Step 1: Create S3 Buckets

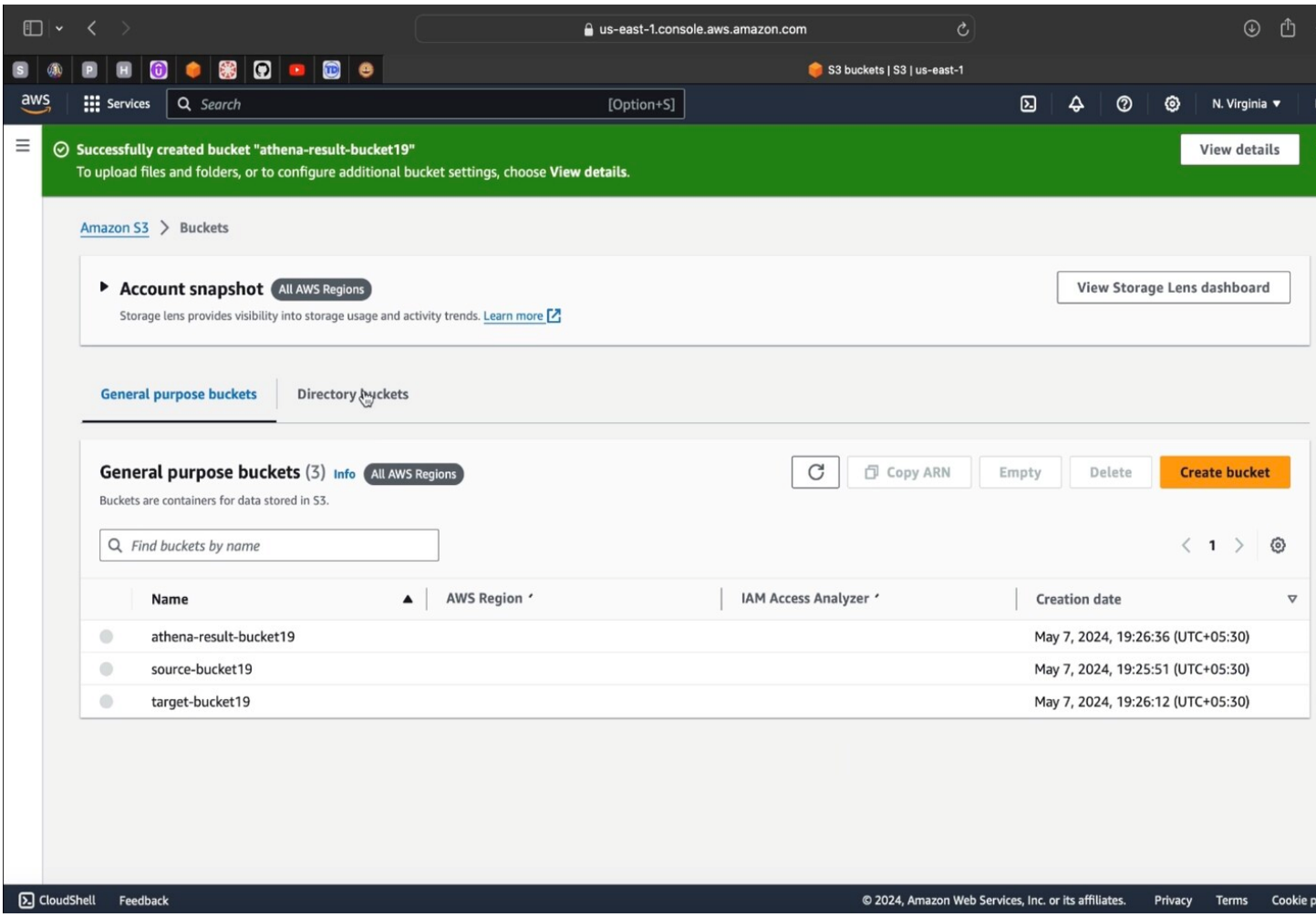
1. Create Source Bucket:

- Click “Create bucket” and follow the prompts to create a new S3 bucket. This will be your source bucket containing the CSV files.

2. Create Target Bucket:

- Similarly, create another S3 bucket that will serve as the target for your transformed data.

