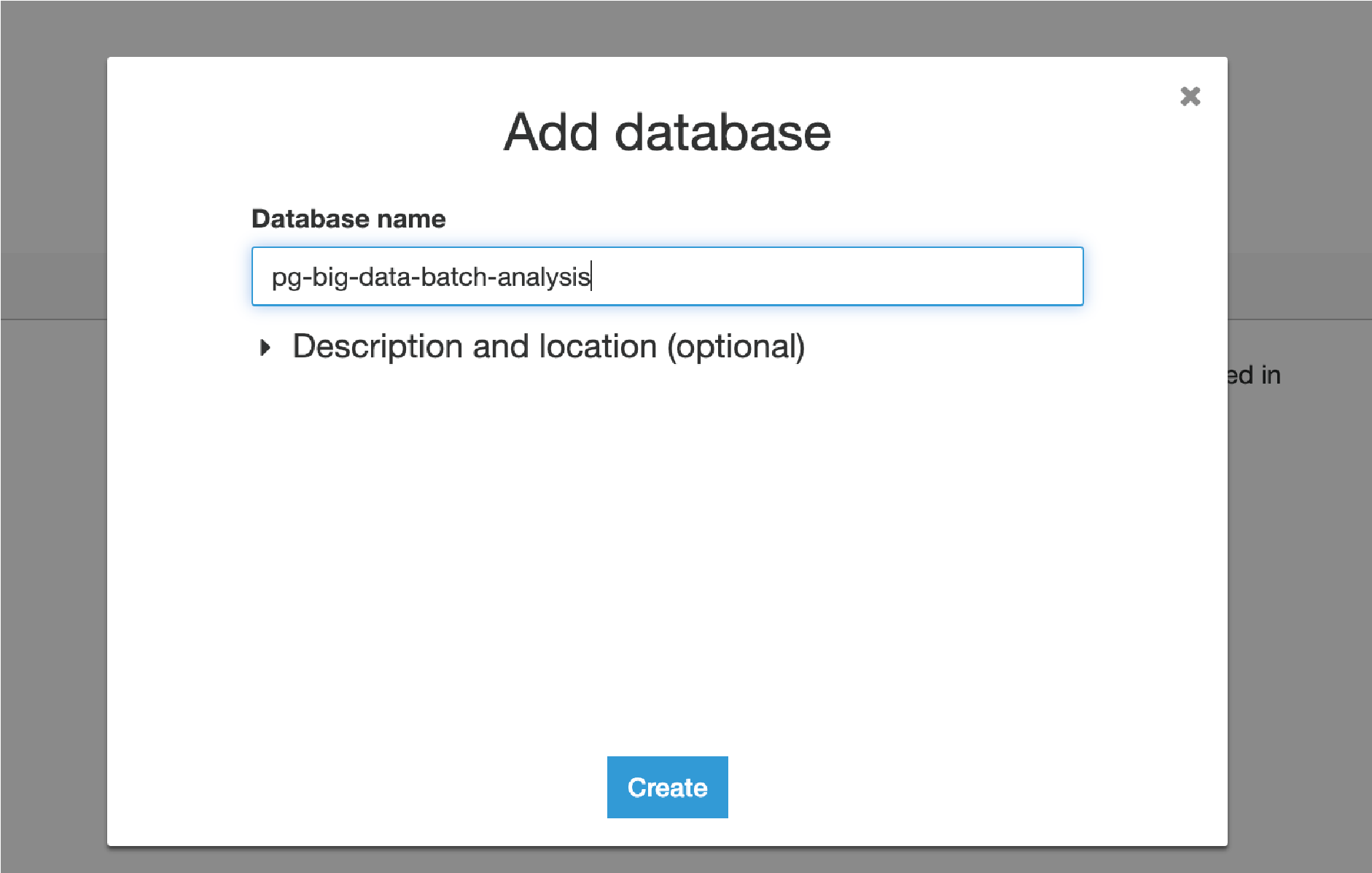




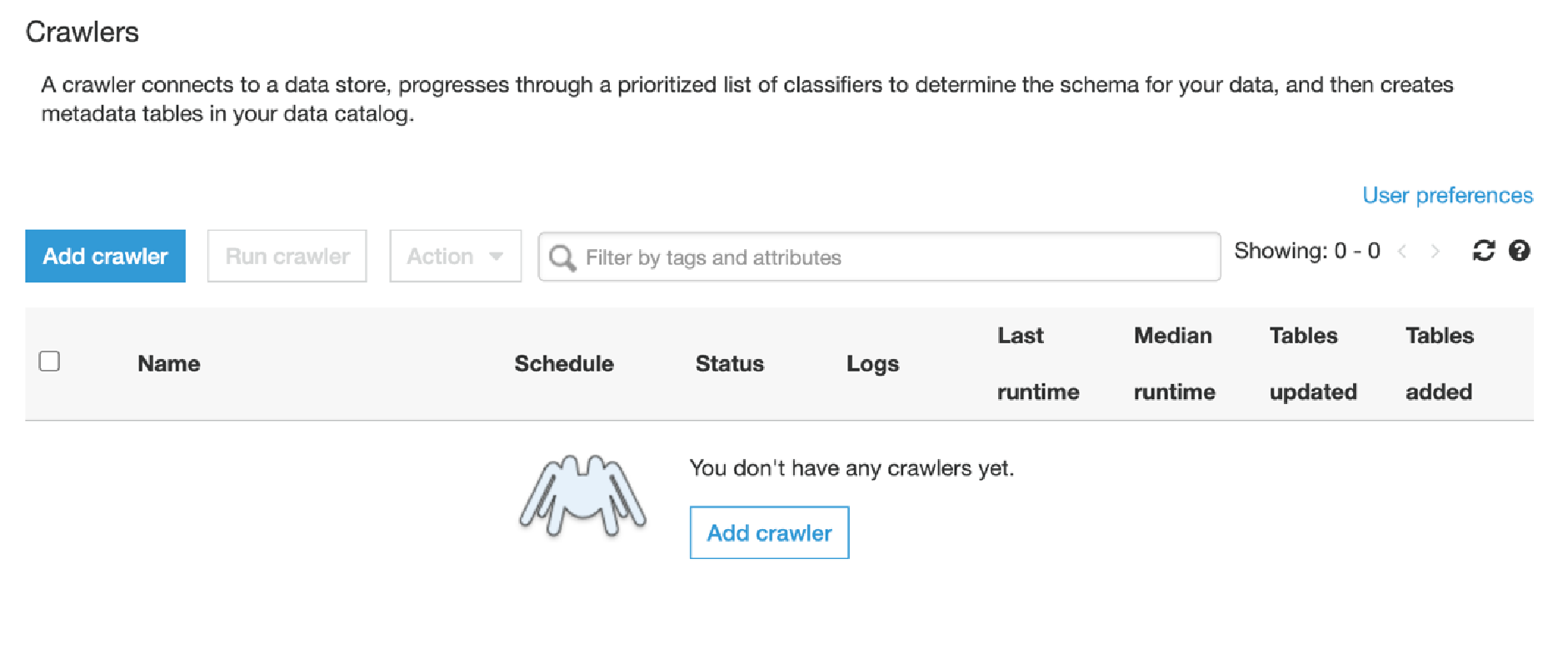
## 2. Create a crawler to crawl the CSV data and generate a metadata catalog

