

## Step 1: Create S3 Buckets

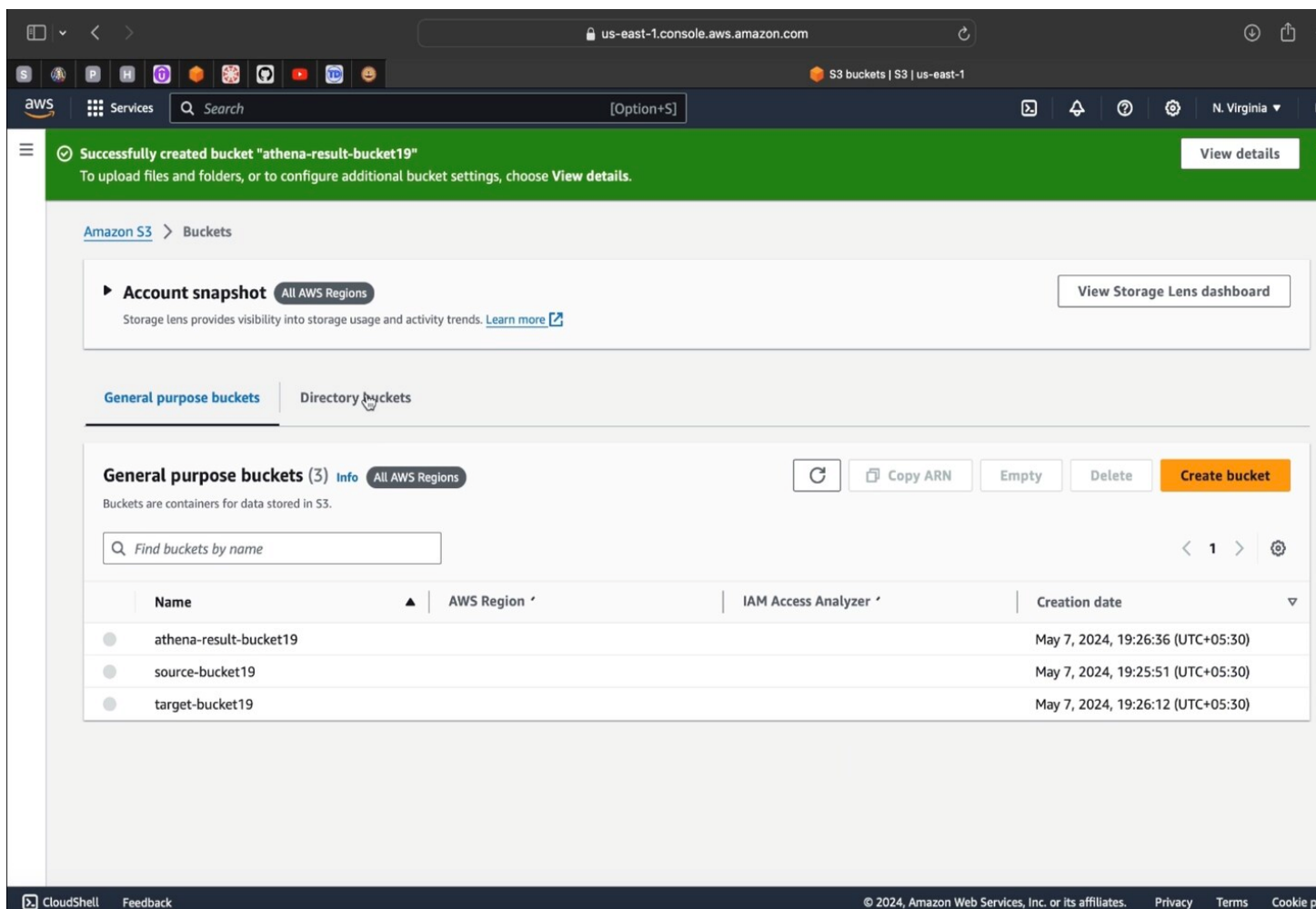
### 1. Create Source Bucket:

- 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

- Choose a use case,” select “Glue.”