3. Create Athena Bucket to Store results :

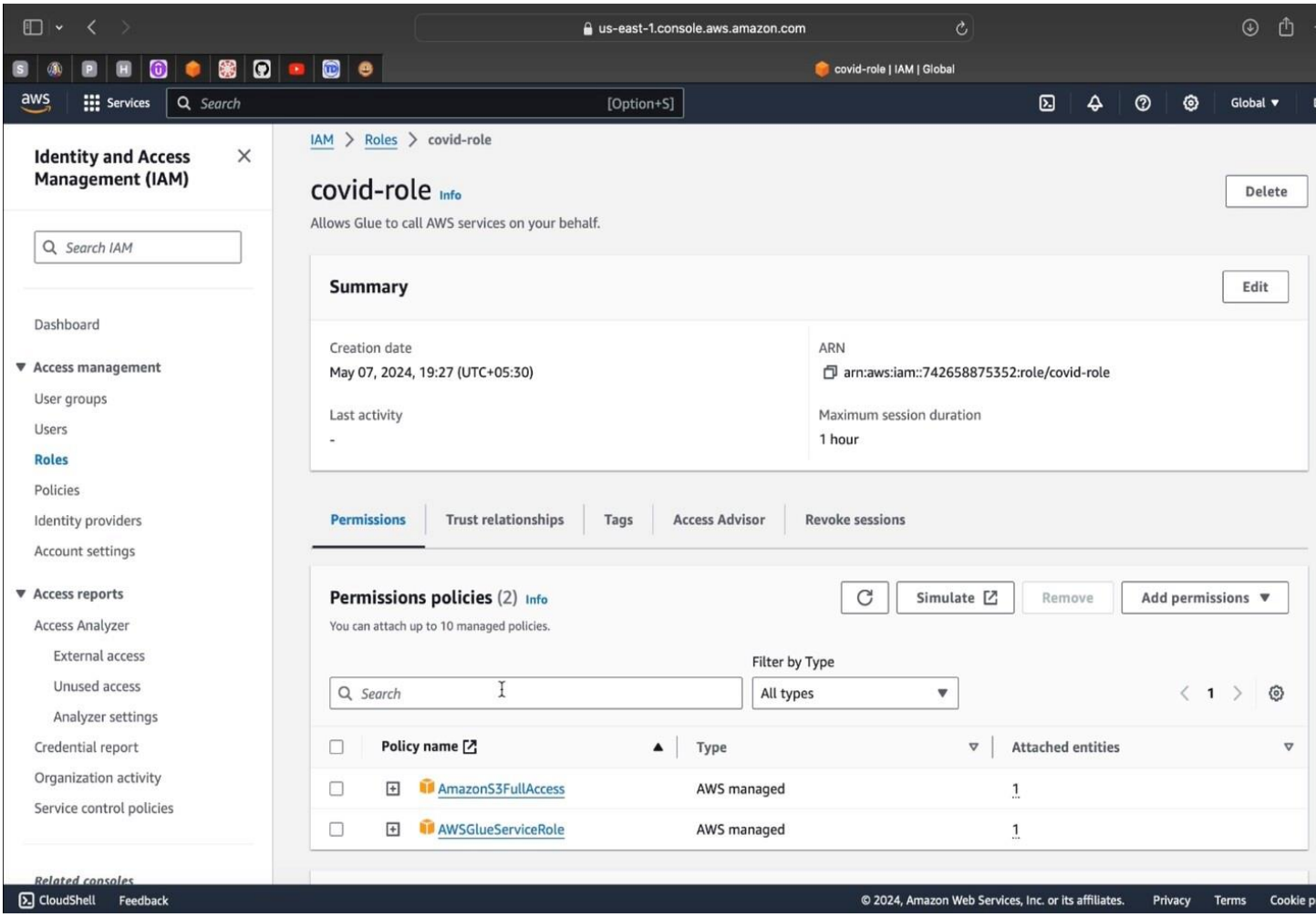


Step 2: Create IAM Role for AWS Glue

1. Choose a use case,” select “Glue.”

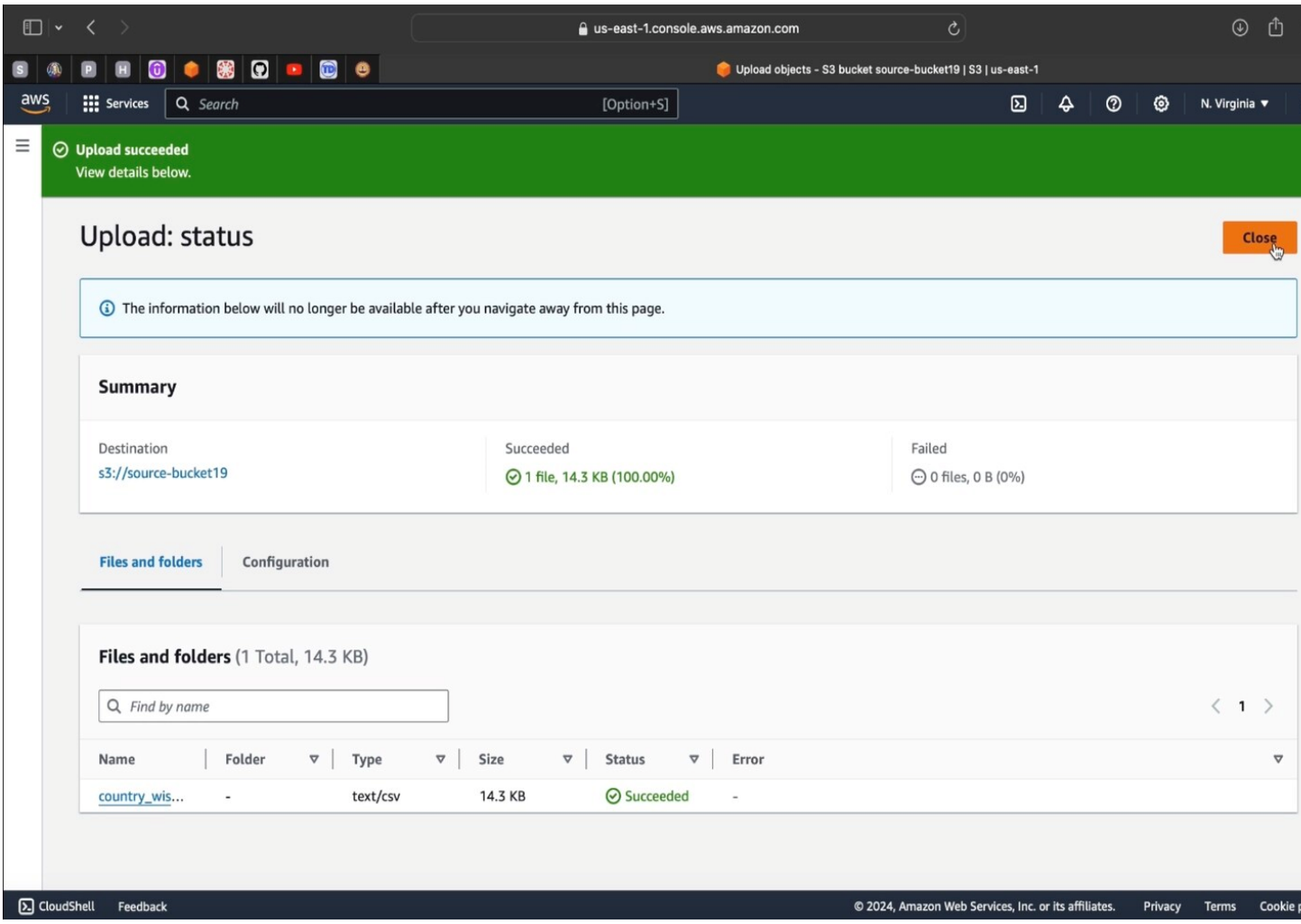
2. Attach Permissions Policies
3. Search and attach the policies:

