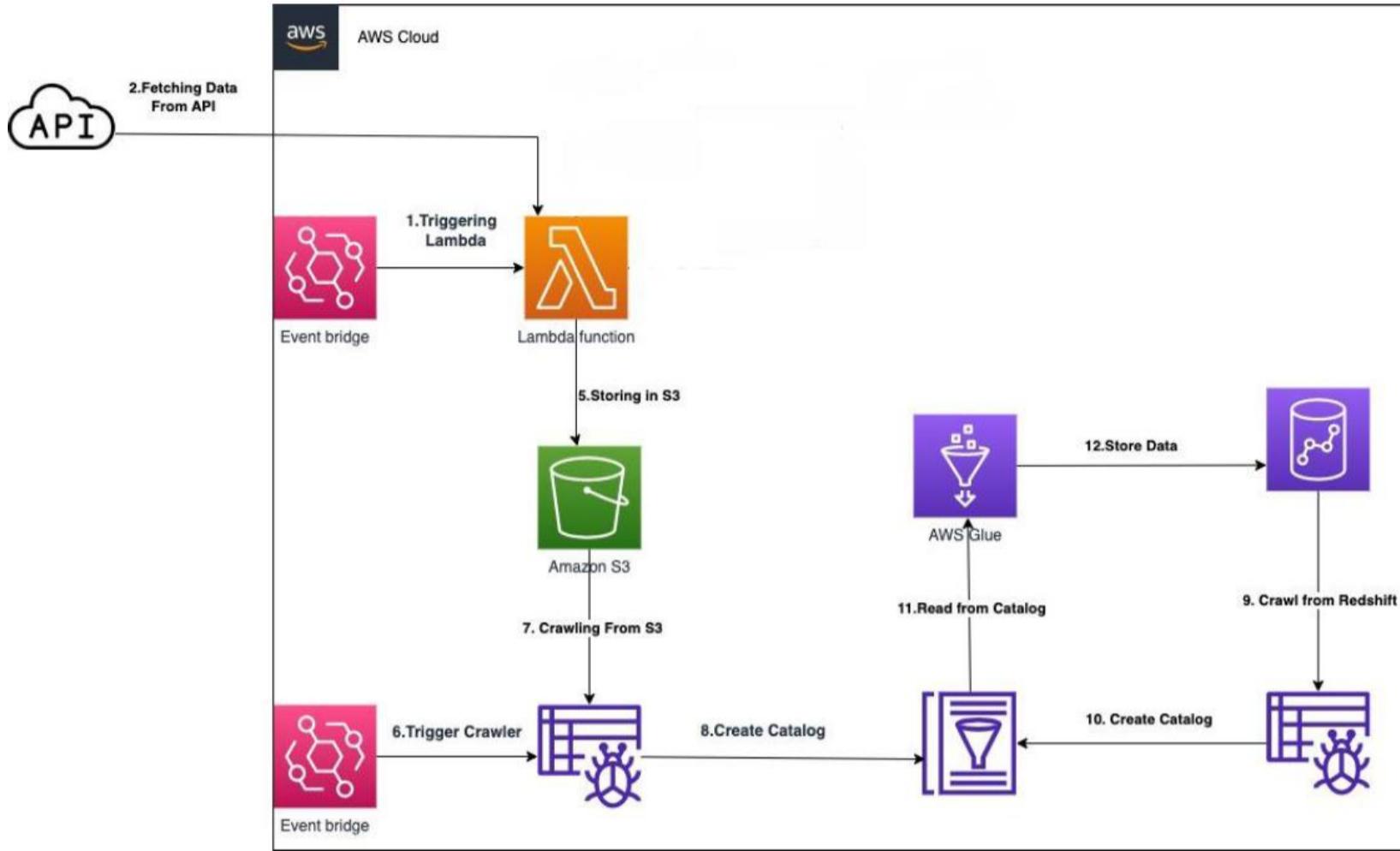


ETL PIPELINE IN AWS CLOUD

1 Introduction

The Aim of the project is to perform ETL Pipeline in AWS using different services of AWS.

Architecture



Explanation:

First, we will extract the data from the API (<http://universities.hipolabs.com/search?country=United+Kingdom>) and dumping data in s3 by AWS lambda using python request while dumping data into s3 configured AWS event bridge to dump data for every 2 min with the help of scheduler.

Next, will create a two individual **crawlers** one for s3 bucket and other for redshift cluster, and we run both crawlers to get schema catalogue for **university.csv** in s3 bucket

To make connect with redshift cluster and crawler need to create **connector** for redshift cluster and by using **JDBC** connection of redshift cluster to get schema catalogue in glue.

Now we need to create a database in AWS redshift and create a table (the table should have equal columns and schema similar to the file in the AWS S3 bucket), We now need to use the crawler to crawl the schema of the table, which was created in AWS Redshift, Now we run crawler to get schema catalogue from AWS redshift cluster

Note that the schemas of both (i.e., one from the S3 file and one from the AWS redshift table) must have equal columns and types to perform data mapping and to store data.

Finally, we create workflow in glue, we will be using AWS event bridge.

Implementation

Created bucket with name **rapidapidata**

The screenshot shows the AWS S3 console interface. The top navigation bar has tabs for AWS Management Console, rapidapidata - S3 bucket, rapidapidata1 - Lambda, Amazon EventBridge Sched, AWS Glue Console, Redshift, and others. The main title is "rapidapidata". Below it, there are tabs for Objects, Properties, Permissions, Metrics, Management, and Access Points. The "Objects" tab is selected. A sub-header "Objects (0)" is displayed. A message states, "Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more." Below this are buttons for Copy S3 URI, Copy URL, Download, Open, Delete, Actions (with a dropdown), Create folder, and Upload. A search bar says "Find objects by prefix". A table header includes columns for Name, Type, Last modified, Size, and Storage class. A message "No objects" indicates "You don't have any objects in this bucket." A large "Upload" button is at the bottom. The footer contains links for Feedback, Language, and copyright information: "© 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences".

Create lambda function with name **rapidapidata1**

The screenshot shows the AWS Lambda console interface. The top navigation bar has tabs for AWS Management Console, rapidapidata - S3 bucket, rapidapidata1 - Lambda, Amazon EventBridge Sched, AWS Glue Console, Redshift, and others. The main title is "rapidapidata1". Below it, there are tabs for Functions and rapidapidata1. The "Functions" tab is selected. A sub-header "Function overview" is shown. The function name is "rapidapidata1". It has "Layers" (2) listed. Triggers include "EventBridge (CloudWatch Events)". There are buttons for Throttle, Copy ARN, and Actions. To the right, there are sections for Description (empty), Last modified (3 minutes ago), Function ARN (arn:aws:lambda:ap-northeast-1:107838789361:function:rapidapidata1), and Function URL (info). At the bottom, there are tabs for Code, Test, Monitor, Configuration, Aliases, and Versions. The "Code" tab is selected. A "Code source" section with an "Upload from" button is shown. The footer contains links for Feedback, Language, and copyright information: "© 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences".