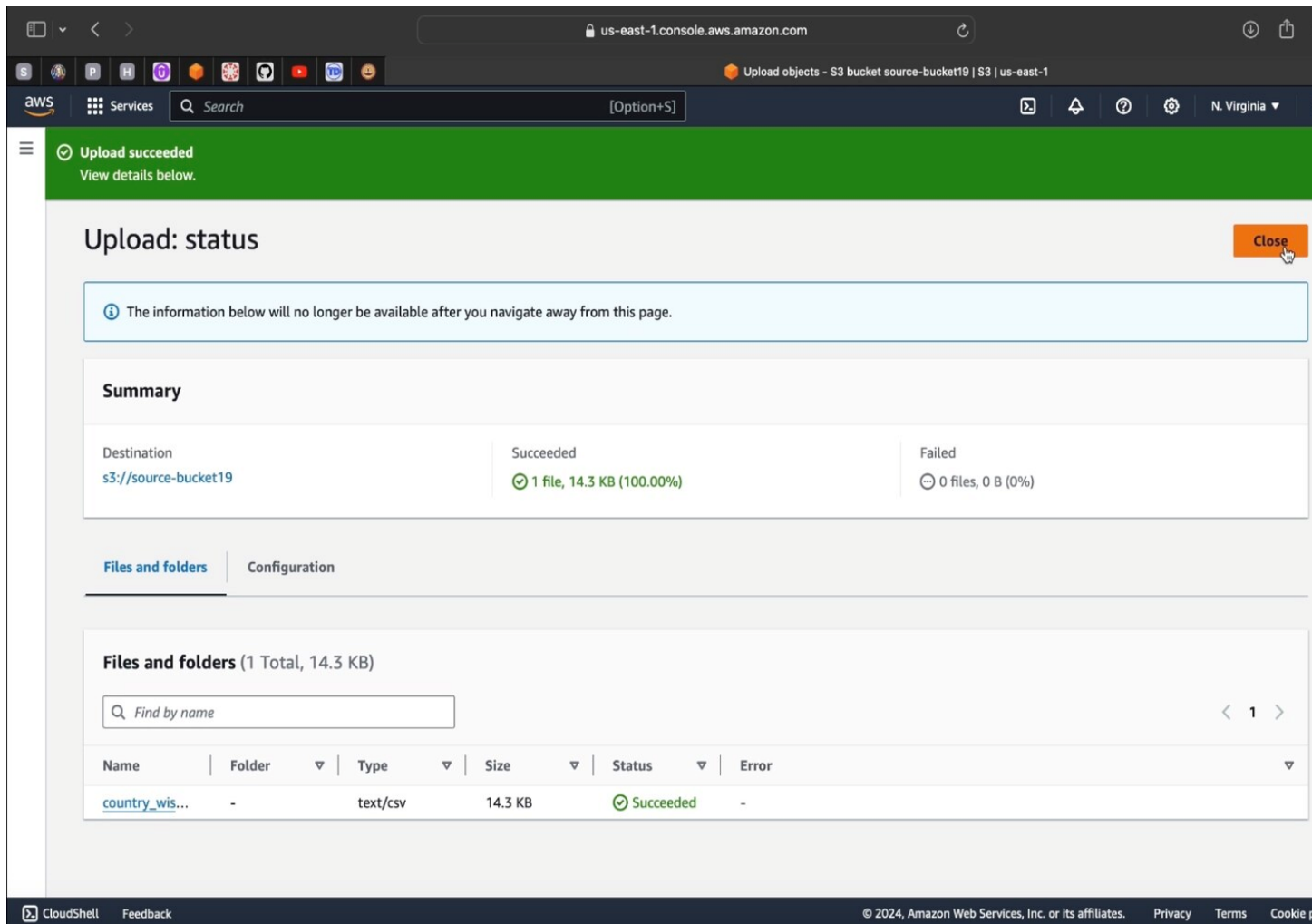
2. Attach Permissions Policies
3. Search and attach the policies:
  - Search for and attach the policy AmazonS3FullAccess
  - Search for and attach the policy AWSGlueServiceRole

