

COVID-19 DATA ANALYSIS

OBJECTIVE

- To perform data analysis and visualization on Covid-19 data using ETL tools.
- To perform data analysis on Covid_19 data and extract the meaningful information from the dataset which will help in taking quick and informed decision.

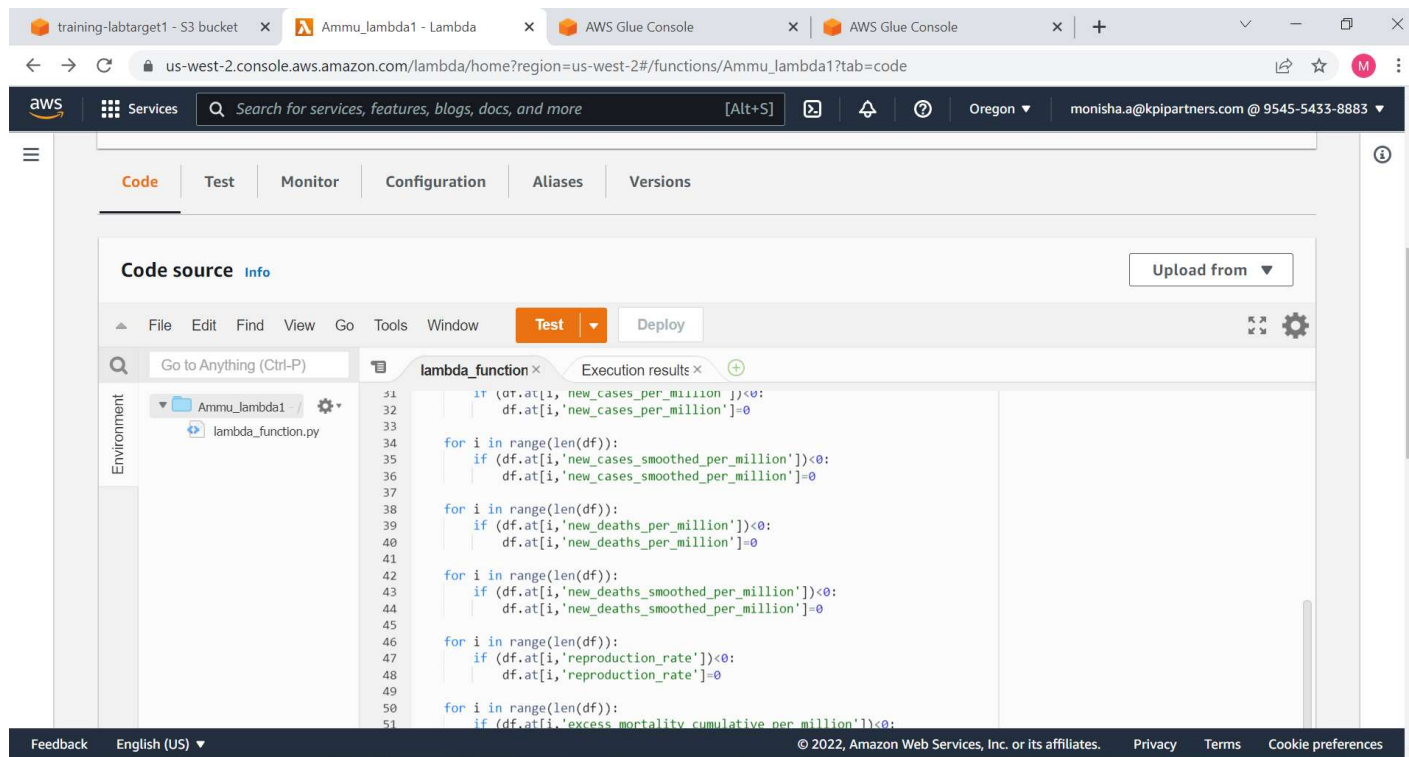
TECHNOLOGIES USED

- AWS Glue
- SparkSQL
- Redshift
- AWS S3
- SparkSQL
- Crawler
- AWS Lambda
- PySpark

Problem Statements

1. Clean and transform data for processing

job name: Ammu_lambda1



The screenshot displays the AWS Lambda console interface for the function 'Ammu_lambda1'. The 'Code source' tab is active, showing a Python script named 'lambda_function.py'. The script contains logic for processing data, likely from a CSV file, and calculating various metrics. The code includes several loops and conditional statements to handle data cleaning and transformation.

```
51 if (df.at[i, 'new_cases_per_million'] < 0):
52     df.at[i, 'new_cases_per_million'] = 0
53
54 for i in range(len(df)):
55     if (df.at[i, 'new_cases_smoothed_per_million'] < 0):
56         df.at[i, 'new_cases_smoothed_per_million'] = 0
57
58 for i in range(len(df)):
59     if (df.at[i, 'new_deaths_per_million'] < 0):
60         df.at[i, 'new_deaths_per_million'] = 0
61
62 for i in range(len(df)):
63     if (df.at[i, 'new_deaths_smoothed_per_million'] < 0):
64         df.at[i, 'new_deaths_smoothed_per_million'] = 0
65
66 for i in range(len(df)):
67     if (df.at[i, 'reproduction_rate'] < 0):
68         df.at[i, 'reproduction_rate'] = 0
69
70 for i in range(len(df)):
71     if (df.at[i, 'excess_mortality_cumulative_per_million'] < 0):
```

1.Clean and transform data for processing Cont...