Sign into the AWS Management Console and open the AWS Glue console. The first step to discovering the data is to add a database. A database is a collection of tables.

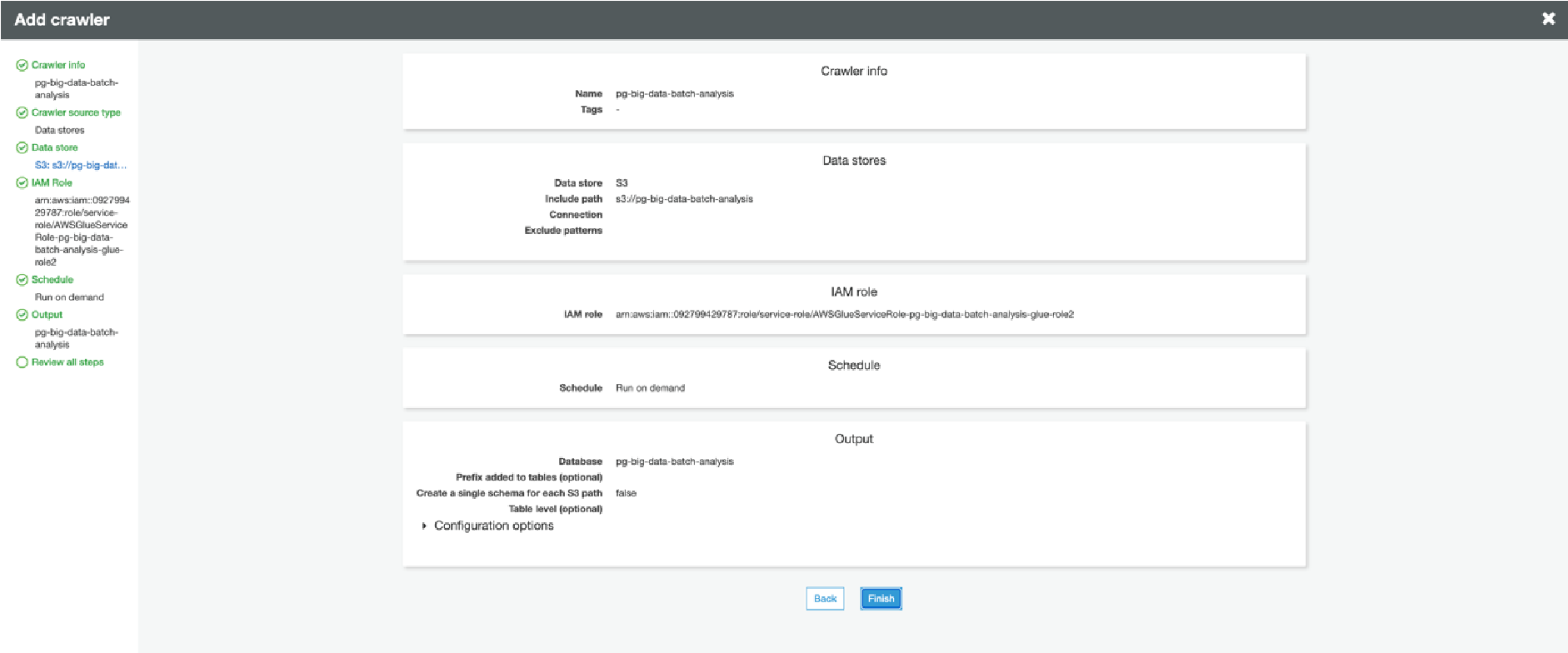
In the console, choose Add database. In Database name, type pg-big-data-batch-analysis, and choose Create.