The screenshot displays the AWS IAM console for a role named 'covid-role'. The role's summary indicates it allows Glue to call AWS services. Under the 'Permissions policies' tab, two policies are listed:

Policy name	Type	Attached entities
AmazonS3FullAccess	AWS managed	1
AWSGlueServiceRole	AWS managed	1

## 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.

us-east-1.console.aws.amazon.com

AWS Glue | us-east-1

Databases - AWS Glue Console

Search [Option+S]

N. Virginia

### AWS Glue

- Getting started
- ETL jobs
  - Visual ETL
  - Notebooks
  - Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)
- ▼ Data Catalog
  - Databases
  - Tables
  - Stream schema registries
  - Schemas
- Connections
- Crawlers
  - Classifiers
- Catalog settings
- ▼ Data Integration and ETL
  - ETL jobs
    - Visual ETL
    - Notebooks
    - Job run monitoring
  - Interactive Sessions
  - Data classification tools
  - Sensitive data detection

CloudShell Feedback

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie p

## covid-19-data

Last updated (UTC)  
May 7, 2024 at 14:03:33

Run crawler Edit Delete

### Crawler properties

Name covid-19-data	IAM role covid-role	Database covid-db	State STOPPING
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

► Advanced settings

Crawler runs Schedule Data sources Classifiers Tags

### Crawler runs (1)

The list of crawler runs for this crawler.

Filter data Filter by a date and time range

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
<input type="radio"/>	May 7, 2024 at 14:01:43	May 7, 2024 at 14:03:32	01 min 48 s	Completed	-	-

Status is now completed

Again go inside database there will be tables

us-east-1.console.aws.amazon.com

AWS Glue | us-east-1

Table Detail - AWS Glue Console

Search [Option+S]

N. Virginia

### AWS Glue

- Getting started
- ETL jobs
  - Visual ETL
  - Notebooks
  - Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)
- ▼ Data Catalog
  - Databases
  - Tables
  - Stream schema registries
  - Schemas
- Connections
- Crawlers
  - Classifiers
- Catalog settings
- ▼ Data Integration and ETL
  - ETL jobs
    - Visual ETL
    - Notebooks
    - Job run monitoring
  - Interactive Sessions
  - Data classification tools
  - Sensitive data detection

CloudShell Feedback

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie p

## source\_bucket19

Last updated (UTC)  
May 7, 2024 at 14:03:44

Version 0 (Current version) Actions

Table overview Data quality New

### Table details

Name source_bucket19	Description -	Database covid-db	Classification CSV
Location s3://source-bucket19/	Connection -	Deprecated -	Last updated May 7, 2024 at 14:03:31
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive ql.io.HiveLongKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema Partitions Indexes Column statistics - new

### Schema (15)

View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key	Comment
1	country/region	string	-	-
2	confirmed	bigint	-	-
3	deaths	bigint	-	-
4	recovered	bigint	-	-

## 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

Runs

Data quality - updated

Schedules

Version Control

+

Data source - S3 bucket

Amazon S3

Transform - SelectFields

Select Fields

Data target - S3 bucket

Amazon S3

Data preview

Output schema

Start a data preview session

IAM role

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

CloudShell

Feedback

© 2024, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie p

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	Id	Run status	Glue version
covid-job	jr_1c3118ba7c5058cb32bee19a43ce24799c5fc200961a865410a88db5e760a9cc	Succeeded	4.0
Retry attempt number	Start time (Local)	End time (Local)	Start time (UTC)
Initial run	05/07/2024 19:41:53	05/07/2024 19:42:57	2024/05/07 14:11:53
End time (UTC)	Start-up time	Execution time	Last modified on (Local)
2024/05/07 14:12:57	12 seconds	51 seconds	05/07/2024 19:42:57
Last modified on (UTC)	Trigger name	Security configuration	Timeout
2024/05/07 14:12:57	-	-	2880 minutes
Max capacity	Number of workers	Worker type	Execution class
10 DPUs	10	G.1X	Standard
Log group name	Cloudwatch logs	Performance and debugging recommendations	
/aws-glue/jobs	All logs	View in CloudWatch	

CloudShell

Feedback

© 2024, Amazon Web Services, Inc. or its affiliates.