Status

✓ Successfully returned 5 records in 572 ms

Bytes returned: 2671 B

Raw

Formatted

| < 1 > | | | | | | |
|----------|-----------|-------------|------------|-------------|-----------|--------------------|
| iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed |
| AFG | Asia | Afghanistan | 2020-02-24 | 5.0 | 5.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-25 | 5.0 | 0.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-26 | 5.0 | 0.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-27 | 5.0 | 0.0 | 0.0 |

2. ETL operations on dataset



The screenshot displays a code editor with a file named `lambda_function.py` in the `Environment` pane. The code is a Python lambda function that processes a dataset (likely a pandas DataFrame) and performs several ETL operations. The code is as follows:

```
48 df.at[i, 'reproduction_rate'] = 0
49
50 for i in range(len(df)):
51     if (df.at[i, 'excess_mortality_cumulative_per_million']) < 0:
52         df.at[i, 'excess_mortality_cumulative_per_million'] = 0
53
54 for i in range(len(df)):
55     if (df.at[i, 'excess_mortality']) < 0:
56         df.at[i, 'excess_mortality'] = 0
57
58 for i in range(len(df)):
59     if (df.at[i, 'excess_mortality_cumulative']) < 0:
60         df.at[i, 'excess_mortality_cumulative'] = 0
61
62 for i in range(len(df)):
63     if (df.at[i, 'excess_mortality_cumulative_absolute']) < 0:
64         df.at[i, 'excess_mortality_cumulative_absolute'] = 0
65
66 for i in range(len(df)):
67     if (df.at[i, 'excess_mortality_cumulative']) < 0:
68         df.at[i, 'excess_mortality_cumulative'] = 0
69
70
71
72 for i in range(len(df)):
73     if (df.at[i, 'continent']) == 0:
74         df.drop([i], axis=0, inplace=True)
75
76
77
78 wr.s3.to_csv(df, "s3://training-labtarget1/TrainingLab/monishaD/OUTPUT/Covid/clean_covid_filtered.csv", index = False)
79
80 return {
81     'statusCode': 200,
82     'body': json.dumps('Successful')
83 }
84
```

The code editor shows the file `lambda_function.py` in the `Environment` pane. The code is a Python lambda function that processes a dataset (likely a pandas DataFrame) and performs several ETL operations. The code is as follows:

78:67 Python Spaces: 4

Code properties

2. ETL operations on dataset Cont...

Status

✓ Successfully returned 5 records in 572 ms

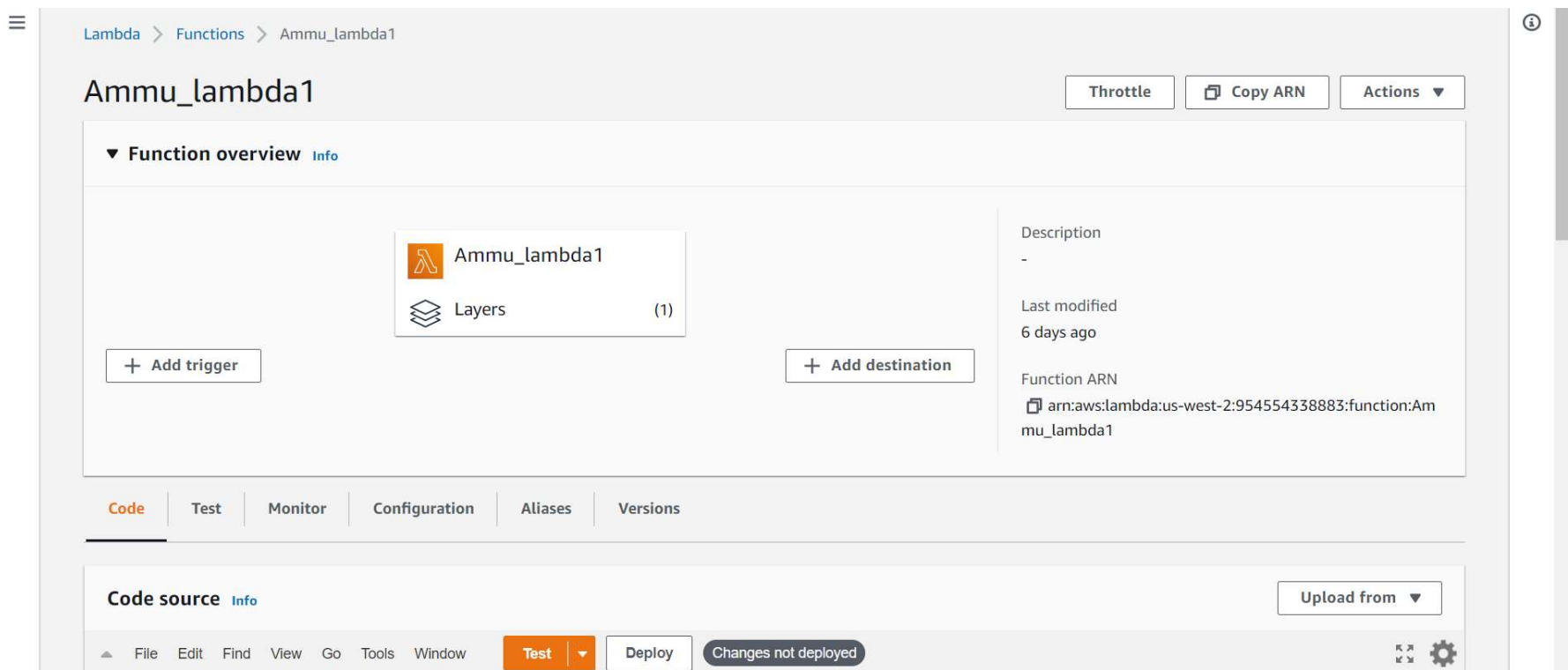
Bytes returned: 2671 B

Raw

Formatted

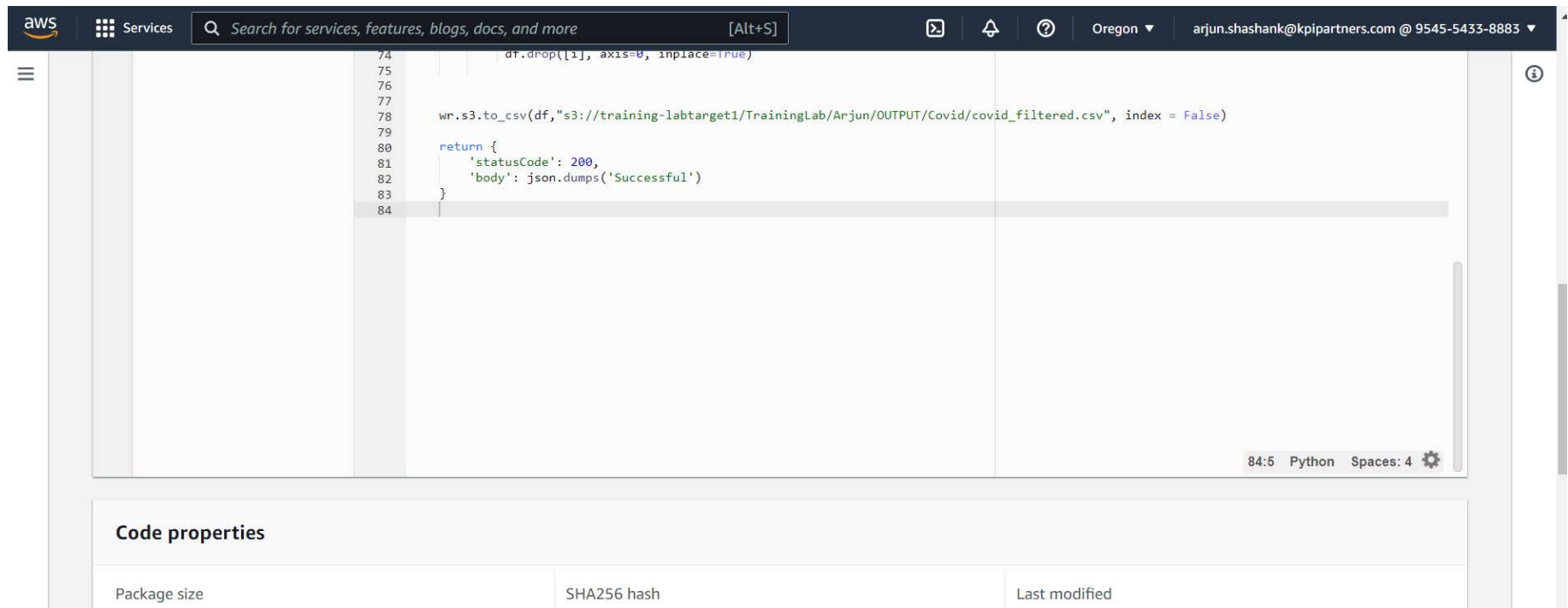
| < 1 > | | | | | | |
|----------|-----------|-------------|------------|-------------|-----------|--------------------|
| iso_code | continent | location | date | total_cases | new_cases | new_cases_smoothed |
| AFG | Asia | Afghanistan | 2020-02-24 | 5.0 | 5.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-25 | 5.0 | 0.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-26 | 5.0 | 0.0 | 0.0 |
| AFG | Asia | Afghanistan | 2020-02-27 | 5.0 | 0.0 | 0.0 |

3. Storing the modified data in AWS



The screenshot displays the AWS Lambda console interface for a function named 'Ammu_lambda1'. The breadcrumb navigation at the top shows 'Lambda > Functions > Ammu_lambda1'. The function name 'Ammu_lambda1' is prominently displayed at the top left of the main content area. To the right of the name are three buttons: 'Throttle', 'Copy ARN', and 'Actions'. Below the function name, the 'Function overview' section is expanded, showing a card for 'Ammu_lambda1' with a 'Layers (1)' link. To the left of this card is a '+ Add trigger' button, and to the right is a '+ Add destination' button. On the right side of the overview, the following details are listed: 'Description' (empty), 'Last modified' (6 days ago), and 'Function ARN' (arn:aws:lambda:us-west-2:954554338883:function:Ammu_lambda1). Below the overview, a horizontal tab bar includes 'Code' (selected), 'Test', 'Monitor', 'Configuration', 'Aliases', and 'Versions'. The 'Code source' section is visible below the tabs, featuring an 'Upload from' dropdown menu. At the bottom of the console, a toolbar contains a menu (File, Edit, Find, View, Go, Tools, Window), a 'Test' button, a 'Deploy' button, and a 'Changes not deployed' status indicator.

3. Storing the modified data in AWS Cont...



The screenshot displays the AWS Lambda console interface. At the top, there's a navigation bar with the AWS logo, a 'Services' menu, a search bar, and user information. The main area shows a code editor with a Python function. The code includes a line to drop a row from a DataFrame and another to write the DataFrame to an S3 bucket. Below the code editor, there's a 'Code properties' section with a table containing columns for 'Package size', 'SHA256 hash', and 'Last modified'.

```
74         df.drop([i], axis=0, inplace=True)
75
76
77
78     wr.s3.to_csv(df, "s3://training-labtarget1/TrainingLab/Anjun/OUTPUT/Covid/covid_filtered.csv", index = False)
79
80     return {
81         'statusCode': 200,
82         'body': json.dumps('Successful')
83     }
84
```

84:5 Python Spaces: 4

| Code properties | | |
|-----------------|-------------|---------------|
| Package size | SHA256 hash | Last modified |

4.Display total cases ,new cases , recovered cases and deaths.



CREATING CRAWLER

AWS Glue

Data catalog

- Databases
- Tables
- Connections
- Crawlers
- Classifiers
- Schema registries
- Schemas
- Settings

ETL

- AWS Glue Studio [↗](#)
- Jobs [↗](#) - **New**
- Jobs (legacy)
- ML Transforms
- Blueprints
- Workflows