AWS Management Console | rapidapidata - S3 bucket | rapidapidata1 - Lambda | Amazon EventBridge Sched | AWS Glue Console | Redshift | Paused

File Edit View Go Tools Window Test Deploy

Go Anything (Ctrl-P) lambda_function Execution results

```

1 import json
2 import os
3 import boto3
4 import requests
5 import pandas as pd
6 from pandas import DataFrame
7 from io import StringIO
8
9
10 BUCKET_NAME=os.getenv("BUCKET_NAME") #Bucket name mention in environmental variable
11 DATA_SOURCE_URL = "http://universities.hipolabs.com/search?country=United+Kingdom"
12
13
14 def lambda_handler(event, context):
15     print(event,context)
16
17     url = DATA_SOURCE_URL
18     data = requests.get(url)
19     data1 = data.json()
20
21     data2 = pd.DataFrame(data1) #Converting json data into data frame
22     data3 = data2.apply(lambda x: x.explode() if x.name in ['domains', 'web_pages'] else x) #Applying explode function to explode list values by using lamdba
23     data3.rename(columns = {'state-province':'state_province'},inplace=True) #Renaming column for easing mapping purpose in glue and redshift
24
25     csv_buffer = StringIO()
26     data3.to_csv(csv_buffer, index=False)
27     s3_resource = boto3.resource('s3')
28     s3_resource.Object(BUCKET_NAME, 'University.csv').put(Body=csv_buffer.getvalue())
29     print("Data Write Successfull in S3")
30
31
32     return {
33         #'statusCode': 200,
34         'body': 'hello' #returning the response (i.e, printing the out as json list)
35     }
36
37 }
```

Feedback Language © 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences

Adding layers to lambda function

AWS Management Console | rapidapidata - S3 bucket | rapidapidata1 - Lambda | Amazon EventBridge Sched | AWS Glue Console | Redshift | Paused

File Edit View Go Tools Window Test Deploy

Runtime settings Info Edit Edit runtime management configuration

Runtime Python 3.7	Handler Info lambda_function.lambda_handler	Architecture Info x86_64
-----------------------	--	-----------------------------

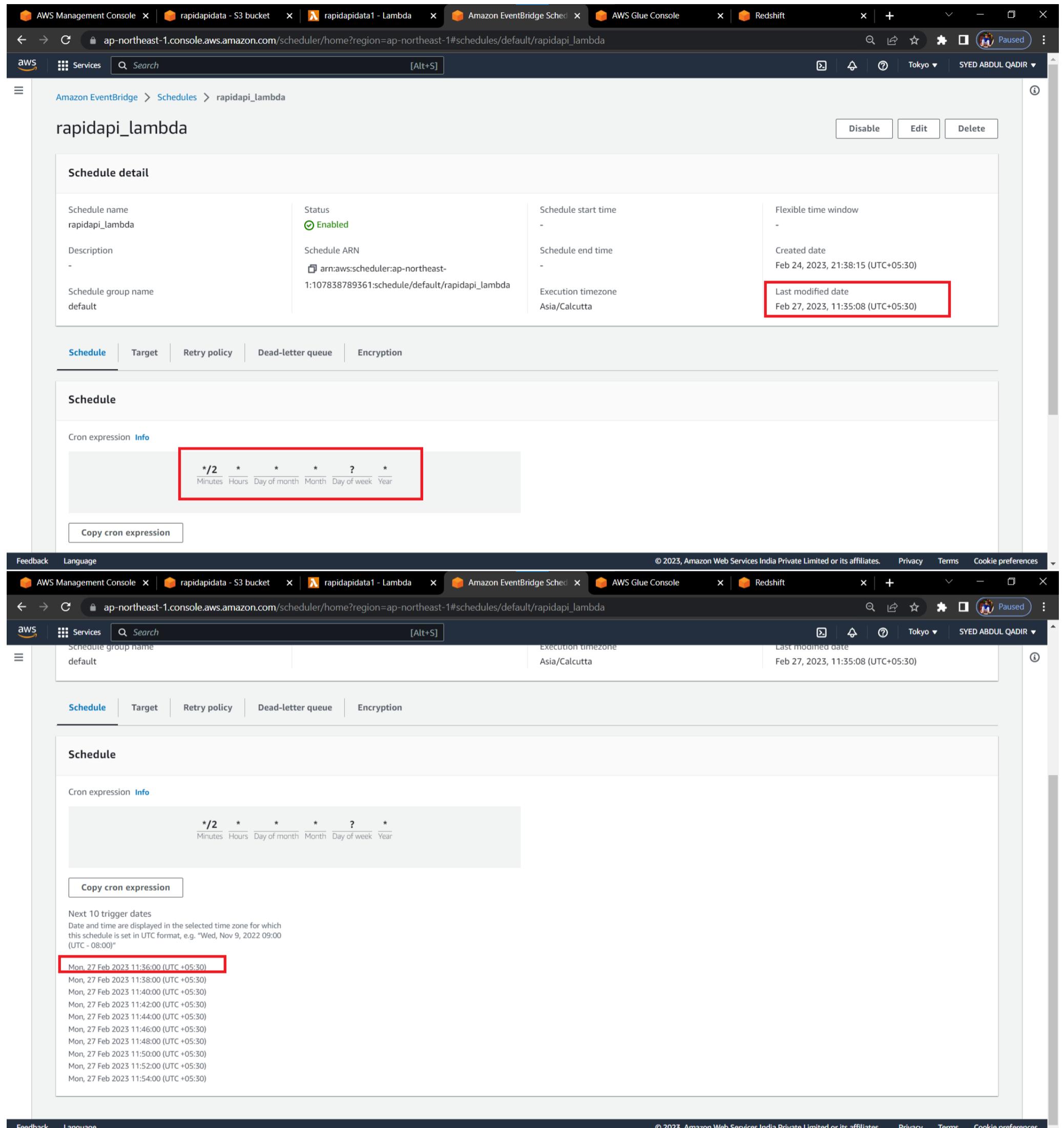
▶ Runtime management configuration

Layers Info

Merge order	Name	Layer version	Compatible runtimes	Compatible architectures
1	pandas_layer	1	python3.7	x86_64
2	AWSLambda-Python37-SciPy1x	118	python3.7	-

Feedback Language © 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences

In Event Bridge creating Schedule to dump data for every two minutes by using cron expression.