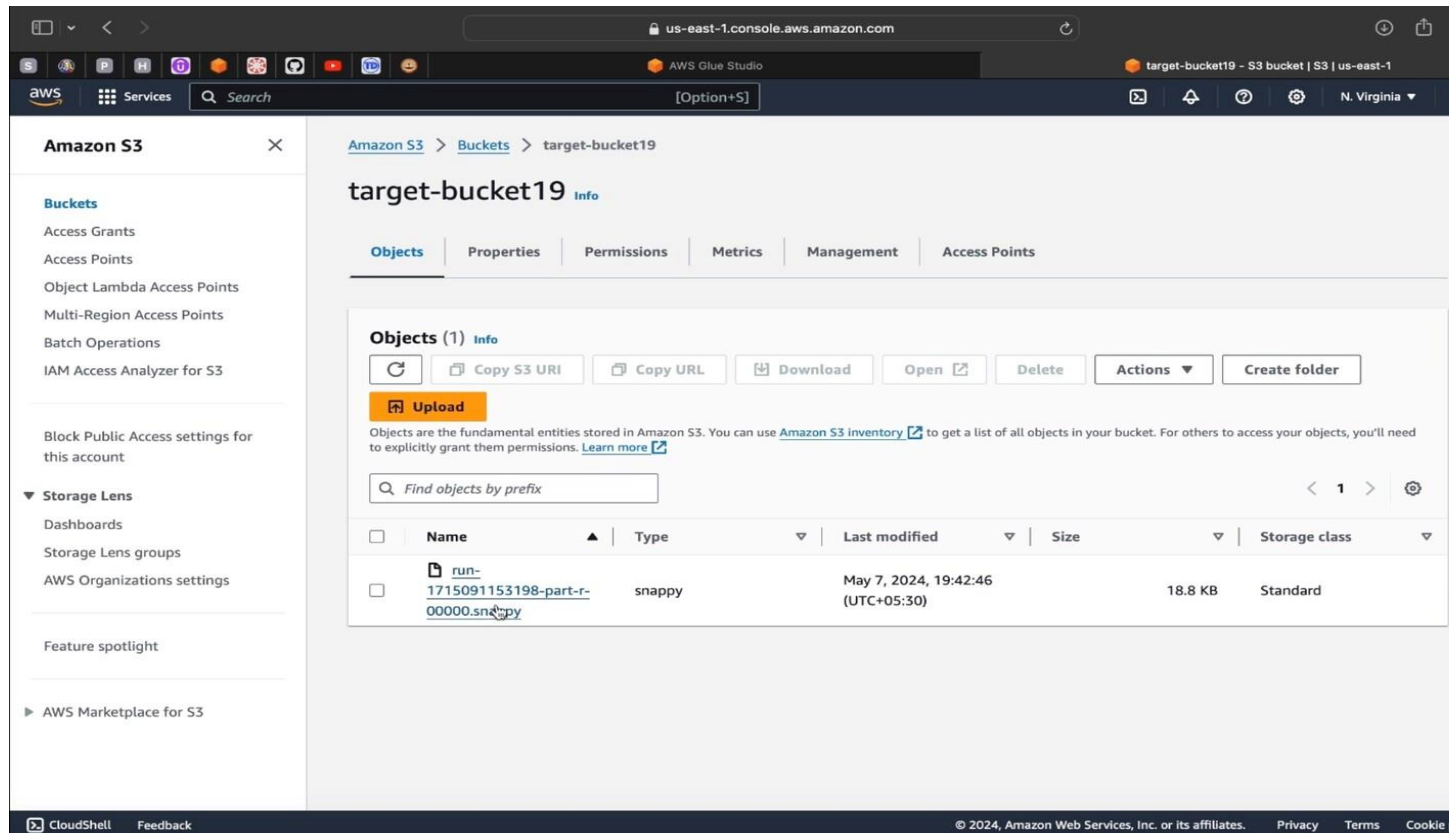
Privacy

Terms

Cookie p

## Step 6: Check Transformed Data in Target Bucket

- Verify that the transformed data is successfully loaded into the target S3 bucket.
- Goto your 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