Tables > clean_covid_filtered_csv

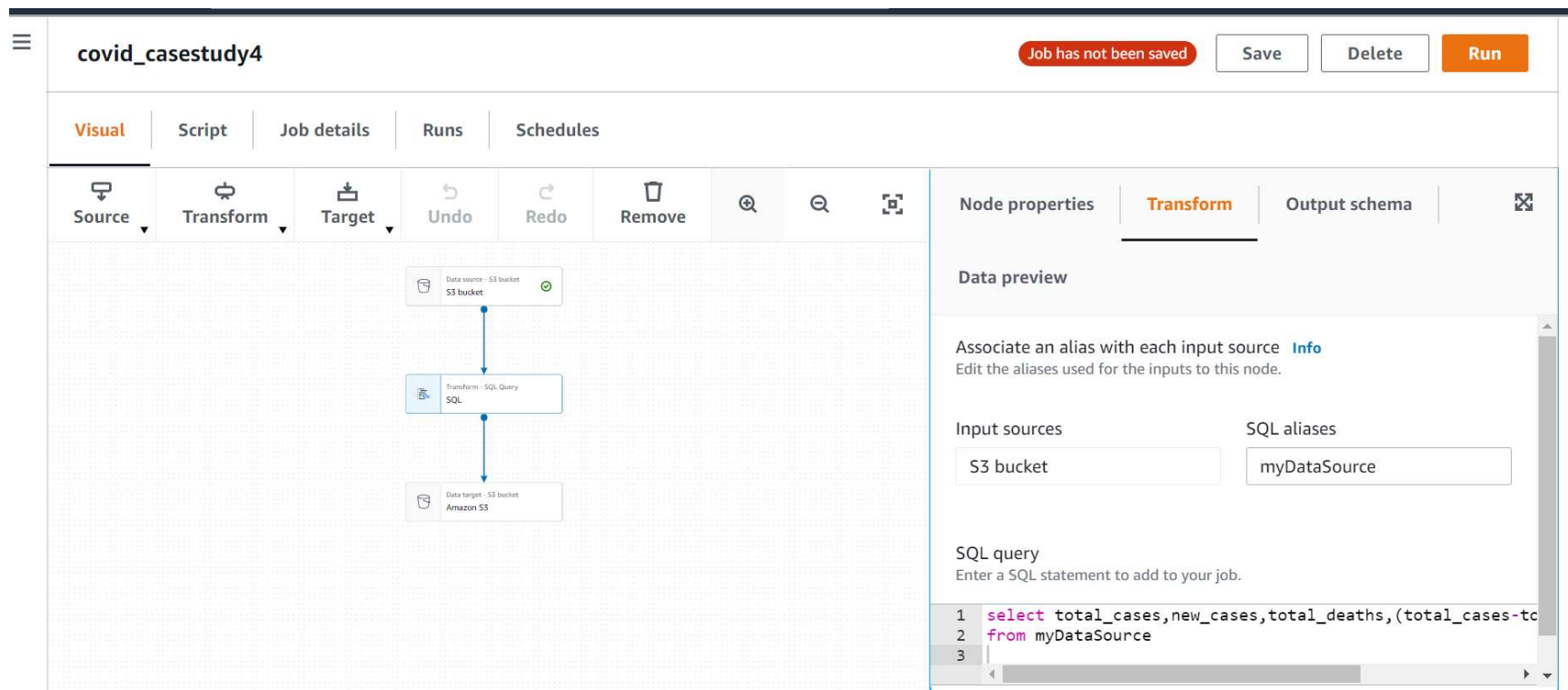
Last updated 26 Feb 2022 09:22 PM **Table** Version (Current version) ▼

[Edit table](#) [Delete table](#) [View properties](#) [Compare versions](#) [Edit schema](#)

| | |
|--------------------------------|--|
| Name | clean_covid_filtered_csv |
| Description | |
| Database | mounisha_db_crawler |
| Classification | csv |
| Location | s3://training-labtarget1/TrainingLab/Monisha/OUTPUT/Covid/clean_covid_filtered.csv |
| Connection | |
| Deprecated | No |
| Last updated | Sat Feb 26 21:22:49 GMT+530 2022 |
| Input format | org.apache.hadoop.mapred.TextInputFormat |
| Output format | org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat |
| Serde serialization lib | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe |
| Serde parameters | field.delim , |
| | skip.header.line.count 1 sizeKey 57098243 objectCount 1 |
| | UPDATED_BY_CRAWLER mouni_covid_crawler |

4. Display total cases ,new cases ,recovered cases and deaths.

Job name:covid_casestudy4



The screenshot shows the KPI PARTNERS job configuration interface for a job named 'covid_casestudy4'. The interface includes a top navigation bar with tabs for 'Visual', 'Script', 'Job details', 'Runs', and 'Schedules'. Below this is a toolbar with icons for 'Source', 'Transform', 'Target', 'Undo', 'Redo', 'Remove', and search functions. The main workspace displays a workflow diagram with three nodes: 'Data source - S3 bucket S3 bucket', 'Transform - SQL Query SQL', and 'Data target - S3 bucket Amazon S3'. The right-hand panel is currently on the 'Transform' tab, showing 'Node properties', 'Data preview', and 'SQL query' sections. The 'Data preview' section includes instructions to associate aliases with input sources. The 'Input sources' field is set to 'S3 bucket', and the 'SQL aliases' field is set to 'myDataSource'. The 'SQL query' section contains a SQL statement:

```
1 select total_cases,new_cases,total_deaths,(total_cases-to
2 from myDataSource
3
```

4.OUTPUT

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

3

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Download results

Status

✔ Successfully returned 5 records in 419 ms

Bytes returned: 109 B

Raw

Formatted

total_cases,new_cases,total_deaths,recovered

5.0,5.0,0.0,5.0

5.0,0.0,0.0,5.0

5.0,0.0,0.0,5.0

5.0,0.0,0.0,5.0

Close

5. Which country in Distinct WHO region has highest cases till date.
JOB NAME : covid_casestudy5



covid_casestudy5

Job has not been saved

Save

End session

Delete

Run

Visual

Script

Job details

Runs

Schedules

Source

Transform

Target

Undo

Redo

Remove

Data source - S3 bucket

Transform - SQL Query

Data target - S3 bucket Amazon S3

Node properties

Transform

Output schema

Data preview

Associate an alias with each input source [Info](#)
Edit the aliases used for the inputs to this node.

Input sources

SQL aliases

S3 bucket

covid5

SQL query

Enter a SQL statement to add to your job.

1 select location


2 from covid5

3 where total_cases in (select max(total_cases)

4 from covid5

5.OUTPUT





Query results

Download results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Status

✔ Successfully returned 2 records in 327 ms

Bytes returned: 16 B

Raw

Formatted

location

Poland

Close

6. Total no of confirmed cases over between a certain date.

JOB NAME : mounisha_casestudy6

mounisha_casestudy6

Last Saved at 3/4/2022, 7:05:38 PM

Save

Delete

Run

Visual

Script

Job details

Runs

Schedules

Source

Transform

Target

Undo

Redo

Remove

Data source - S3 bucket

S3 bucket

Transform - SQL Query

SQL

Data target - S3 bucket

Amazon S3

Node properties

Transform

Output schema

Data preview

Associate an alias with each input source

Info

Edit the aliases used for the inputs to this node.

Input sources

SQL aliases

S3 bucket

covid6

SQL query

Enter a SQL statement to add to your job.