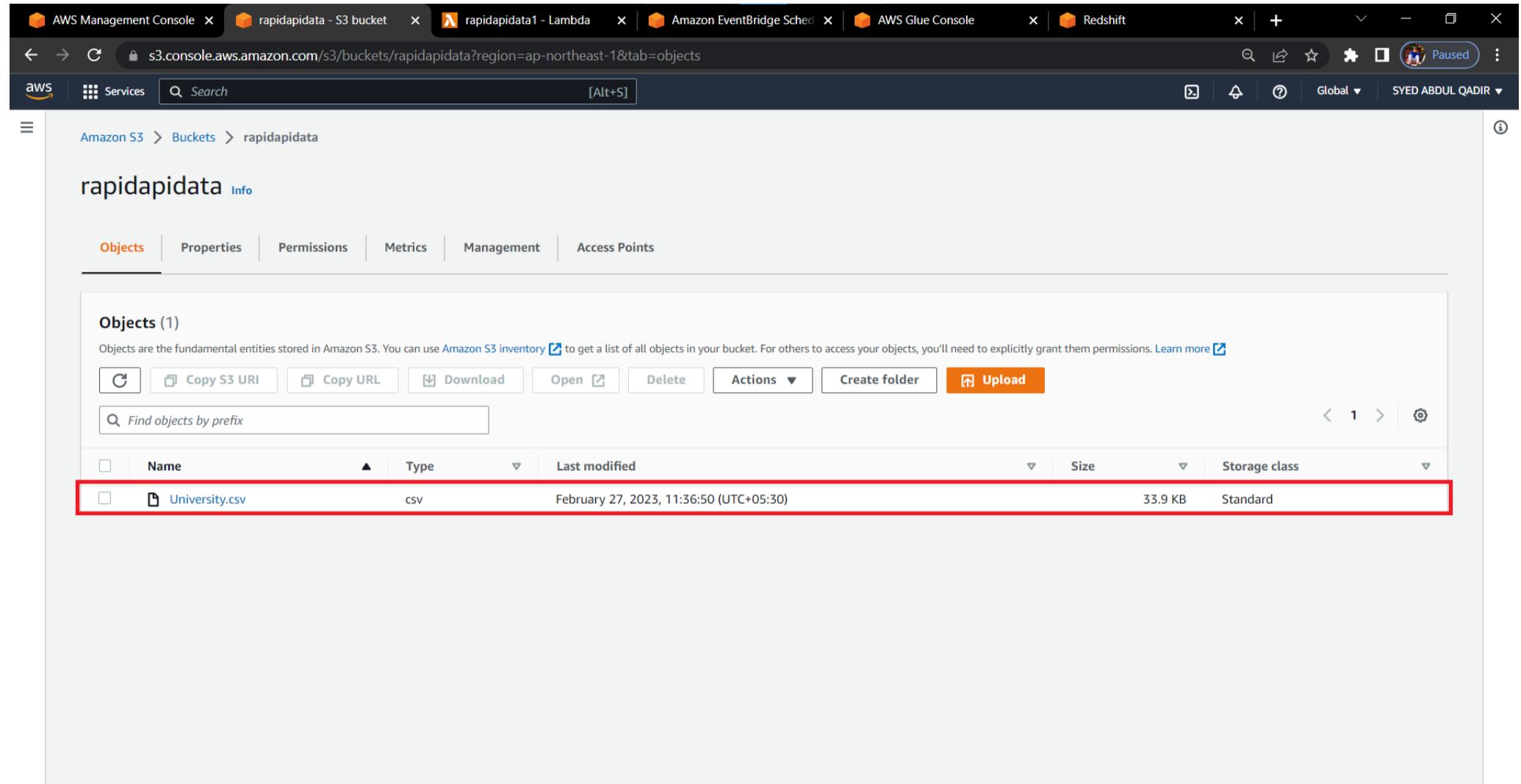
The screenshot shows the AWS EventBridge Schedule configuration page for a schedule named "rapidapi_lambda". The "Schedule detail" section displays the following information:

Schedule name	Status	Schedule start time	Flexible time window
rapidapi_lambda	Enabled	-	-
Description	Schedule ARN	Schedule end time	Created date
-	arn:aws:scheduler:ap-northeast-1:107838789361:schedule/default/rapidapi_lambda	-	Feb 24, 2023, 21:38:15 (UTC+05:30)
Schedule group name	Execution timezone	Last modified date	
default	Asia/Calcutta	Feb 27, 2023, 11:35:08 (UTC+05:30)	

The "Schedule" tab is selected, showing the cron expression `*/2 * * * ? *` (Minutes, Hours, Day of month, Month, Day of week, Year) highlighted with a red box. A "Copy cron expression" button is visible below the cron field.