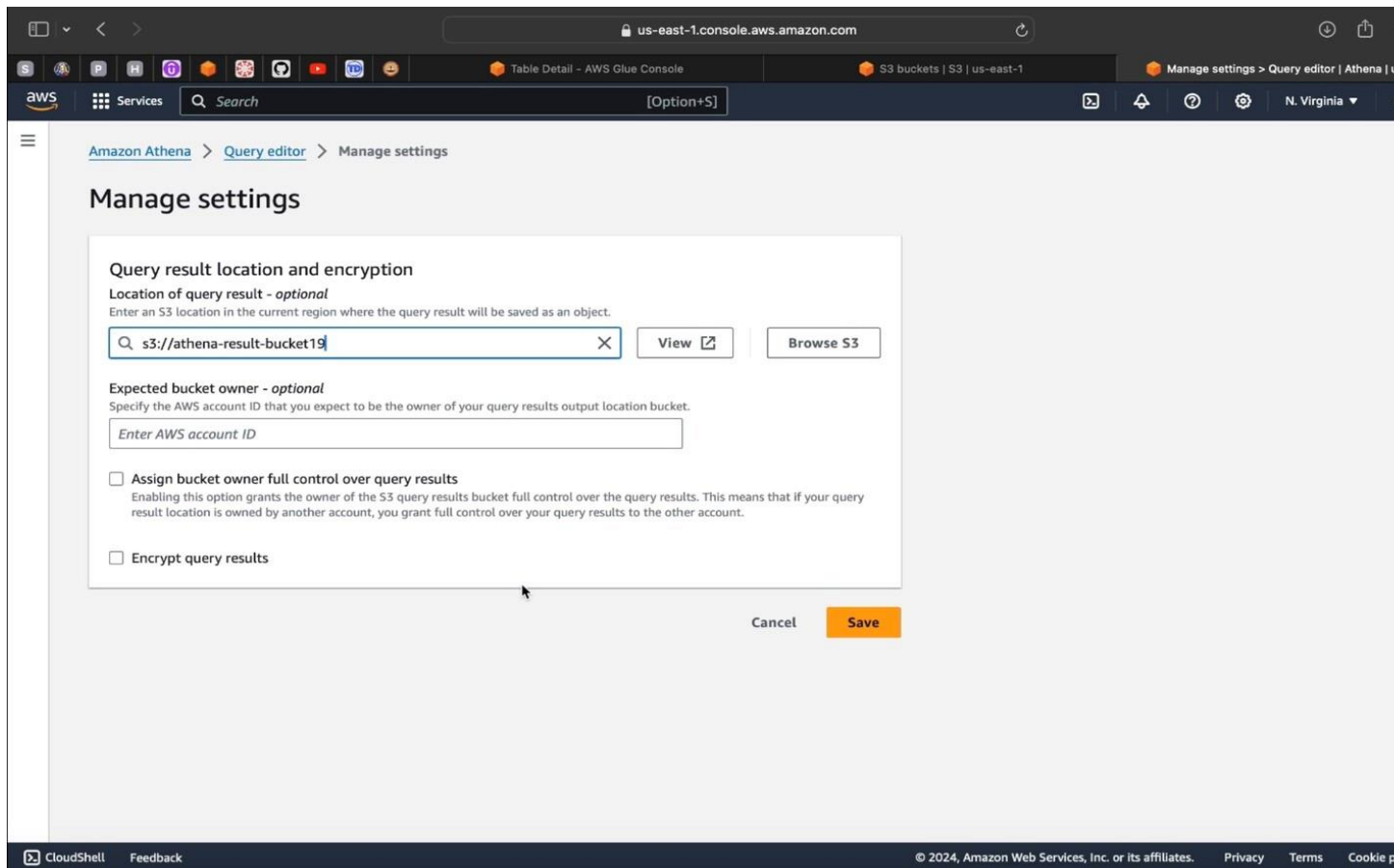
Initially Query will not run so,

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.