1

select sum(new_cases) from covid6 where date between '2-17-2020' and '3-21-2020'

2


6.OUTPUT






Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

 Download results

Status


 Successfully returned 2 records in 1415 ms

Bytes returned: 28 B

Raw


Formatted

```
sum(new_cases)
4.18499208E8
```



8. Date of first confirmed case in a particular region.



JOB NAME :mouni_covid_8

 **mouni_covid_8** Last Saved at 2/28/2022, 11:53:36 AM Save Delete Run 

Script Job details Runs Schedules

Script [Info](#)

```
33     "recurse": True,
34 },
35     transformation_ctx="S3bucket_node1",
36 )
37
38 # Script generated for node SQL
39 SqlQuery0 = ""
40 select first(date) from covid8 where continent='Asia'
41
42 ""
43 SQL_node1646029254800 = sparkSqlQuery(
44     glueContext,|
45     query=SqlQuery0,
46     mapping={"covid8": S3bucket_node1},
47     transformation_ctx="SQL_node1646029254800",
48 )
```

Python Ln 44, Col 17  Errors: 0  Warnings: 0 

8.OUTPUT



Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

3

Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Download results

Status

Successfully returned 2 records in 314 ms

Bytes returned: 23 B

Raw

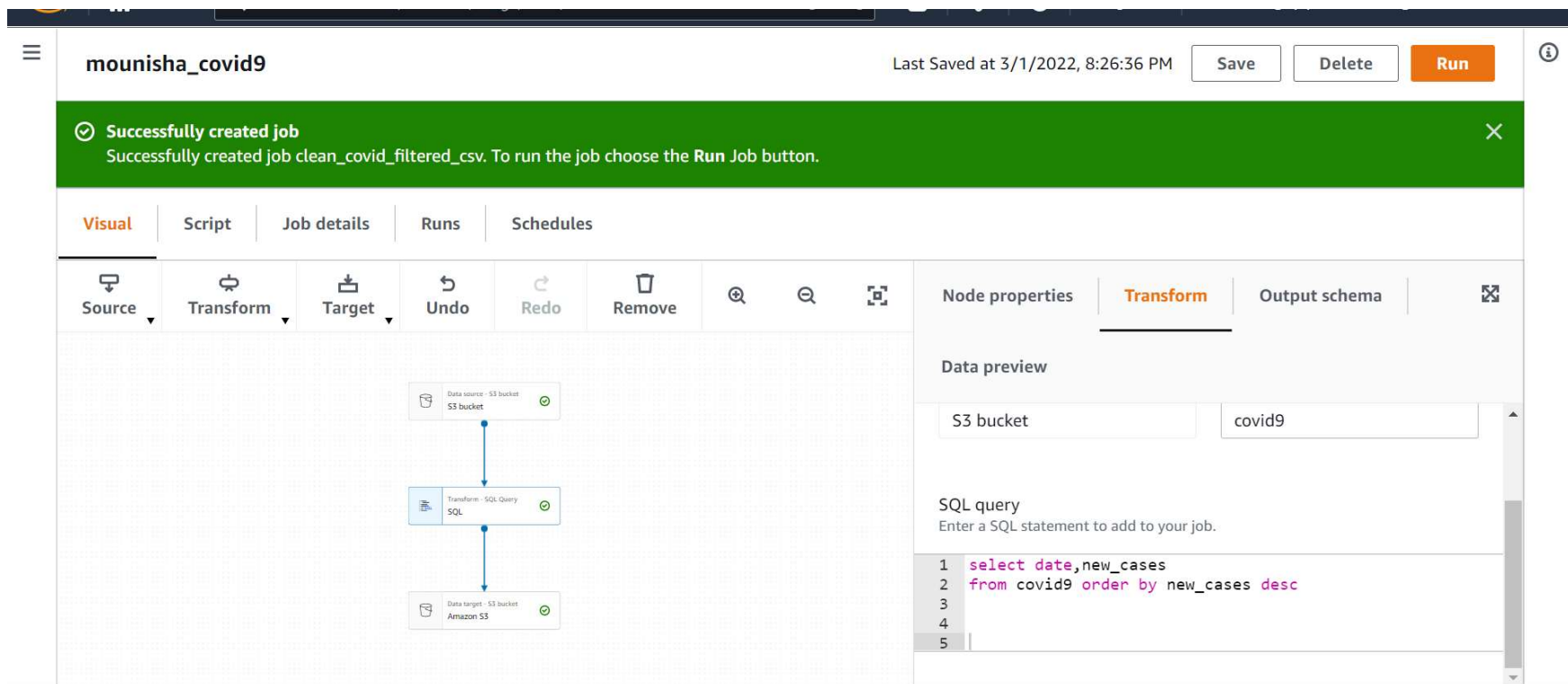
Formatted

first(date)

2020-02-24

9.Date on which max no. cases were reported in a country.

JOB NAME : mounisha_covid9



The screenshot shows the KPI PARTNERS job configuration interface for a job named 'mounisha_covid9'. The interface includes a top bar with a menu icon, the job name, the last saved time (3/1/2022, 8:26:36 PM), and buttons for Save, Delete, and Run. A green notification banner at the top states: 'Successfully created job. Successfully created job clean_covid_filtered_csv. To run the job choose the Run Job button.' Below the notification are tabs for Visual, Script, Job details, Runs, and Schedules. The Visual tab is active, showing a workflow diagram with three nodes: 'Data source - S3 bucket S3 bucket', 'Transform - SQL Query SQL', and 'Data target - S3 bucket Amazon S3'. The right sidebar contains 'Node properties', 'Transform' (selected), and 'Output schema'. The 'Data preview' section shows 'S3 bucket' and 'covid9'. The 'SQL query' section contains the following SQL statement:

```
1 select date,new_cases
2 from covid9 order by new_cases desc
3
4
5
```

9.OUTPUT



Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Download results

Status

Successfully returned 3 records in 353 ms

Bytes returned: 57 B

Raw

Formatted

< 1 >

| date | new_cases |
|------------|-----------|
| 2022-01-10 | 1368563.0 |
| 2022-01-18 | 1113068.0 |

10. Line chart showing total cases, deaths & recoveries of a particular country .

JOB NAME: mouni_q10



Visual Script Job details Runs Schedules

Source Transform Target Undo Redo Remove

mouni_q10 Last Saved at 3/4/2022, 1:21:17 PM Save Delete Run

Data source - S3 bucket S3 bucket

Transform - SQL Query SQL

Data target - Redshift Amazon Redshift

Node properties Transform Output schema Data preview

Associate an alias with each input source [Info](#)
Edit the aliases used for the inputs to this node.