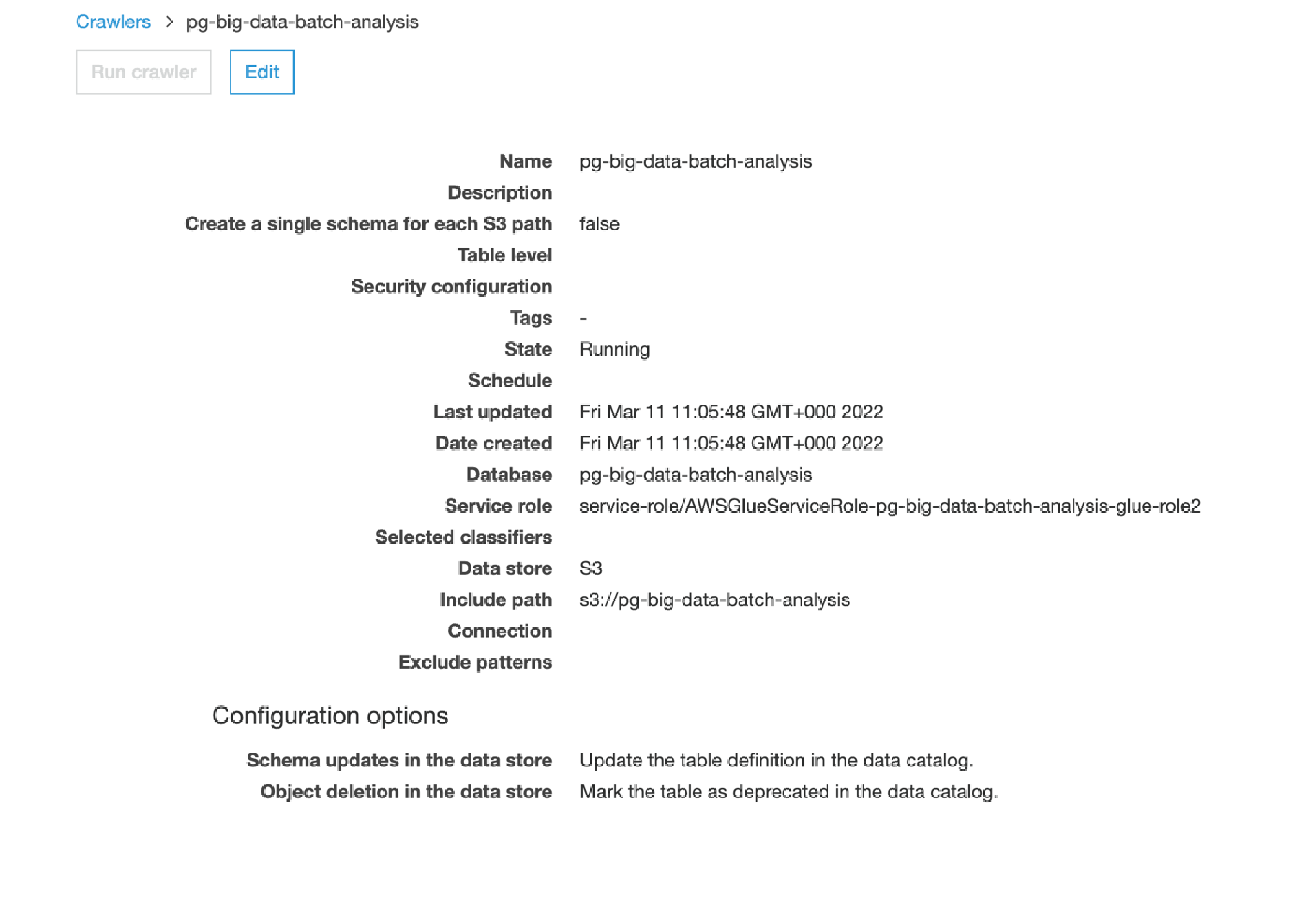
Add a table to the database pg-big-data-batch-analysis. We can add a table manually or by using a crawler. A crawler is a program that connects to a data store and progresses through a prioritized list of classifiers to determine the schema for your data. AWS Glue provides classifiers for common file types like CSV, JSON, Avro, and others. You can also write your own classifier using a grok pattern.



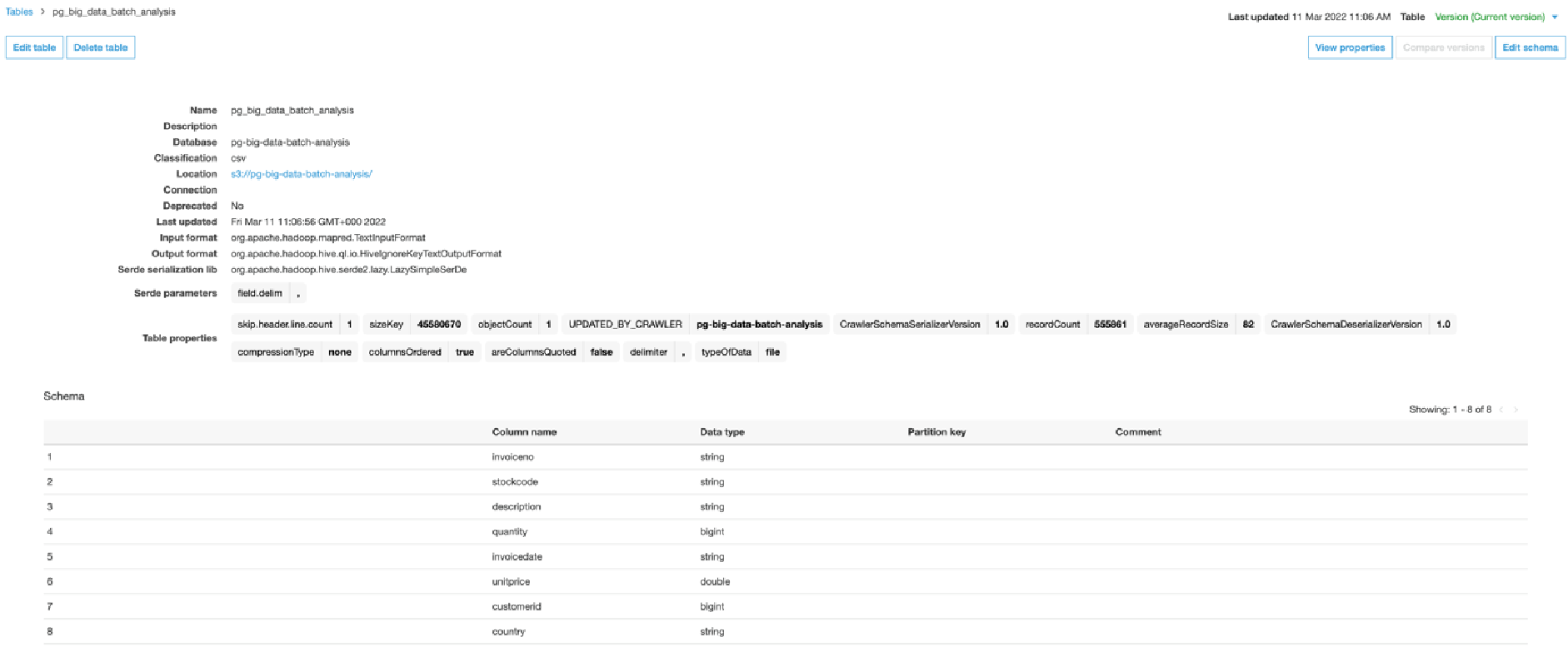
To add a crawler, enter the data source: an Amazon S3 bucket named s3://pg-big-databatch-analysis /. This S3 bucket contains the CSV data file. For IAM role, we created a new role. For Frequency, choose Run on demand. The crawler can be run on demand or set to run on a schedule. For Database, choose pg-big-data-batch-analysis, the one we created in the previous steps.