- Search for and attach the policy AmazonS3FullAccess
- Search for and attach the policy AWSGlueServiceRole



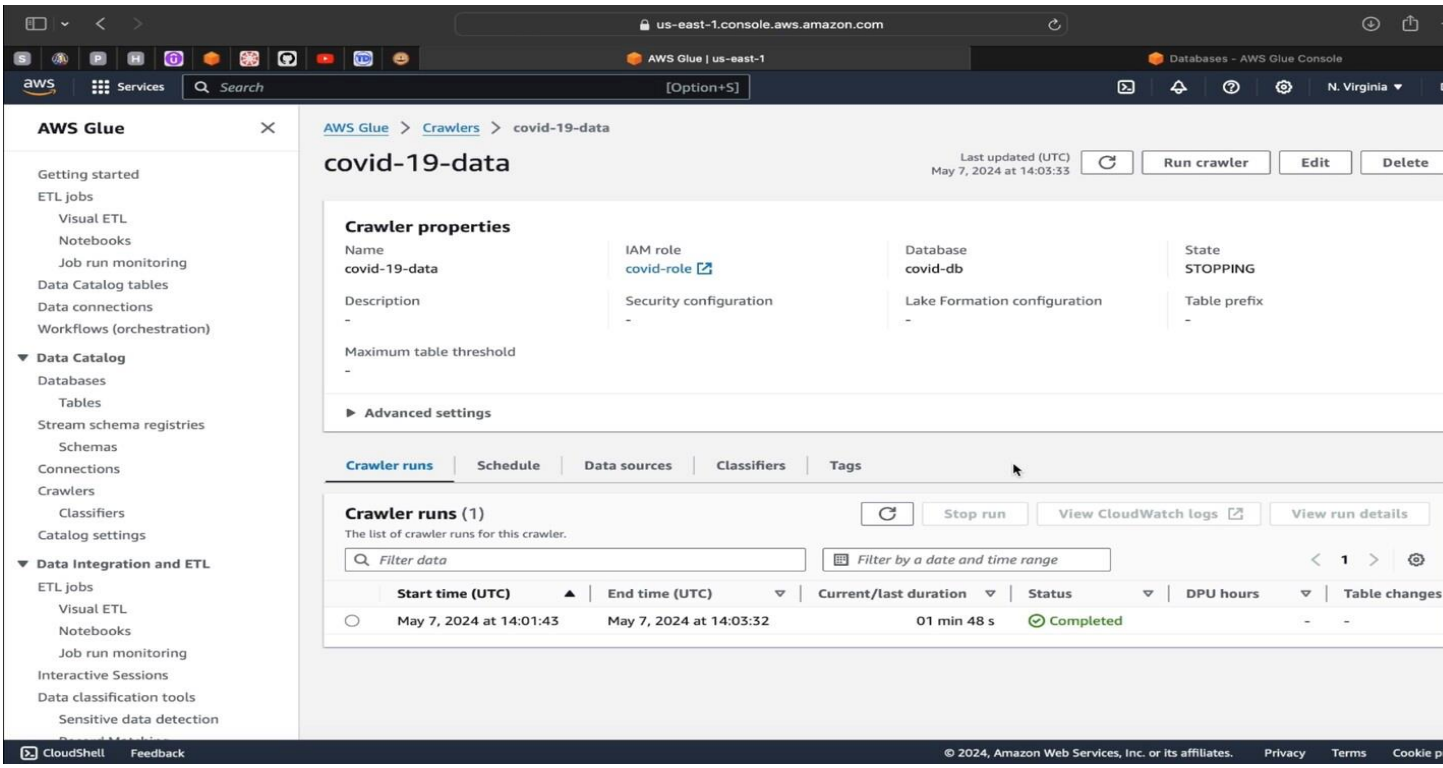
Step 3: Upload CSV Files to Source Bucket

- Upload the CSV files that you want to process into the source S3 bucket.Note(Create a folder inside the Bucket and then upload CSV file inside that folder because sometimes when you tried to run the query from Athena you might be getting RETURNS ZERO RECORDS)