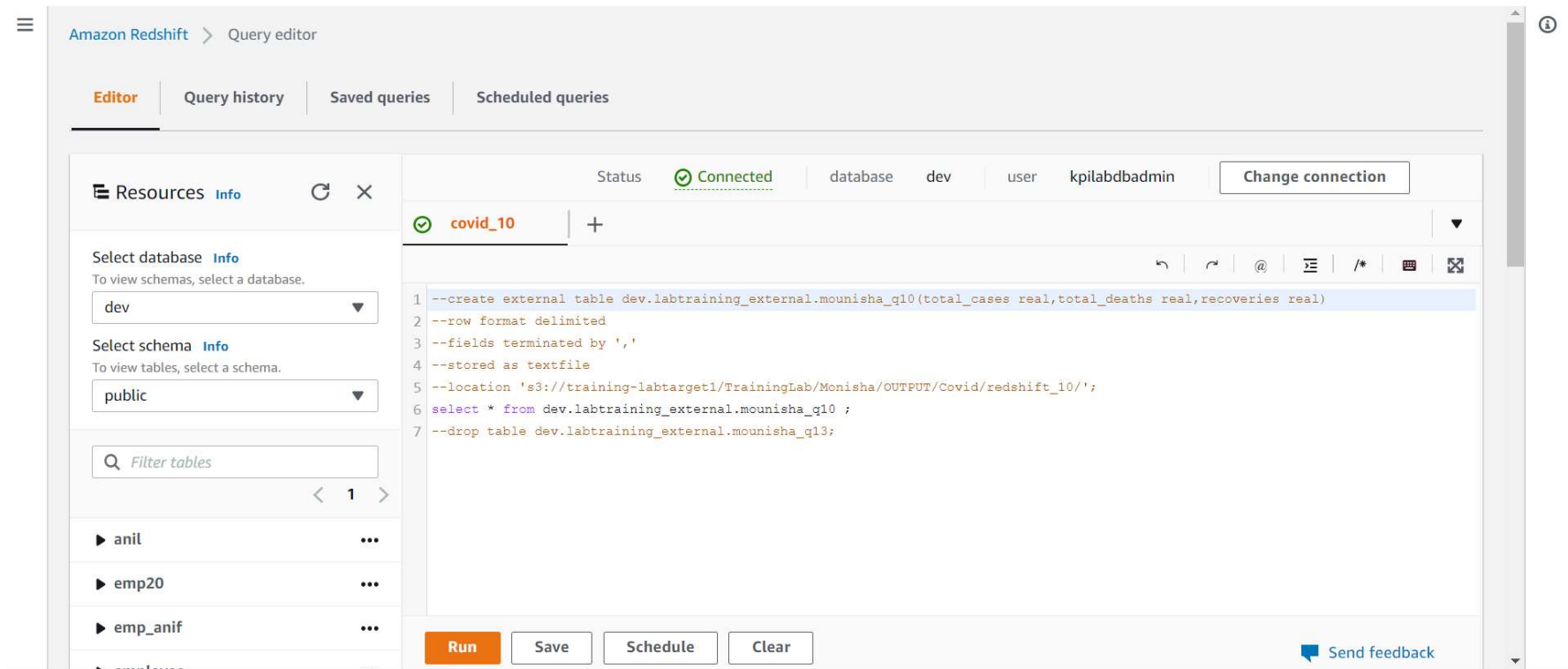
Input sources SQL aliases

S3 bucket covid10

SQL query
Enter a SQL statement to add to your job.

```
1 select total_cases,total_deaths,(total_cases-total_deaths) as recoveries from
2
```

10 . Creating external table



The screenshot displays the Amazon Redshift Query Editor interface. The top navigation bar shows "Amazon Redshift" and "Query editor". Below this, there are tabs for "Editor", "Query history", "Saved queries", and "Scheduled queries". The "Editor" tab is active.


On the left side, there is a "Resources" panel with a search bar and a list of tables. The "Select database" dropdown is set to "dev", and the "Select schema" dropdown is set to "public". The table list shows "anil", "emp20", "emp_anif", and "employee".

The main query editor area shows a SQL script for creating an external table. The status bar at the top indicates "Status: Connected" and "database: dev user: kpiabdbadmin".

```
1 --create external table dev.labtraining_external.mounisha_q10(total_cases real,total_deaths real,recoveries real)
2 --row format delimited
3 --fields terminated by ','
4 --stored as textfile
5 --location 's3://training-labtarget1/TrainingLab/Monisha/OUTPUT/Covid/redshift_10/';
6 select * from dev.labtraining_external.mounisha_q10 ;
7 --drop table dev.labtraining_external.mounisha_q13;
```

At the bottom, there are buttons for "Run", "Save", "Schedule", and "Clear". A "Send feedback" link is also present.