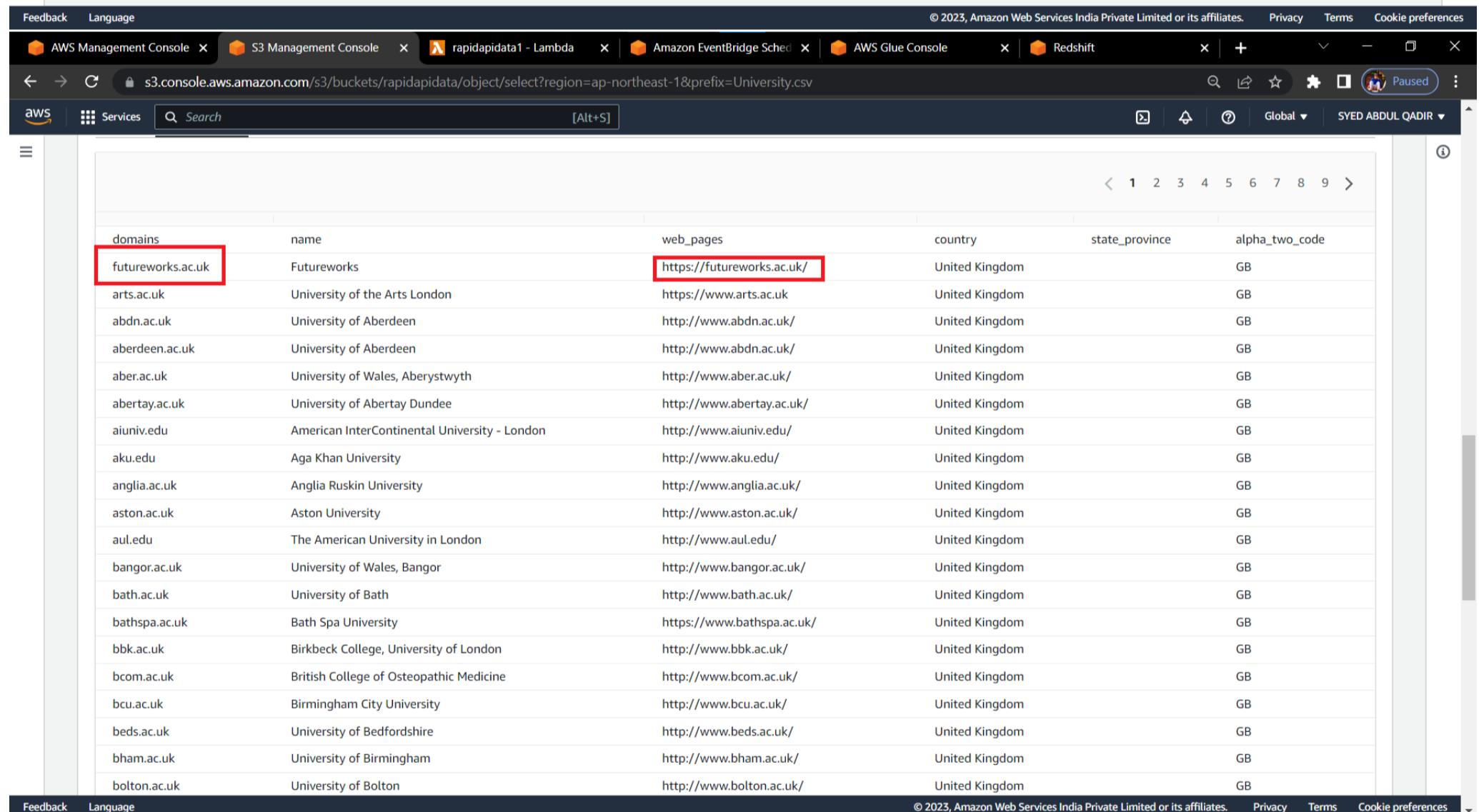
At the bottom of the page, the "Feedback", "Language", and "Cookie preferences" buttons are visible, along with the copyright notice "© 2023, Amazon Web Services India Private Limited or its affiliates." and links for "Privacy", "Terms", and "Cookie preferences".

Fetch data from Open source API to AWS s3 bucket as university.csv



The screenshot shows the AWS S3 console with the 'Objects' tab selected. The bucket 'rapidapidata' contains one object, 'University.csv', which is highlighted with a red border. The object details are as follows:

Name	Type	Last modified	Size	Storage class
University.csv	csv	February 27, 2023, 11:36:50 (UTC+05:30)	33.9 KB	Standard



domains	name	web_pages	country	state_province	alpha_two_code
futureworks.ac.uk	Futureworks	https://futureworks.ac.uk/	United Kingdom		GB
arts.ac.uk	University of the Arts London	https://www.arts.ac.uk	United Kingdom		GB
abdn.ac.uk	University of Aberdeen	http://www.abdn.ac.uk/	United Kingdom		GB
aberdeen.ac.uk	University of Aberdeen	http://www.abdn.ac.uk/	United Kingdom		GB
aber.ac.uk	University of Wales, Aberystwyth	http://www.aber.ac.uk/	United Kingdom		GB
abertay.ac.uk	University of Abertay Dundee	http://www.abertay.ac.uk/	United Kingdom		GB
aiuniv.edu	American InterContinental University - London	http://www.aiuniv.edu/	United Kingdom		GB
aku.edu	Aga Khan University	http://www.aku.edu/	United Kingdom		GB
anglia.ac.uk	Anglia Ruskin University	http://www.anglia.ac.uk/	United Kingdom		GB
aston.ac.uk	Aston University	http://www.aston.ac.uk/	United Kingdom		GB
aul.edu	The American University in London	http://www.aul.edu/	United Kingdom		GB
bangor.ac.uk	University of Wales, Bangor	http://www.bangor.ac.uk/	United Kingdom		GB
bath.ac.uk	University of Bath	http://www.bath.ac.uk/	United Kingdom		GB
bathspa.ac.uk	Bath Spa University	https://www.bathspa.ac.uk/	United Kingdom		GB
bbk.ac.uk	Birkbeck College, University of London	http://www.bbk.ac.uk/	United Kingdom		GB
bcom.ac.uk	British College of Osteopathic Medicine	http://www.bcom.ac.uk/	United Kingdom		GB
bcu.ac.uk	Birmingham City University	http://www.bcu.ac.uk/	United Kingdom		GB
beds.ac.uk	University of Bedfordshire	http://www.beds.ac.uk/	United Kingdom		GB
bham.ac.uk	University of Birmingham	http://www.bham.ac.uk/	United Kingdom		GB
bolton.ac.uk	University of Bolton	http://www.bolton.ac.uk/	United Kingdom		GB

Created crawler for s3 bucket to get schema metadata catalogue for university.csv, started crawler to run.

Crawler properties

Name s3_crawler	IAM role crawler_gluefullaccess	Database redshift	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			
Advanced settings			

Crawler runs (4)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
February 27, 2023 at 06:09:40	-	06 s	Running	-	-
February 25, 2023 at 06:47:37	February 25, 2023 at 06:48:22	44 s	Completed	0.082	1 table change, 0 partition changes
February 22, 2023 at 12:33:17	February 22, 2023 at 12:33:56	39 s	Completed	0.080	1 table change, 0 partition changes
February 22, 2023 at 12:27:16	February 22, 2023 at 12:28:05	48 s	Completed	0.068	1 table change, 0 partition changes

Crawler properties

Name s3_crawler	IAM role crawler_gluefullaccess	Database redshift	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			
Advanced settings			

Crawler runs (4)

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
February 27, 2023 at 06:09:40	February 27, 2023 at 06:10:23	42 s	Completed	-	1 table change, 0 partition changes
February 25, 2023 at 06:47:37	February 25, 2023 at 06:48:22	44 s	Completed	0.082	1 table change, 0 partition changes
February 22, 2023 at 12:33:17	February 22, 2023 at 12:33:56	39 s	Completed	0.080	1 table change, 0 partition changes
February 22, 2023 at 12:27:16	February 22, 2023 at 12:28:05	48 s	Completed	0.068	1 table change, 0 partition changes

Able to verify schema metadata in database section Tables.