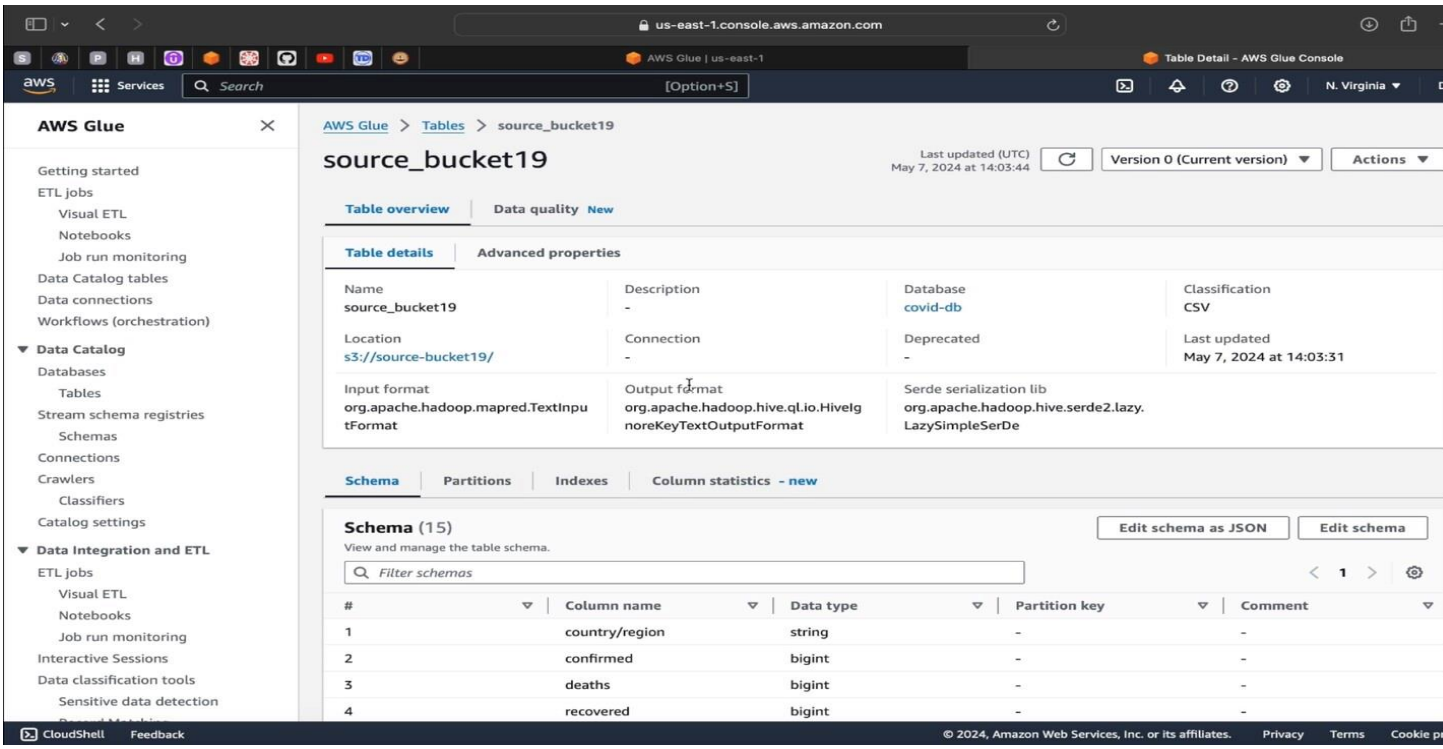
Step 4: Set Up AWS Glue Crawler

1. Go to AWS Glue Console
2. Go to Crawlers -> Create a Crawler
3. Click “Add crawler” and follow the wizard to configure the crawler
4. Specify a name for the crawler.
5. Choose “S3” as the data store.
6. Specify the S3 path to your source bucket.
7. Finish the wizard to create the crawler.



Status is now completed

Again go inside database there will be tables



Step 5: Create an ETL Job in AWS Glue

1. Go to AWS Glue Console:

- Click on “Jobs” in the left sidebar.
- Click “Add job” and follow the wizard to configure the ETL job:
- Specify a name for the job.
- Choose the source and target connections.
- Define the transformation logic using the Glue ETL script.
- Finish the wizard to create the job.

1. Run the ETL Job:

- After creating the job, run it to perform the ETL process on the data.

2. Select Source bucket — Amazon S3 with customer — data catalog table

3.Select Target bucket — Select field

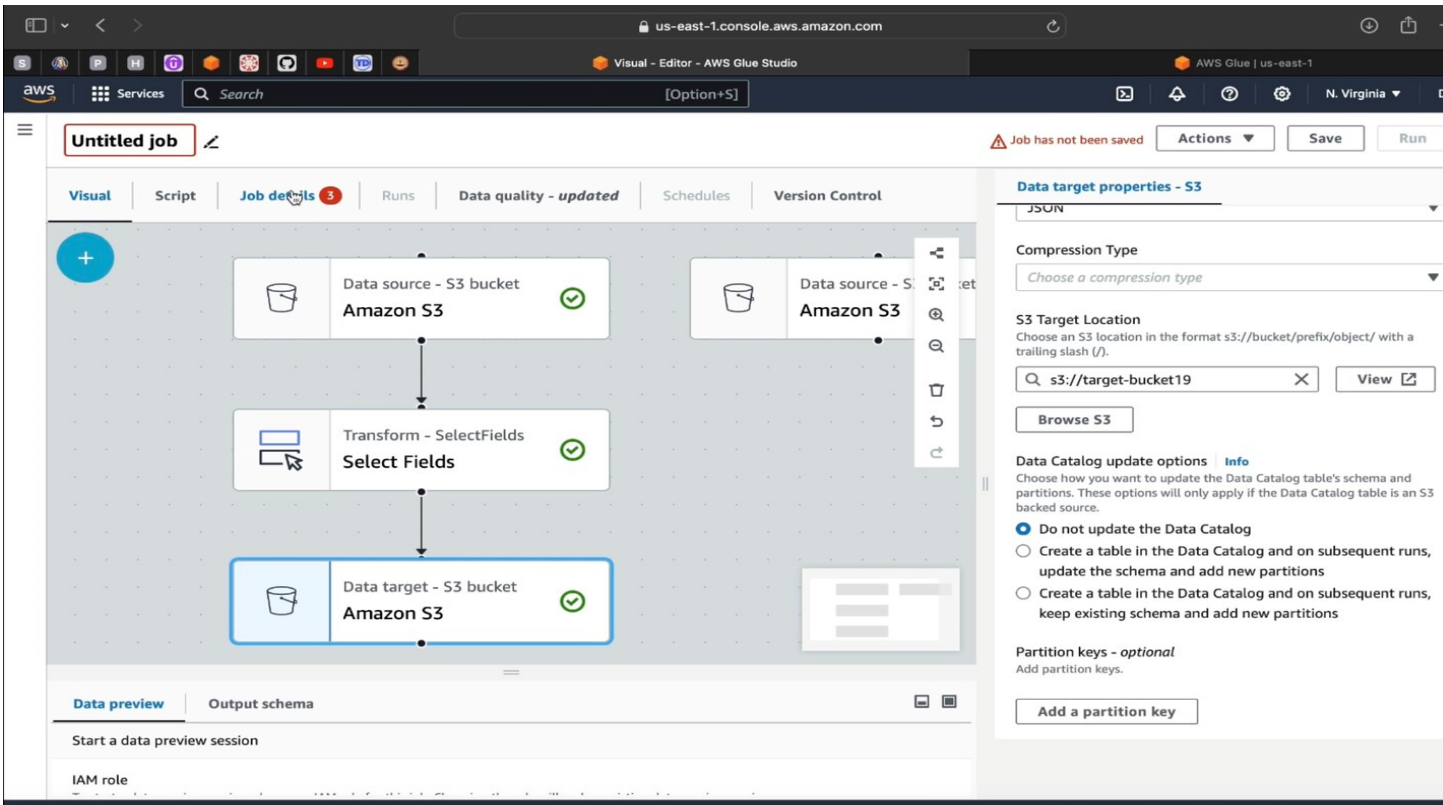
4.Select Transform bucket — Amazon S3 with customer id