10. OUTPUT



Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Status

✔ Successfully returned 5 records in 397 ms

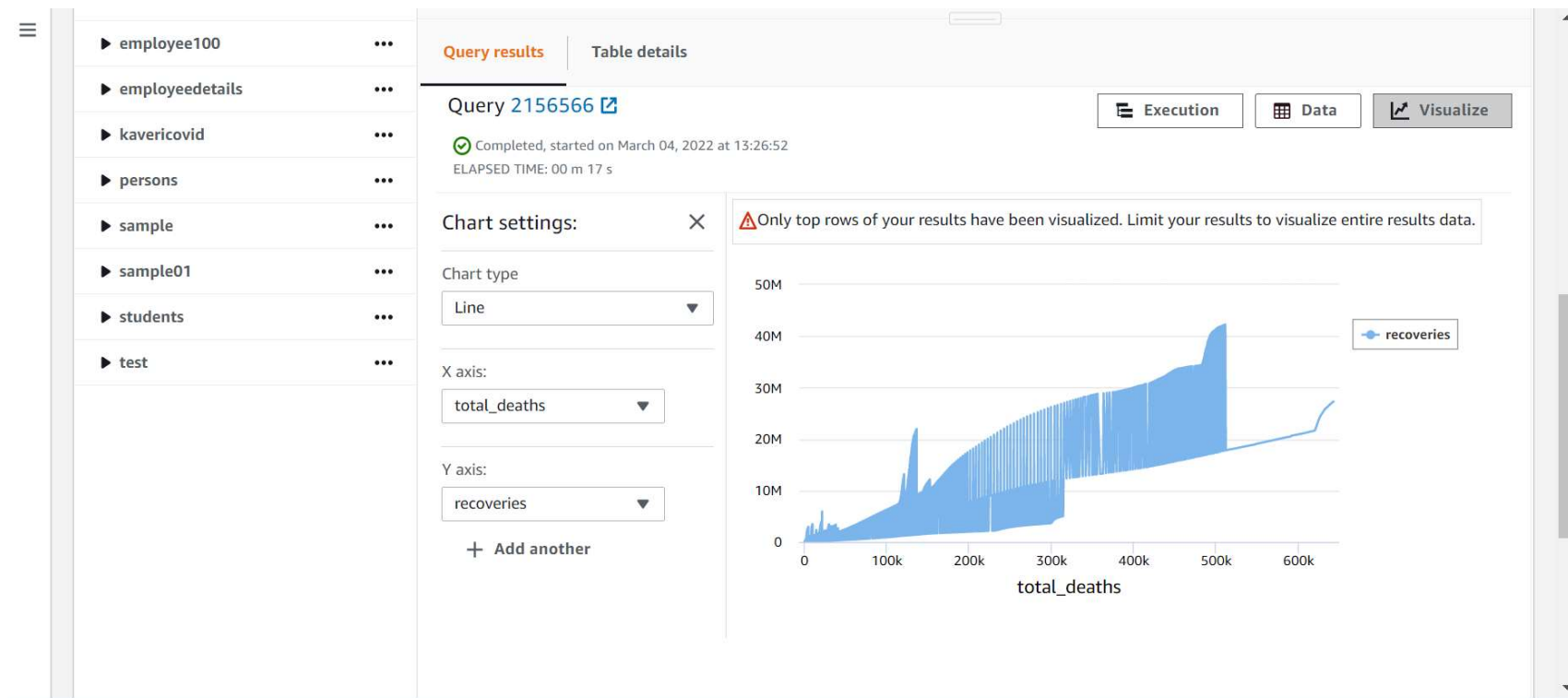
Bytes returned: 84 B

Raw | Formatted

```
total_cases,total_deaths,recoveries
5.0,0.0,5.0
5.0,0.0,5.0
5.0,0.0,5.0
5.0,0.0,5.0
```

Close

10. Line chart



13.No. of active cases vs critical cases of a country.

JOB NAME: mounisha_redshift

mounisha_redshift

Last Saved at 3/4/2022, 12:54:56 PM

Save

Delete

Run

Visual

Script

Job details

Runs

Schedules

Source

Transform

Target

Undo

Redo

Remove

Node properties

Transform

Output schema

Data preview

Associate an alias with each input source [Info](#)
Edit the aliases used for the inputs to this node.

Input sources

SQL aliases

S3 bucket

covid13

SQL query

Enter a SQL statement to add to your job.

1 select location,sum(total_cases) as active_cases, sum(icu_patients) as critica

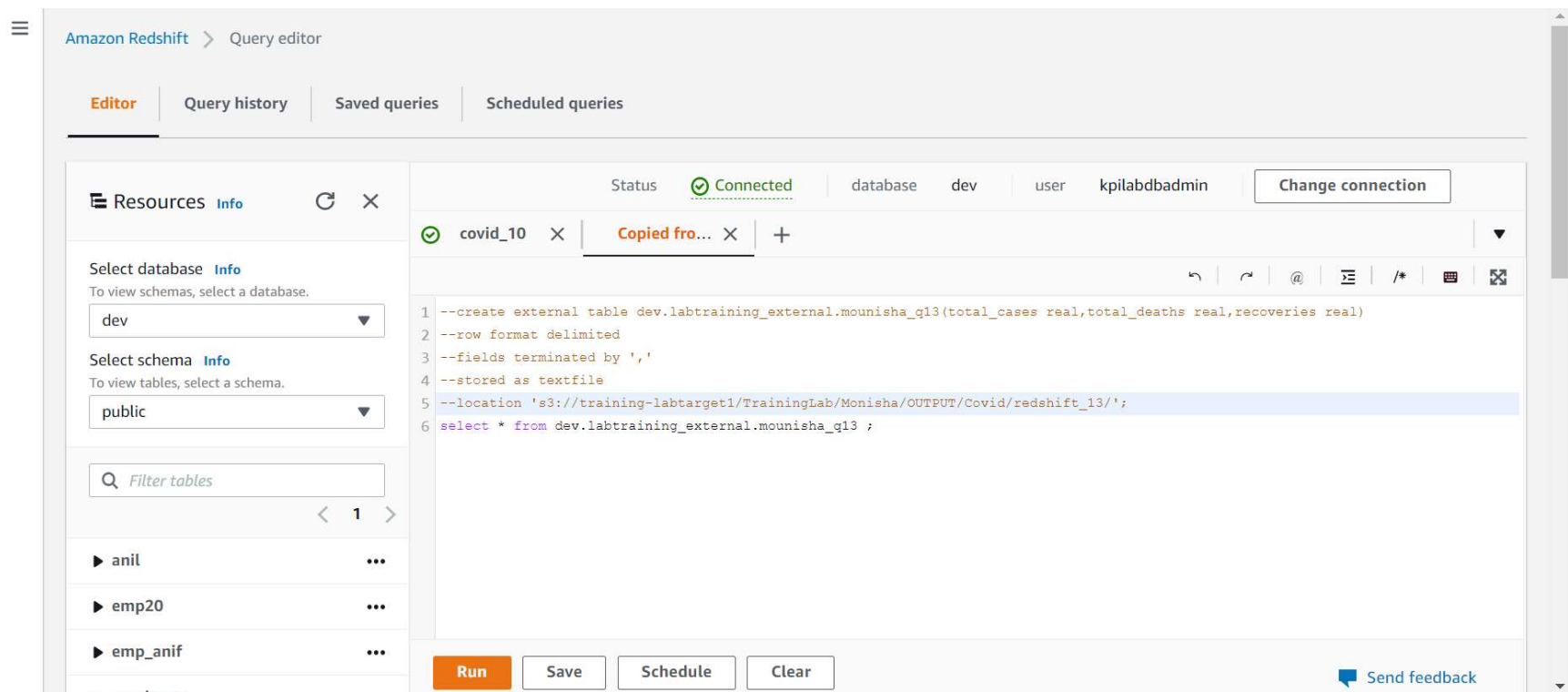
2

Data source - S3 bucket
S3 bucket

Transform - SQL Query
SQL

Data target - Redshift
Amazon Redshift

13. Creating external table



The screenshot displays the Amazon Redshift Query Editor interface. The top navigation bar shows "Amazon Redshift" and "Query editor". Below this, there are tabs for "Editor", "Query history", "Saved queries", and "Scheduled queries". The "Editor" tab is active.