The screenshot shows the AWS Glue Console interface. The top navigation bar includes tabs for AWS Management Console, S3 Management Console, Lambda, Amazon EventBridge Sched, AWS Glue Console, Redshift, and other services. The main content area displays the 'rapidapidata' table under the 'Tables' section. The 'Table overview' tab is selected, showing basic information like Name (rapidapidata), Location (s3://rapidapidata/), Input format (org.apache.hadoop.mapred.TextInputFormat), and Output format (org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat). The 'Database' is listed as redshift and 'Classification' as csv. The 'Schema' tab is also visible, showing a table structure with 6 columns: domains, name, web_pages, country, state_province, and alpha_two_code, all defined as string type.

Created connector to make connection for AWS redshift cluster.

The screenshot shows the AWS Glue Studio interface. The top navigation bar includes tabs for AWS Management Console, S3 Management Console, Lambda, Amazon EventBridge, AWS Glue Console, Glue Studio, Redshift, and other services. The main content area displays the 'redshift_cluster_connection' connector under the 'Connectors' section. The 'Connection details' tab is selected, showing the following configuration: Connector type (JDBC), Connection URL (jdbc:redshift://redshift-cluster-1.c41vabe747w1.ap-northeast-1.redshift.amazonaws.com:5439/dev), Username (awsuser), Subnet (subnet-0f7723df21681f75d), Security groups (sg-0518e0f846f9ca087), and Created on (2023-02-25 14:55:53.102000). Below this, there is a section for 'Your jobs (4)' with buttons for Edit, Delete, and Run job.

Created AWS Redshift cluster with free trial version configuration as redshift-cluster-1.

The screenshot shows the AWS Redshift console with the cluster details page for 'redshift-cluster-1'. The 'Properties' tab is selected. Key information includes:

- General information:** Cluster identifier: redshift-cluster-1, Status: Available, Date created: February 25, 2023, 12:24 (UTC+05:30), Storage used: 0.23% (0.36 of 160 GB used), Node type: dc2.large, Number of nodes: 1, Multi-AZ: No.
- Database configurations:** Database name: [REDACTED], Port: [REDACTED], Admin user name: [REDACTED]. Parameter group: [REDACTED] (Defines database parameter and query queues for all the databases). SSH ingestion setting (cluster public key): [REDACTED]. Encryption: Disabled, AWS KMS key ID: -.
- Network and security settings:** Virtual private cloud (VPC): [REDACTED], Subnet: [REDACTED]. Availability Zone: ap-northeast-1c, Enhanced VPC routing: Enabled. VPC security group: sg-0518e0f846f9ca087 (Specified in the screenshot). Publicly accessible: Disabled.

For **VPC security group** redshift cluster attaching Type **All TCP, All traffic**.

The screenshot shows the AWS VPC Management console with the security group details page for 'sg-0518e0f846f9ca087 - default'. The 'Inbound rules' tab is selected. Two rules are listed:

Name	Security group rule...	Type	Protocol	Port range	Source	Description
-	[REDACTED]	All TCP	TCP	0 - 65535	[REDACTED]	-
-	[REDACTED] 0.0.0.0/0	IPv4	All traffic	All	0.0.0.0/0	-

Attaching endpoints to s3 bucket rapidapidata with name my-endpoint-01.

The screenshot shows the AWS VPC Endpoints console. On the left, there's a sidebar with various VPC-related options like 'Virtual private cloud', 'Endpoints', 'Security', 'Network Analysis', 'DNS firewall', etc. The main area displays a table titled 'Endpoints (1/1)'. It has columns for Name, VPC endpoint ID, VPC ID, Service name, Endpoint type, and Status. One row is selected, showing 'my-endpoint-01' with a redacted VPC endpoint ID, a redacted VPC ID, 'com.amazonaws.ap-northeast-1.s3' as the service name, 'Gateway' as the endpoint type, and 'Available' as the status. Below the table, there's a detailed view for 'vpce-056c02a65d5ad7442 / my-endpoint-01' with tabs for Details, Route tables, Policy, and Tags. The 'Details' tab shows information such as Endpoint ID (redacted), Status (Available), Creation time (Wednesday, February 22, 2023 at 13:00:47 GMT+5:30), Service name (com.amazonaws.ap-northeast-1.s3), and Endpoint type (Gateway). The status message is '-'.

Created Table in redshift cluster as University_lists.

The screenshot shows the AWS Redshift query editor v2. On the left, there's a sidebar with 'Editor', 'Queries', 'Notebooks', 'Charts', and 'History'. The main area has a 'Redshift query editor v2' title bar with tabs for 'Create', 'Load data', 'Run' (highlighted in red), 'Limit 100', 'Explain', 'Isolated session', 'redshift-cluster-1' (selected database), and 'dev' (schema). Below this is a code editor window containing the following SQL code:

```
1
2
3 CREATE TABLE University_lists (
4 domains VARCHAR(255),
5 name VARCHAR(255),
6 web_pages VARCHAR(255),
7 country VARCHAR(255),
8 state_province INT,
9 alpha_two_code VARCHAR(255)
10 );
11 select * from University_lists;
```

Below the code editor is a 'Result 1' table with columns: domains, name, web_pages, country, state_province, and alpha_two_code. A message 'No Rows To Show' is displayed. At the bottom right, it says 'Elapsed time: 133 ms 0'.

Created crawler pointing to redshift cluster for **University_lists** table to obtained schema catalogue in glue with help of connector of **redshift_cluster_connection**, with crawler name **redshift_crawler**.

Crawler get successfully executed.

The screenshot shows the AWS Glue Crawler properties and runs page for a crawler named "redshift_crawler".

Crawler properties:

Name	IAM role	Database	State
redshift_cralwer	crawler_gluefullaccess	redshift	READY

Crawler runs (8):

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
February 27, 2023 at 06:18:13	February 27, 2023 at 06:20:22	02 min 08 s	Completed	-	1 table change, 0 partition changes
February 25, 2023 at 11:27:15	February 25, 2023 at 11:29:23	02 min 07 s	Completed	0.238	-
February 25, 2023 at 11:18:43	February 25, 2023 at 11:21:37	02 min 53 s	Completed	0.308	-
February 25, 2023 at 10:50:35	February 25, 2023 at 10:52:39	02 min 04 s	Completed	0.212	1 table change, 0 partition changes
...					

Crawler runs (8):

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
February 27, 2023 at 06:18:13	-	05 s	Running	-	-
...					

Able to fetch schema from redshift cluster.