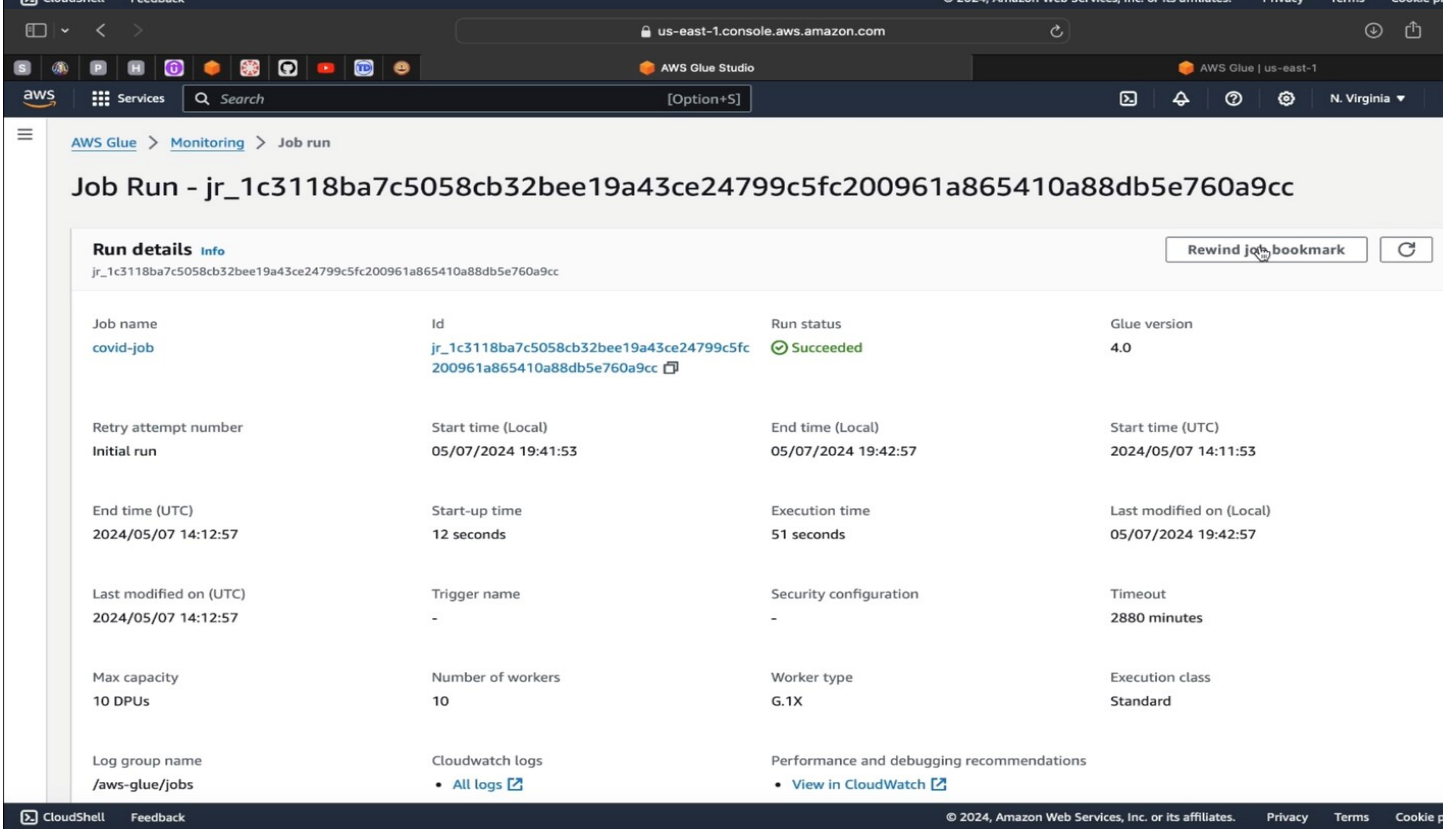
5.Before Run Goto Job details

6.Set the Name of the job and also select the IAM role -> Then save

7.-> Run

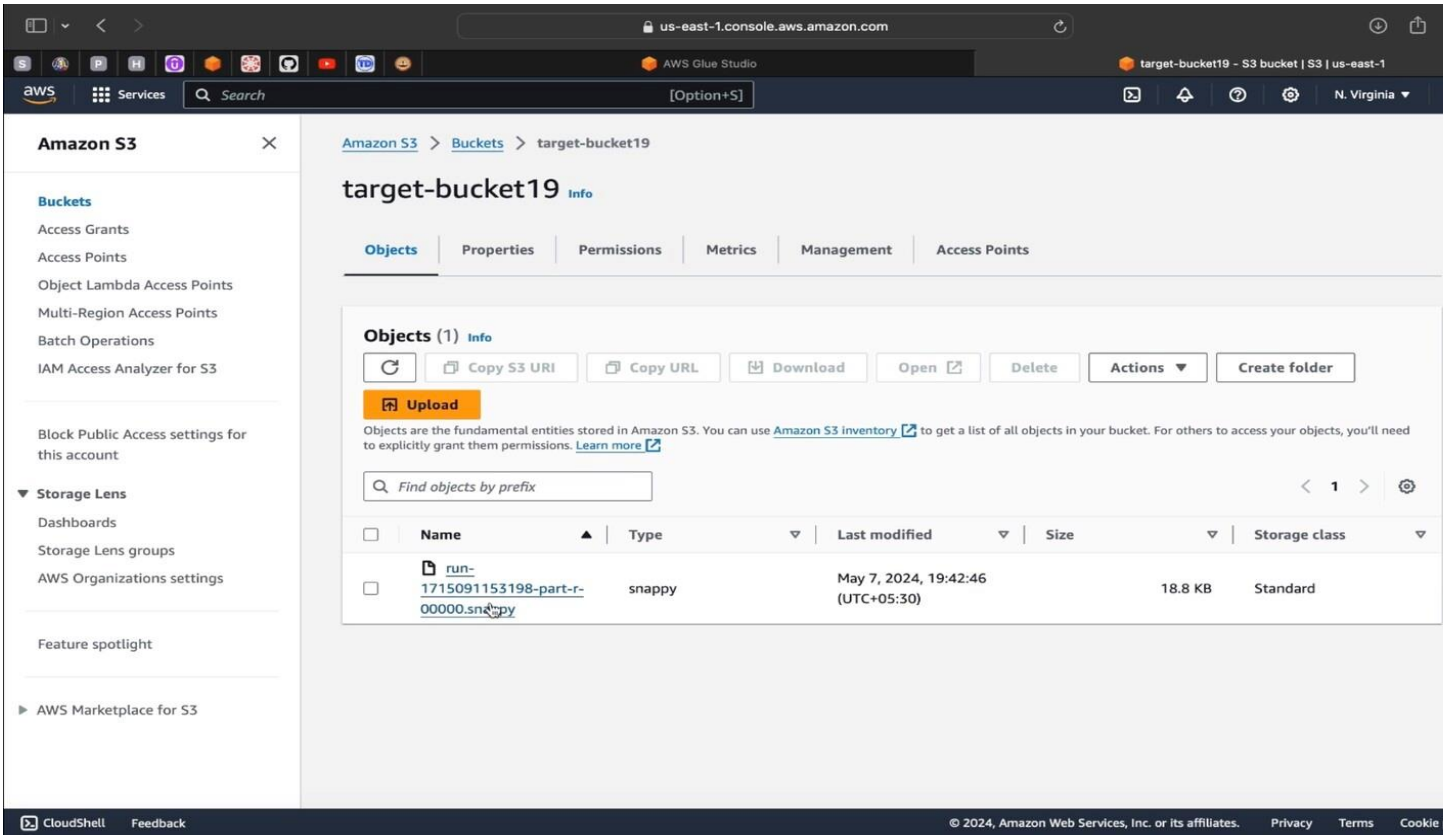
8.Goto Runs Section -> once status succeeded


us-east-1.console.aws.amazon.com
Visual - Editor - AWS Glue Studio
AWS Glue | us-east-1
Services
Search
[Option+S]
N. Virginia
Untitled job
Job has not been saved
Actions
Save
Run
Visual
Script
Job details 3
Runs
Data quality - updated
Schedules
Version Control
Data source - S3 bucket
Amazon S3
Transform - SelectFields
Select Fields
Data target - S3 bucket
Amazon S3
Data target properties - S3
JSON
Compression Type
Choose a compression type
S3 Target Location
Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://target-bucket19
View
Browse S3
Data Catalog update options Info
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
Do not update the Data Catalog
Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions
Partition keys - optional
Add partition keys.
Add a partition key
Data preview
Output schema
Start a data preview session
IAM role
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms


us-east-1.console.aws.amazon.com
AWS Glue Studio
AWS Glue | us-east-1