us-east-1.console.aws.amazon.com

Services Search [Option+S]

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings Workgroup primary

Data

Data source  
AwsDataCatalog

Database  
covid-db

Tables and views Create Filter tables and views

Tables < 1 >  
Loading tables

Views < 1 >  
Loading views

Query 1 Query 2 Query 3 Query 4

1 SELECT \* FROM "AwsDataCatalog"."covid-db"."source\_bucket19" limit 10;

SQL Ln 1, Col 1

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 178 ms Run time: 411 ms Data scanned: 14.25 KB

Results Copy Download results

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie p

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

us-east-1.console.aws.amazon.com

Table Detail - AWS Glue ConsoleS3 buckets | S3 | us-east-1Query editor | Athena | us-east-1Query editor | Athena | N. Virginia

ServicesSearch[Option+S]

Query resultsQuery stats

CompletedTime in queue: 178 msRun time: 411 msData scanned: 14.25 KB

Results (10)CopyDownload results

Search rows

#	country/region	confirmed	deaths	recovered	active	new cases	new deaths	new recovered
1	Afghanistan	36263	1269	25198	9796	106	10	18
2	Albania	4880	144	2745	1991	117	6	63
3	Algeria	27973	1163	18837	7973	616	8	749
4	Andorra	907	52	803	52	10	0	0
5	Angola	950	41	242	667	18	1	0
6	Antigua and Barbuda	86	3	65	18	4	0	5
7	Argentina	167416	3059	72575	91782	4890	120	2057
8	Armenia	37390	711	26665	10014	73	6	187
9	Australia	15303	167	9311	5825	368	6	137
10	Austria	20558	713	18246	1599	86	1	37

CloudShellFeedback© 2024, Amazon Web Services, Inc. or its affiliates.PrivacyTermsCookie

View125%ZoomAdd CategoryPivot TableInsertTableChartTextShapeMediaCommentShare

Sheet 1

8dfcbdf8-07bd-4fa2-956d-4a00a857530d

	country/region	confirmed	deaths	recovered	active	new cases	new deaths	new recovered	deaths / 100 cases	recovered / 100 cases	deaths / 100 recovered	confirmed last
2	Afghanistan	36263	1269	25198	9796	106	10	18	3.5	69.49	5.04	
3	Albania	4880	144	2745	1991	117	6	63	2.95	56.25	5.25	
4	Algeria	27973	1163	18837	7973	616	8	749	4.16	67.34	6.17	
5	Andorra	907	52	803	52	10	0	0	5.73	88.53	6.48	
6	Angola	950	41	242	667	18	1	0	4.32	25.47	16.94	
7	Antigua and Barbuda	86	3	65	18	4	0	5	3.49	75.58	4.62	
8	Argentina	167416	3059	72575	91782	4890	120	2057	1.83	43.35	4.21	
9	Armenia	37390	711	26665	10014	73	6	187	1.9	71.32	2.67	
10	Australia	15303	167	9311	5825	368	6	137	1.09	60.84	1.79	
11	Austria	20558	713	18246	1599	86	1	37	3.47	88.75	3.91	

TableCellTextArrange

Table Styles

Table Options

☒ Title

☐ Caption

Headers & Footer

1

1

0

Rows

11

Columns

15

Table Font Size

A

A

Table Outline

0.35 pt

☐ Outline Table Title

Gridlines

☐ Alternating Row Colour

Row & Column Size

Height

20 pt

Fit

Width

110 pt

Fit

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.