The screenshot shows the AWS Glue Table Overview page for the table 'dev_public_university_lists'. The table details are as follows:

Name	Description	Database	Classification
dev_public_university_lists	-	redshift	redshift
Location	Connection	Deprecated	Last updated
dev.public.university_lists	redshift_cluster_connection	-	February 27, 2023 at 06:20:22
Input format	Output format	Serde serialization lib	-
-	-	-	

The Schema tab is selected, displaying the table schema with 6 columns:

#	Column name	Data type	Partition key	Comment
1	country	string	-	-
2	web_pages	string	-	-
3	name	string	-	-
4	domains	string	-	-
5	state_province	int	-	-
6	alpha_two_code	string	-	-

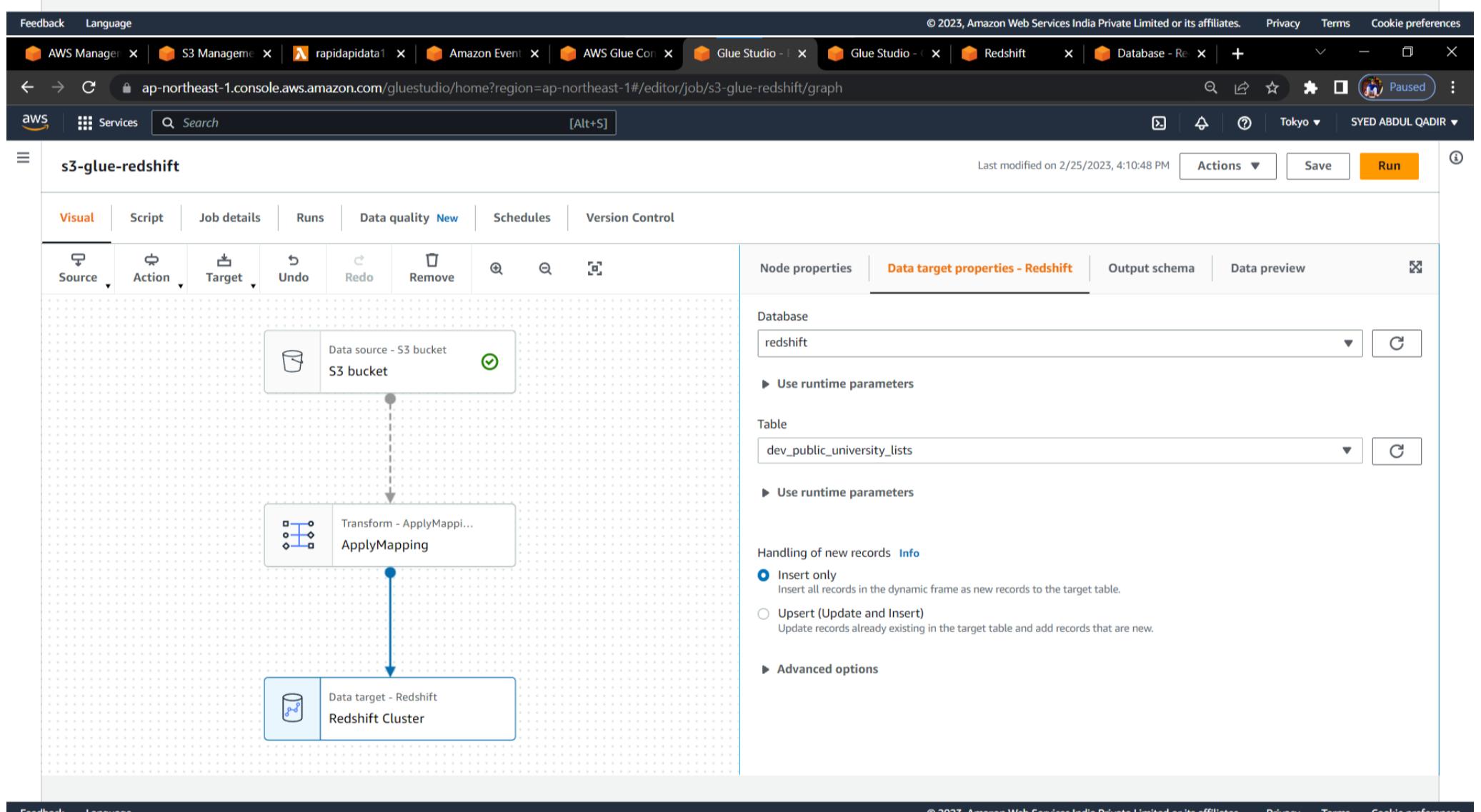
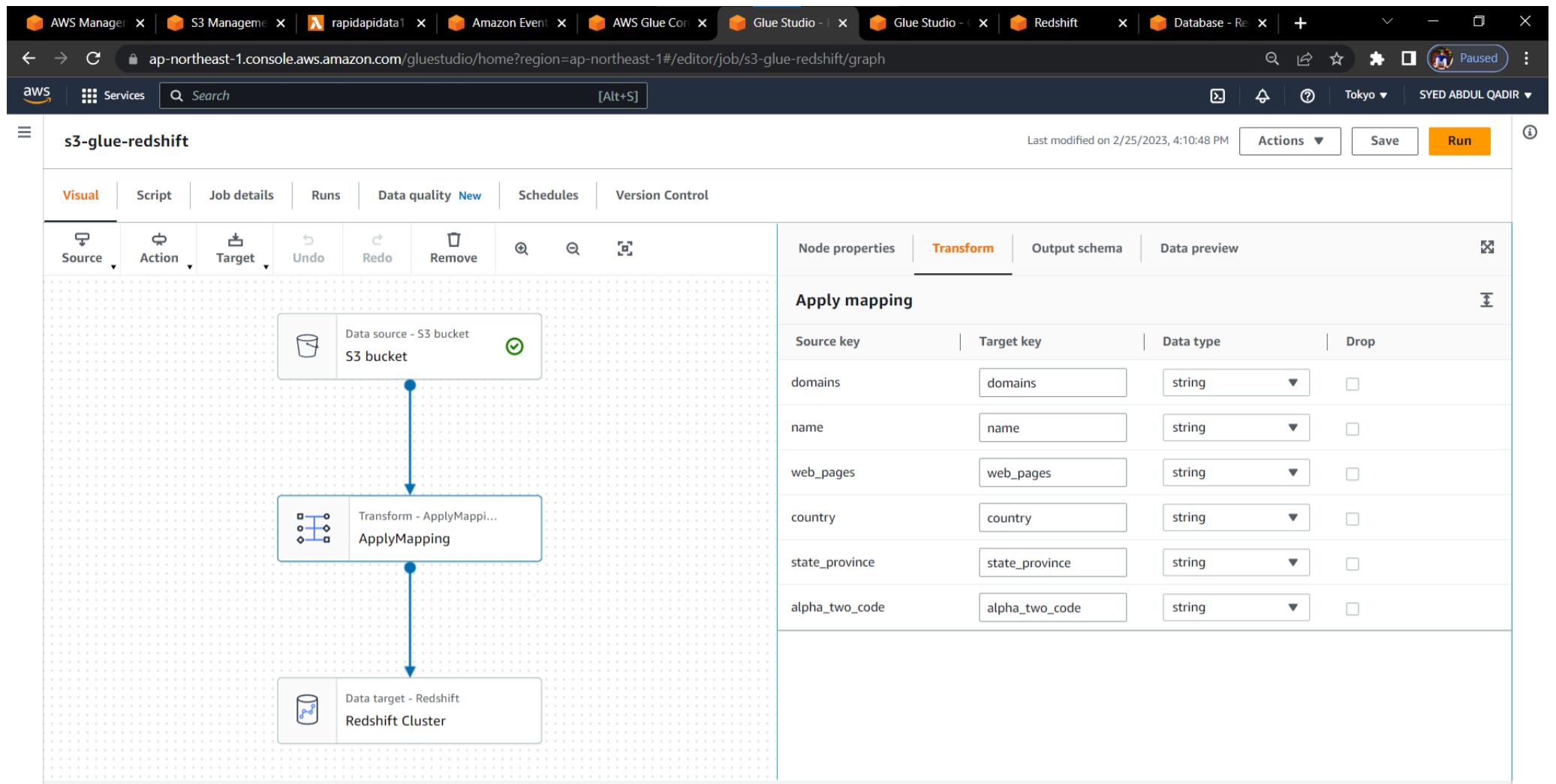
Now run glue job with help visual source:s3 with data catalogue of s3 schema and target would be Redshift cluster.