Services
Search
[Option+S]
N. Virginia
AWS Glue > Monitoring > Job run
Job Run - jr_1c3118ba7c5058cb32bee19a43ce24799c5fc200961a865410a88db5e760a9cc
Run details Info
jr_1c3118ba7c5058cb32bee19a43ce24799c5fc200961a865410a88db5e760a9cc
Rewind job bookmark
Job name
covid-job
Id
jr_1c3118ba7c5058cb32bee19a43ce24799c5fc200961a865410a88db5e760a9cc
Run status
Succeeded
Glue version
4.0
Retry attempt number
Initial run
Start time (Local)
05/07/2024 19:41:53
End time (Local)
05/07/2024 19:42:57
Start time (UTC)
2024/05/07 14:11:53
End time (UTC)
2024/05/07 14:12:57
Start-up time
12 seconds
Execution time
51 seconds
Last modified on (Local)
05/07/2024 19:42:57
Last modified on (UTC)
2024/05/07 14:12:57
Trigger name
-
Security configuration
-
Timeout
2880 minutes
Max capacity
10 DPUs
Number of workers
10
Worker type
G.1X
Execution class
Standard
Log group name
/aws-glue/jobs
Cloudwatch logs
All logs
Performance and debugging recommendations
View in CloudWatch
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms

Step 6: Check Transformed Data in Target Bucket

- Verify that the transformed data is successfully loaded into the target S3 bucket.
- Goto yout target Bucket check your file will automatically added.