Review the steps and choose Finish. The crawler is ready to run. Choose Run it now.

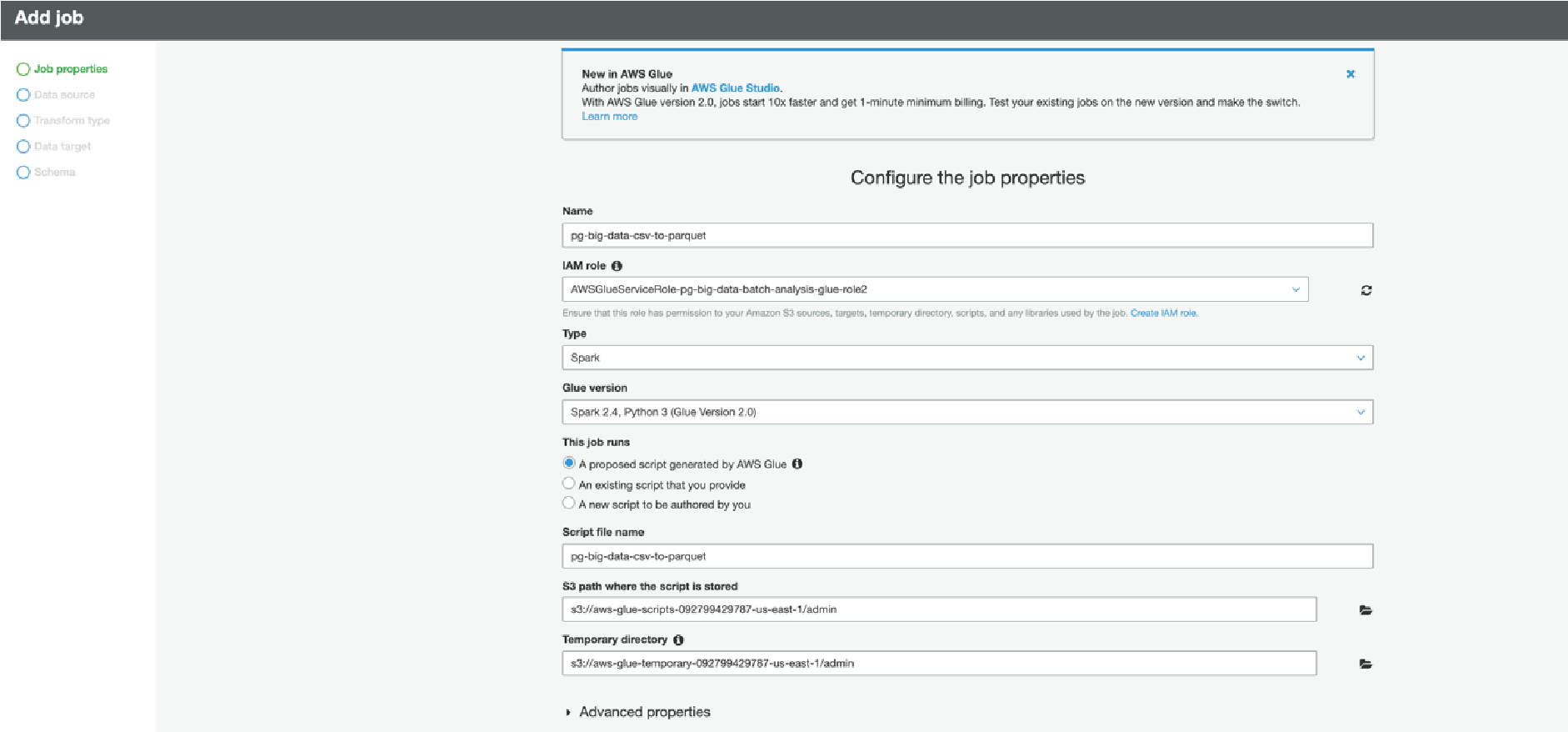


When the crawler has finished, one table has been added. Choose Tables in the left navigation pane, and then choose data. This screen describes the table, including schema, properties, and other valuable information.