The screenshot shows the AWS Glue Studio visual editor for a job named 's3-glue-redshift'. The pipeline consists of three main components:

- Source:** Data source - S3 bucket, S3 bucket
- Action:** Transform - ApplyMapping, ApplyMapping
- Target:** Data target - Redshift, Redshift Cluster

The 'Data source properties - S3' tab is selected, showing the configuration for the S3 source:

- S3 source type: Data Catalog table (selected)
- S3 location: Choose a file or folder in an S3 bucket.
- Database: redshift
- Table: rapidapidata
- Partition predicate - optional: Enter a boolean expression supported by Spark SQL, using only partition columns.



Created workflow in glue job to trigger job.

The screenshot shows the AWS Glue console with the 'Workflows' view. A single workflow named 'workflow_1' is listed. Below it, the 'History' tab is selected, showing a single run entry. The run ID is 'wr_43be5643ecd9667789acfb1c659faca8c861b...', the status is 'Running', and it started on 'Mon, 27 Feb 2023 06:58:35 GMT'. The 'Run ID' column is highlighted with a red box.

Created AWS Event Bridge of rule to trigger the workflow created in glue to run it by every two minutes.

The screenshot shows the AWS EventBridge console with the 'Edit rule' step selected. In the 'Select target(s)' section, the 'Target types' dropdown is set to 'AWS service'. Under 'Glue workflow', the name 'workflow_1' is selected. The 'Execution role' dropdown shows 'Amazon_EventBridge_Invoke_Glue_1696339090'. The 'Next' button at the bottom right is highlighted with a red box.

AWS Manager | S3 Management | rapidapidata1 | Amazon Event | AWS Glue Core | Glue Studio | Glue Studio - | Redshift | Database - | + | Paused | Services | Search | [Alt+S] | AWS | Tokyo | SYED ABDUL QADIR | Rule trigger-workflow was updated successfully

Amazon EventBridge > Rules > trigger-workflow

trigger-workflow

Rule details [Info](#)

Rule name	trigger-workflow	Status	Enabled
Description		Event bus name	default
		Event bus ARN	arn:aws:events:ap-northeast-1:107838789361:rule/trigger-workflow
		Type	Standard

[Edit](#) [Disable](#) [Delete](#) [CloudFormation Template](#)

[Event schedule](#) [Targets](#) [Monitoring](#) [Tags](#)

Event schedule [Info](#)

Cron expression: */2 * * * ? *

Next 10 trigger date(s):

- Mon, Feb 27, 2023, 12:28 PM GMT+5:30
- Mon, Feb 27, 2023, 12:30 PM GMT+5:30
- Mon, Feb 27, 2023, 12:32 PM GMT+5:30
- Mon, Feb 27, 2023, 12:34 PM GMT+5:30
- Mon, Feb 27, 2023, 12:36 PM GMT+5:30
- Mon, Feb 27, 2023, 12:38 PM GMT+5:30
- Mon, Feb 27, 2023, 12:40 PM GMT+5:30
- Mon, Feb 27, 2023, 12:42 PM GMT+5:30
- Mon, Feb 27, 2023, 12:44 PM GMT+5:30
- Mon, Feb 27, 2023, 12:46 PM GMT+5:30

Local time zone: ▾

Feedback Language © 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences

Trigger-1 with type: event bridge.

AWS Manager | S3 Management | rapidapidata1 | Amazon Event | AWS Glue Core | Glue Studio | Glue Studio - | Redshift | Database - | + | Paused | Services | Search | [Alt+S] | AWS | Tokyo | SYED ABDUL QADIR | Workflows (1)

A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.

Add workflow Actions Filter workflows < 1 >

Name	Last run	Last run status	Last modified
workflow_1	-	-	Mon, 27 Feb 2023 06:25:41 GMT

Graph Details History Legend: ● Start ◆ Trigger □ Job ■ Crawler ⚡ Incomplete ✘ Error ✎ Deleting Remove Action

```

graph LR
    T1[Trigger:Trigger-1 Type: EVENT] --> R1[redshift_crawler]
    R1 --> ANY1{ANY}
    ANY1 --> S1[s3-glue-redshift]
    R1 --> T2{Trigger-2}
  
```

Feedback Language © 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences

For Trigger-1 attaching crawler of redshift cluster.

Screenshot of the AWS Glue Workflows console showing a workflow named "workflow_1".

Workflow Graph:

```

graph LR
    Start((Start)) --> Trigger1((Trigger-1))
    Trigger1 --> Crawler[redshift_crawler]
    Crawler --> ANY{ANY}
    ANY --> Trigger2((Trigger-2))
    Trigger2 --> S3[s3-glue-redshift]
    
```

Legend:

- Start
- Trigger
- Job
- Crawler
- Incomplete
- Error
- Deleting

Actions:

- Remove
- Action

Trigger-2 with type: by default event.

Screenshot of the AWS Glue Workflows console showing a workflow named "workflow_1".