Step 7: Set Up Amazon Athena

Now you can directly Query here also using Athena

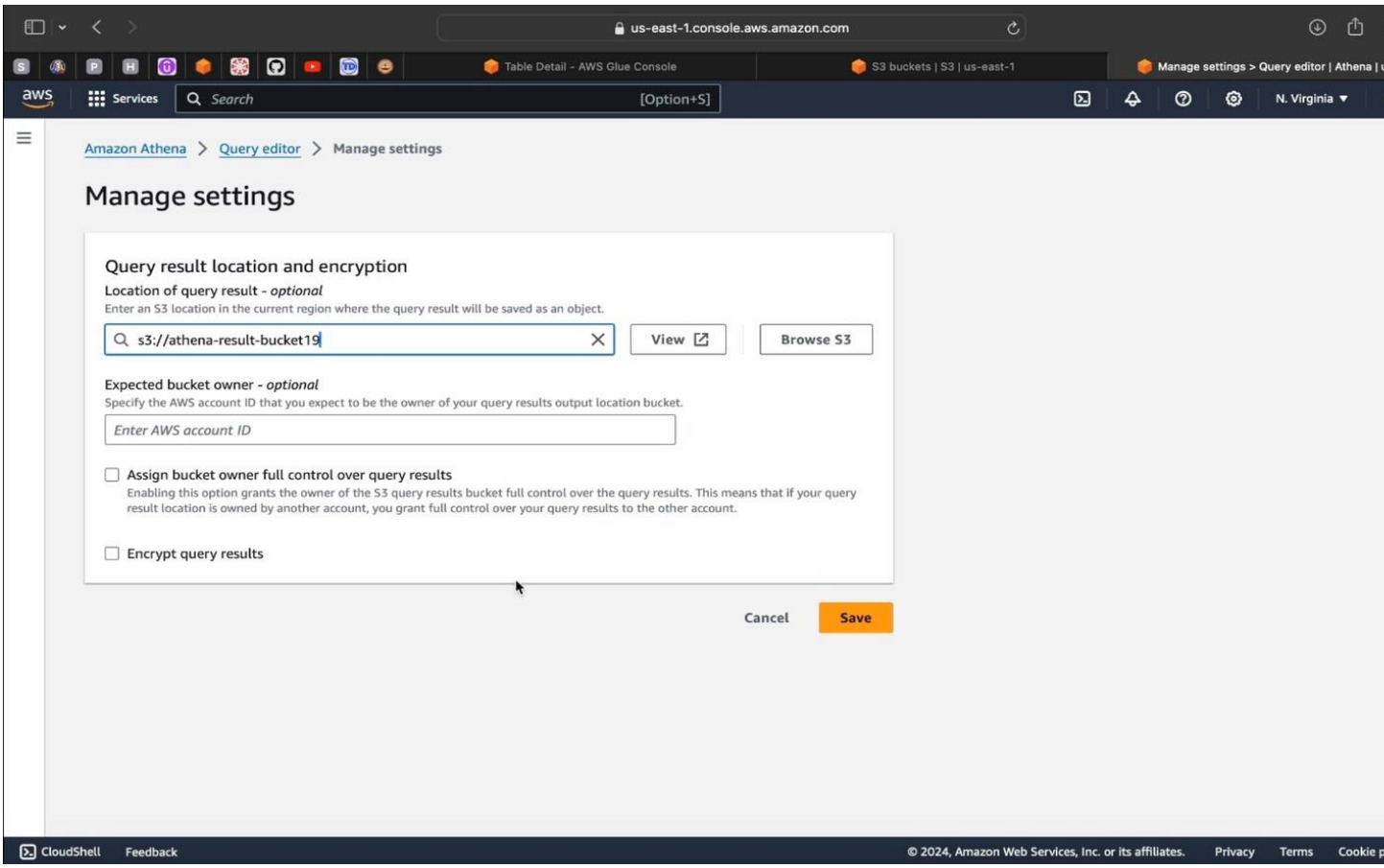
Action -> View data

Initially Query will not run so,

Goto Setting -> Manage

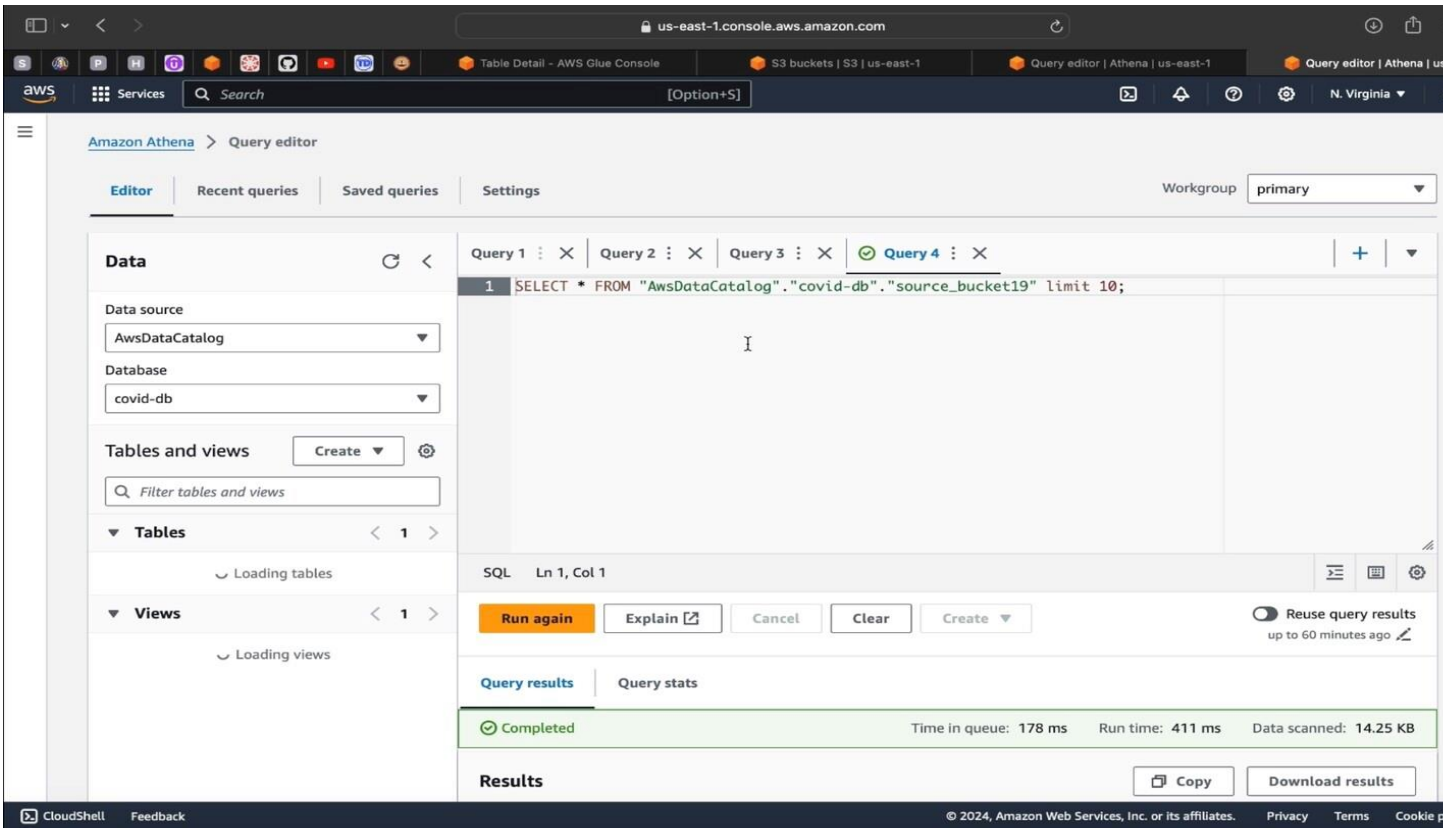
-> Browse S3 Bucket(Select the Athena Bucket)