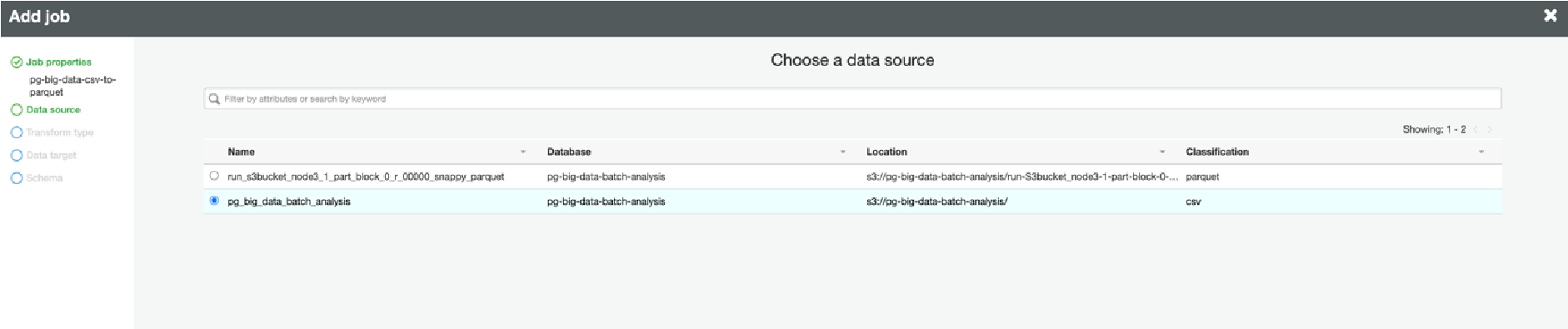


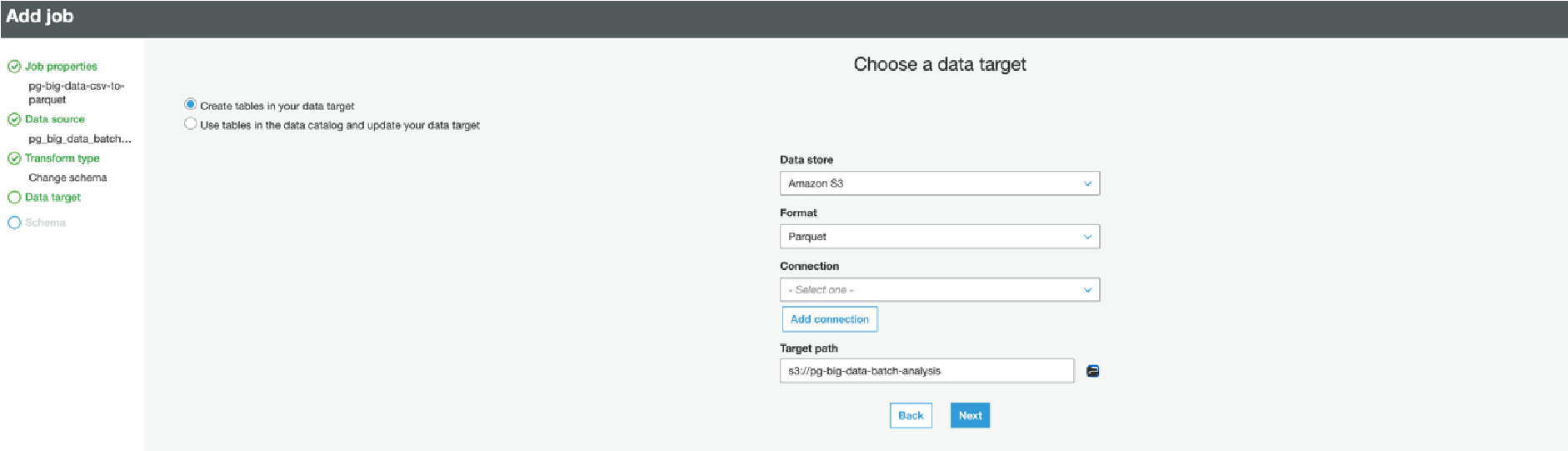
## 3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

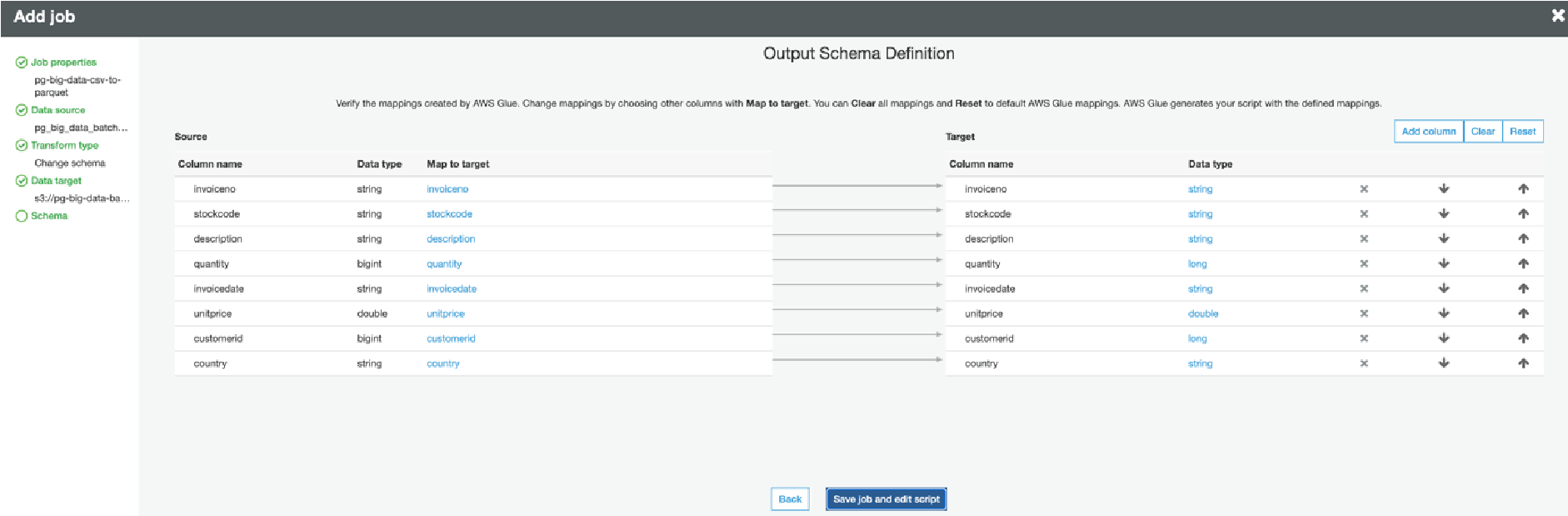
Under ETL in the left navigation pane, choose Jobs, and then choose Add job.