On the left side, there is a "Resources" panel with a search bar and a list of tables. The "Select database" dropdown is set to "dev", and the "Select schema" dropdown is set to "public". The table list shows "anil", "emp20", "emp_anif", and "employees".

The main editor area shows a SQL query being written. The status bar at the top indicates "Status: Connected" and "database: dev, user: kpilabdbadmin". The query is as follows:

```
1 --create external table dev.labtraining_external.mounisha_q13(total_cases real,total_deaths real,recoveries real)
2 --row format delimited
3 --fields terminated by ','
4 --stored as textfile
5 --location 's3://training-labtarget1/TrainingLab/Monisha/OUTPUT/Covid/redshift_13/';
6 select * from dev.labtraining_external.mounisha_q13 ;
```

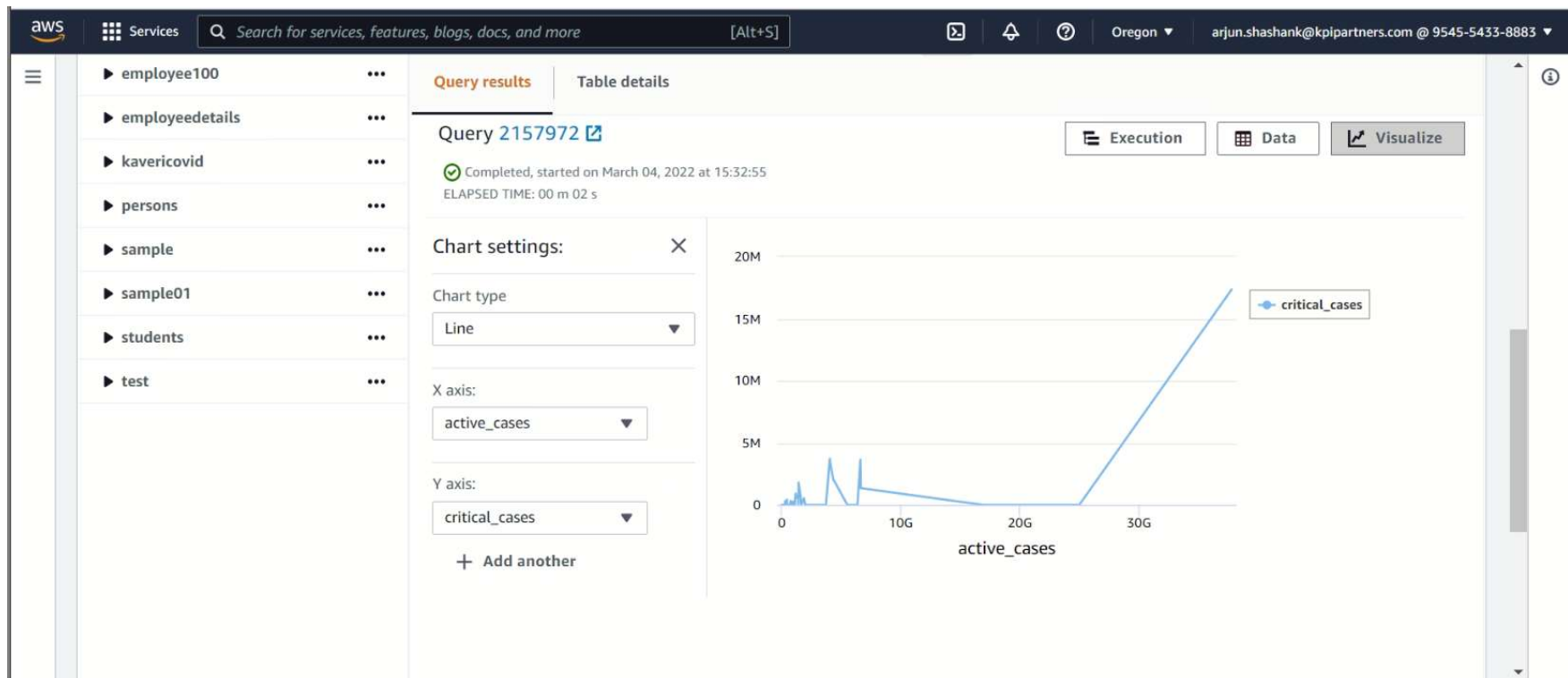
At the bottom of the editor, there are buttons for "Run", "Save", "Schedule", and "Clear". A "Send feedback" link is also present in the bottom right corner.

13. OUTPUT



| | | | |
|--|--|---------------|----------------|
| <div>☰</div> <ul style="list-style-type: none"> ▶ anilrd2 ... ▶ arjcov2 ... ▶ arjun_qstn10 ... ▶ arjun_s_10 ... ▶ arjuntbl1 ... ▶ arjuntbl22 ... ▶ countrybusinessindex ... ▶ covid ... ▶ covid1 ... ▶ covid190001 ... ▶ covid19001 ... ▶ covid_10 ... ▶ covid_13 ... ▶ covid_19 ... ▶ covid_19_india ... | <div>Rows returned (522)</div> <div> <input type="text" value="Search rows"/> <div> < 1 2 3 4 5 6 7 ... 53 > </div> <div>⚙️</div> </div> <div>Export ▼</div> | | |
| | location ▼ | total_cases ▼ | icu_patients ▼ |
| | location | | |
| | Albania | 6.6565288E7 | 0.0 |
| | Austria | 3.55942912E8 | 165958.0 |
| | Ghana | 5.352048E7 | 0.0 |
| | "Isle of Man" | 2719508.0 | 0.0 |
| | Montserrat | 19590.0 | 0.0 |
| | Nepal | 2.7205744E8 | 0.0 |
| | Suriname | 1.318105E7 | 0.0 |
| | Zambia | 6.926164E7 | 0.0 |
| | location | | |
| | | | |
| | | | |
| | | | |

13. active cases vs critical cases





THANK YOU