-> Save

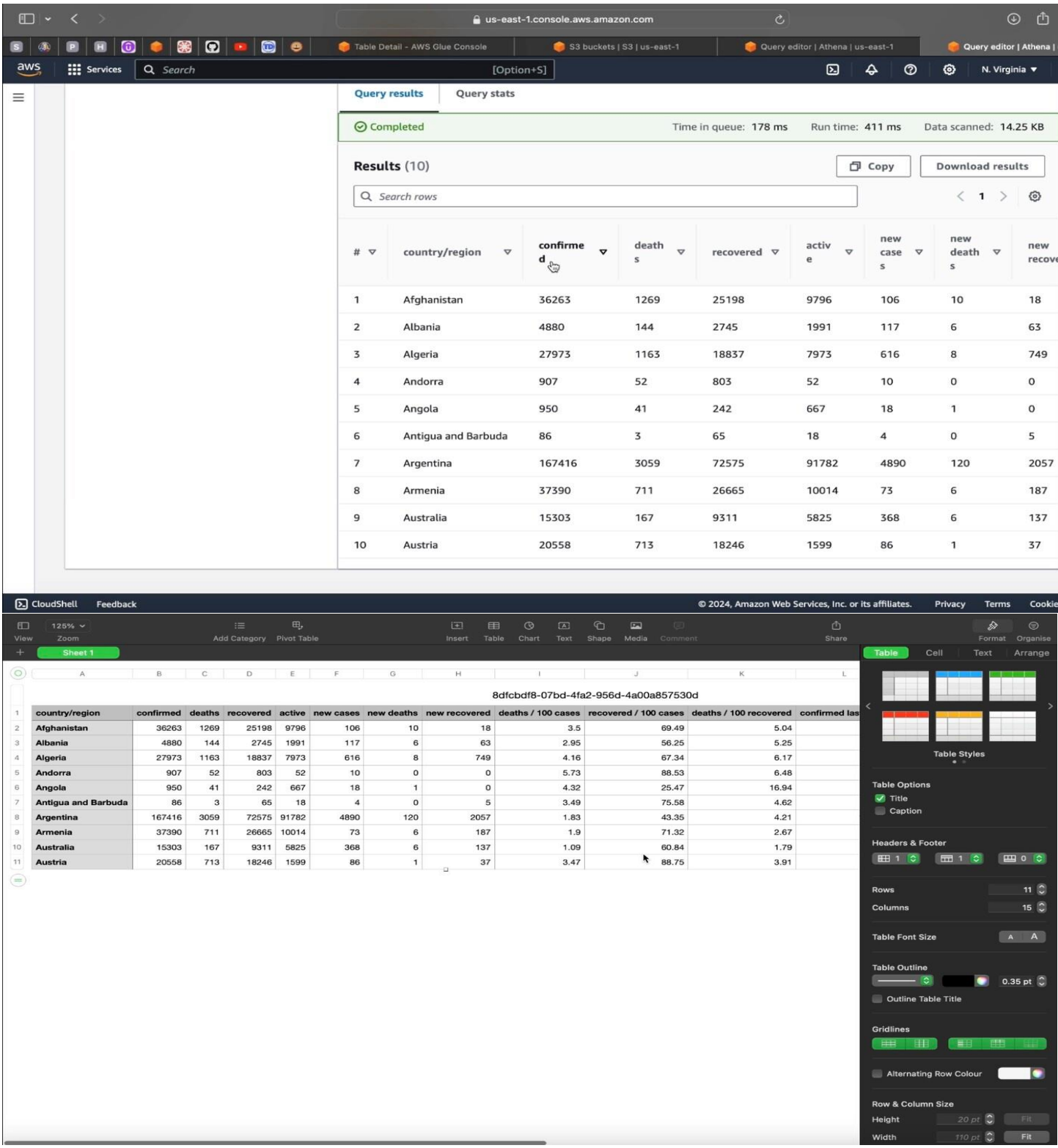


Step 8: Run Queries in Athena

- Use the Athena query editor to run SQL queries on your cataloged tables and analyze the transformed data.



- As per your requirement give the query it will show you the tables.



In this guide, we've traversed the intricacies of constructing a seamless data transformation process, from sourcing CSV files in an S3 bucket through Glue Crawlers, to transforming and loading the data into a designated S3 target bucket.