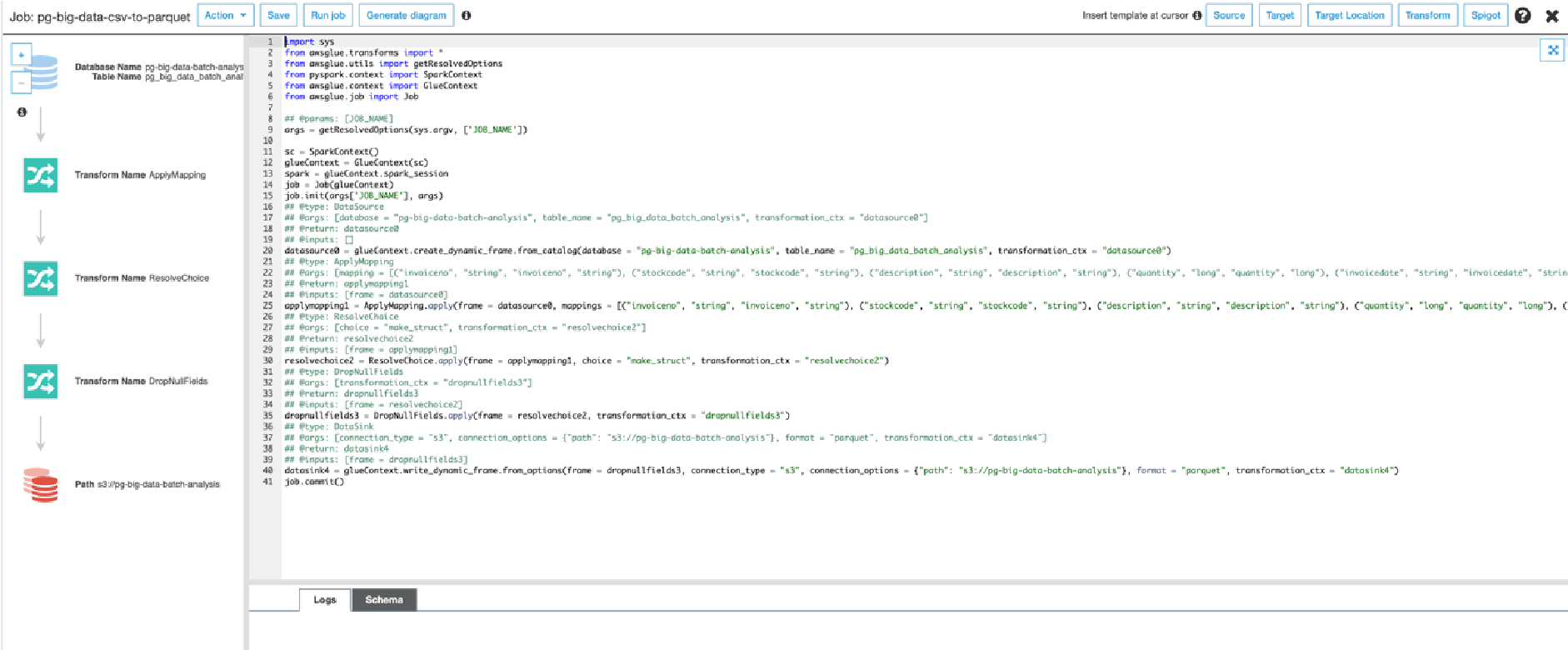
We specify the source S3 bucket which contains our CSV file and, in the target, we specify another S3 bucket where we want our Parquet file to land, in order to transform from CSV to Parquet we need to specify the format type of Parquet for our target.





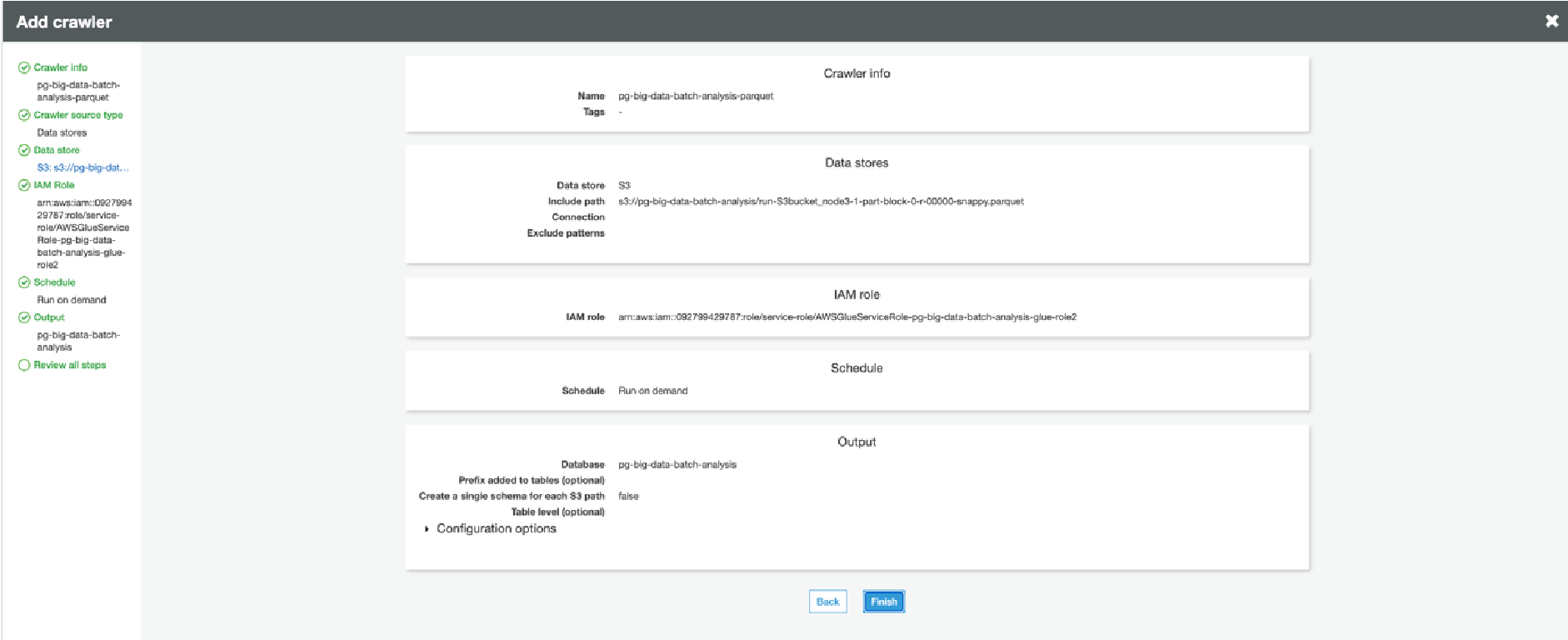


In the script we can see, edit and save the source code of the Glue Job Here we run the job to start the CSV to Parquet conversion



## 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

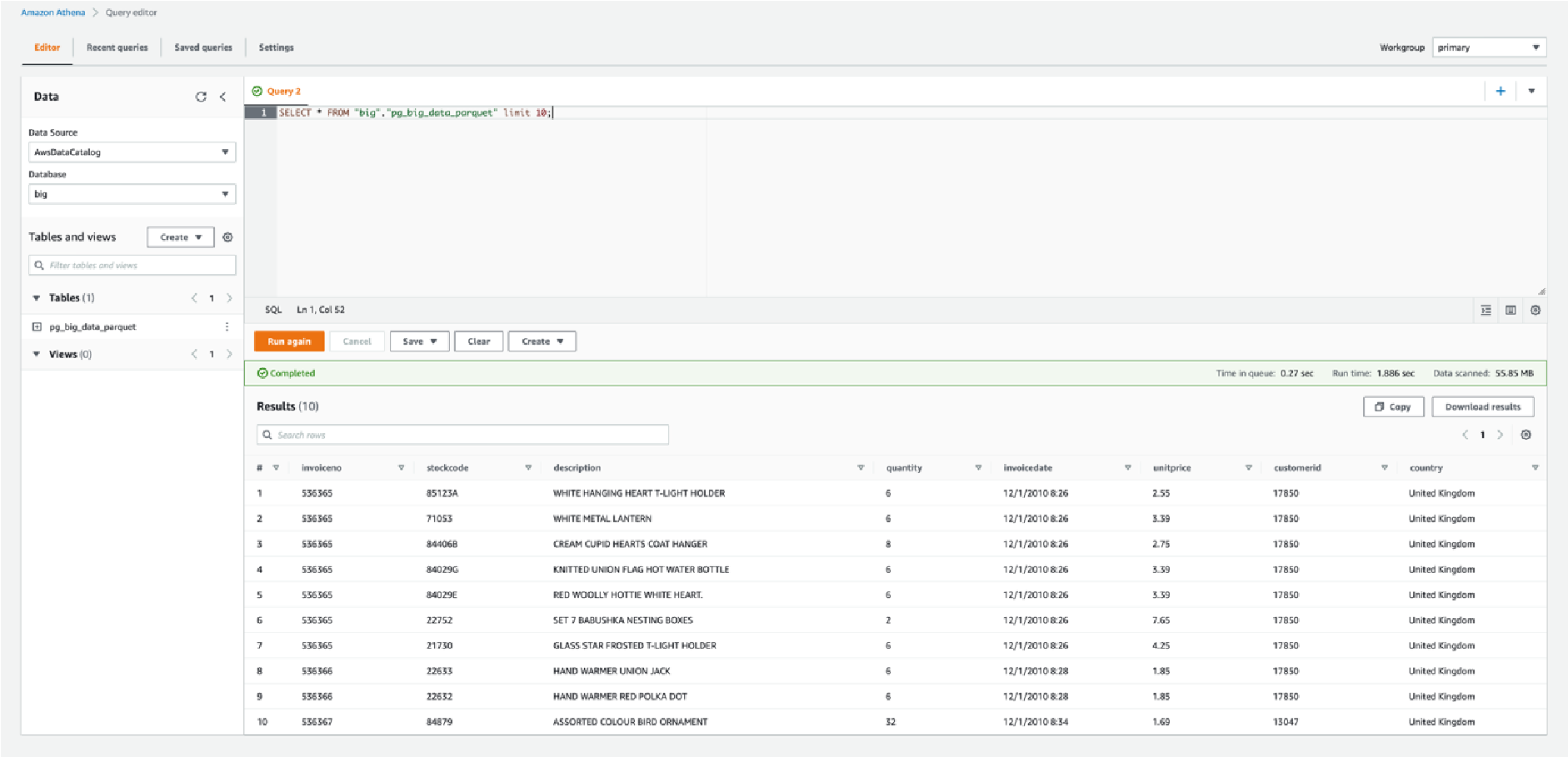
When the job has finished, add a new table for the Parquet data using a crawler.



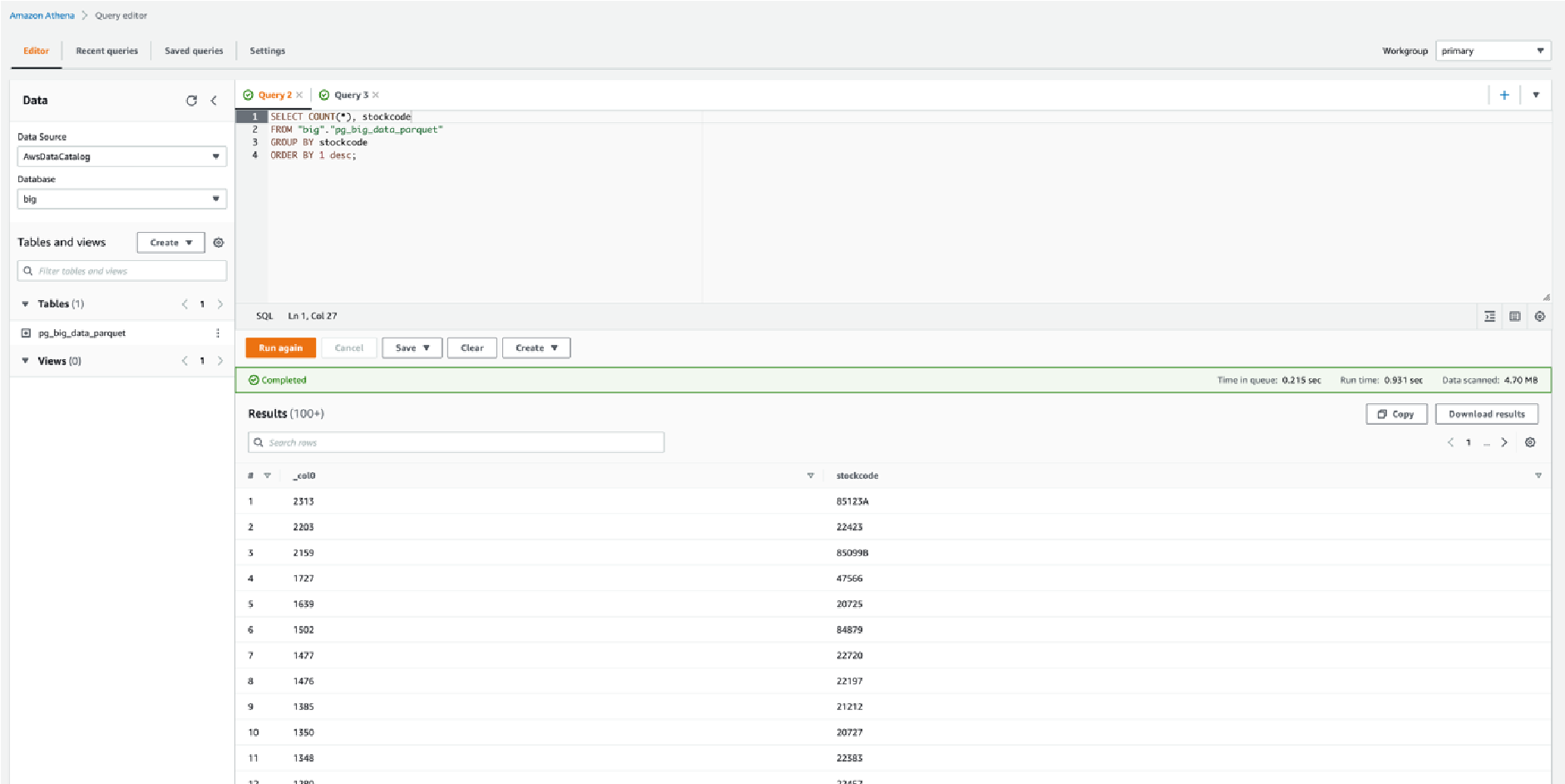
## 5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

Firstly, we query to see if Athena can retrieve the data correctly and that we can query our

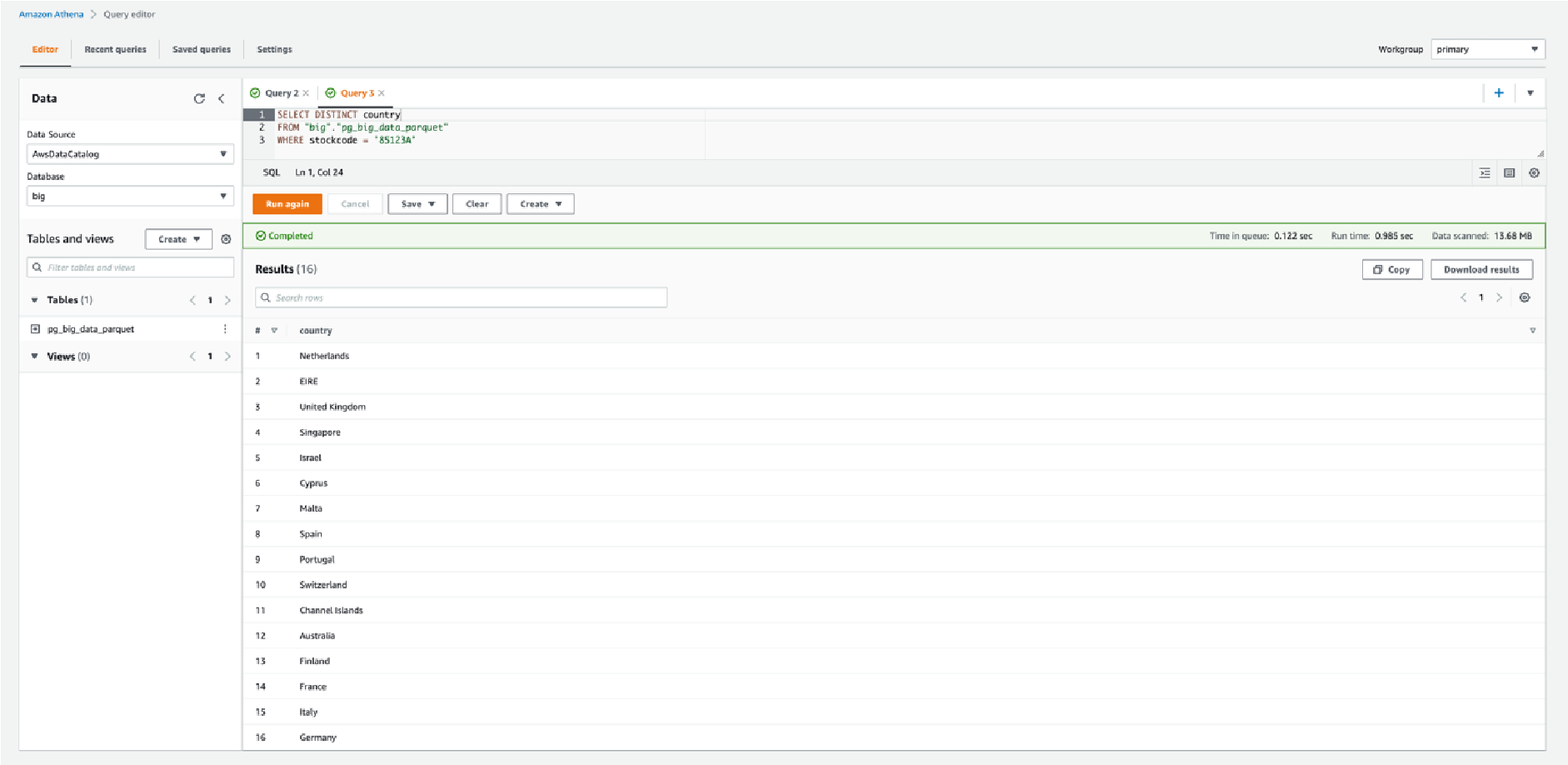
Parquet file



After we have checked that Athena is able to retrieve records, we are going to check which is the most sold item in the file:



We have got back that the item with the Stock Code 85123A is the most sold one, now we can check in which countries this item has been sold:



We got back a list of 16 countries out of the total 38 possible countries that appear in all the data.