Workflow Graph:

```

graph LR
    Start((Start)) --> Trigger1((Trigger-1))
    Trigger1 --> Crawler[redshift_crawler]
    Crawler --> ANY{ANY}
    ANY --> Trigger2((Trigger-2))
    Trigger2 --> S3[s3-glue-redshift]
    
```

Trigger-2 Properties:

Type: CONDITIONAL

Legend:

- Start
- Trigger
- Job
- Crawler
- Incomplete
- Error
- Deleting

Actions:

- Remove
- Action

Attaching glue job with name s3-glue-redshift to Trigger-2.

Screenshot of the AWS Glue Workflow service showing the creation of a workflow named "workflow_1". The workflow graph includes a "Trigger-1" node (Start), a "redshift_crawler" node (Job), an "ANY" trigger node, and an "s3-glue-job" node (Job). A pink box highlights the "s3-glue-job" node.

After Trigger-1 succeeded disenable, the rule created at event bridge.

Screenshot of the AWS Glue Workflow service showing the status of a workflow run. The "Trigger-1" node is highlighted with a red box and has a green checkmark indicating it has succeeded. The "redshift_crawler" node is running, and the "Trigger-2" node is waiting. A pink box highlights the "s3-glue-job" node.

Overall workflow will get executed succeeded.

The screenshot shows the AWS Glue Workflow Studio interface. On the left, a sidebar lists various AWS services like AWS Manager, S3 Management, Amazon Event, AWS Glue Con, Glue Studio, Redshift, and Database. The main area displays a workflow named 'workflow_1' with a run ID of 'wr_43be5643ecd9667789acfb1c659faca8c861be7c2c66441593207e74f47f204d'. The run status is 'Completed' with a start time of 'Mon, 27 Feb 2023 06:58:35 GMT' and end time of 'Mon, 27 Feb 2023 07:07:09 GMT'. Below this, a 'Graph' tab is selected, showing a sequence of four nodes: 'Trigger-1' (green circle), 'redshift_crawler' (green square), 'ANY' (green diamond), and 's3-glue-redshift' (green square). Each node has a green checkmark indicating success. A legend at the bottom defines symbols for Succeeded (green checkmark), Running (blue circle), Stopped (grey circle), Failed (red X), Timeout (red X), Error (red X), Warning (yellow triangle), Resume (purple circle), and Not started (grey square).

Glue job with name **s3-glue-redshift** run status **Succeeded**.

The screenshot shows the AWS Glue Studio interface for a job named 's3-glue-redshift'. The 'Runs' tab is selected. It shows a recent job run from 'February 27, 2023 12:35:11 PM'. The 'Run status' is highlighted with a red box and shows 'Succeeded'. Other details for this run include: Job name 's3-glue-redshift', Id 'jr_b5ec15f2cb2e06d5d78001d7c654c8f624d2f846f05f45d8549109070197e1c5', Start time 'February 27, 2023 12:35:11 PM', End time 'February 27, 2023 12:37:09 PM', Trigger name 'Trigger-2', Number of workers '2', and Glue version '3.0'. Below this, another run from 'February 25, 2023 5:00:46 PM' is listed with a similar structure. The interface includes tabs for Visual, Script, Job details, Runs, Data quality, Schedules, and Version Control.

Finally able to dump s3 data into redshift cluster with table name **University_lists**.

The screenshot shows the AWS SQL Workbench interface with a query results table titled "Result 1 (100)". The table has seven columns: domains, name, web_pages, country, state_province, alpha_two_code, and state_province_string. The data consists of 100 rows of university information from the United Kingdom. The "Elapsed time: 38 ms" and "Total rows: 100" are displayed at the bottom right of the results table.

domains	name	web_pages	country	state_province	alpha_two_code	state_province_string
futureworks.ac.uk	Futureworks	https://futureworks.ac.uk/	United Kingdom	NULL	GB	
arts.ac.uk	University of the Arts Lon...	https://www.arts.ac.uk	United Kingdom	NULL	GB	
abdn.ac.uk	University of Aberdeen	http://www.abdn.ac.uk/	United Kingdom	NULL	GB	
aberdeen.ac.uk	University of Aberdeen	http://www.aberdeen.ac.uk/	United Kingdom	NULL	GB	
aber.ac.uk	University of Wales, Aber...	http://www.aber.ac.uk/	United Kingdom	NULL	GB	
abertay.ac.uk	University of Abertay Dun...	http://www.abertay.ac.uk/	United Kingdom	NULL	GB	
aiuniv.edu	American InterContinental...	http://www.aiuniv.edu/	United Kingdom	NULL	GB	
aku.edu	Aga Khan University	http://www.aku.edu/	United Kingdom	NULL	GB	
anglia.ac.uk	Anglia Ruskin University	http://www.anglia.ac.uk/	United Kingdom	NULL	GB	
aston.ac.uk	Aston University	http://www.aston.ac.uk/	United Kingdom	NULL	GB	
aul.edu	The American University i...	http://www.aul.edu/	United Kingdom	NULL	GB	
bangor.ac.uk	University of Wales, Bangor	http://www.bangor.ac.uk/	United Kingdom	NULL	GB	
bath.ac.uk	University of Bath	http://www.bath.ac.uk/	United Kingdom	NULL	GB	
bathspa.ac.uk	Bath Spa University	https://www.bathspa.ac.uk/	United Kingdom	NULL	GB	
bbk.ac.uk	Birkbeck College, Univers...	http://www.bbk.ac.uk/	United Kingdom	NULL	GB	
bcom.ac.uk	British College of Osteop...	http://www.bcom.ac.uk/	United Kingdom	NULL	GB	
bcu.ac.uk	Birmingham City University	http://www.bcu.ac.uk/	United Kingdom	NULL	GB	
beds.ac.uk	University of Bedfordshire	http://www.beds.ac.uk/	United Kingdom	NULL	GB	
bham.ac.uk	University of Birmingham	http://www.bham.ac.uk/	United Kingdom	NULL	GB	
bolton.ac.uk	University of Bolton	http://www.bolton.ac.uk/	United Kingdom	NULL	GB	
bournemouth.ac.uk	Bournemouth University	http://www.bournemouth.ac.uk/	United Kingdom	NULL	GB	
brad.ac.uk	University of Bradford	http://www.brad.ac.uk/	United Kingdom	NULL	GB	
brijnet.org	London School of Jewish ...	http://www.brijnet.org/ljsj/	United Kingdom	NULL	GB	
bristol.ac.uk	University of Bristol	https://www.bristol.ac.uk/	United Kingdom	NULL	GB	
bris.ac.uk	University of Bristol	https://www.bristol.ac.uk/	United Kingdom	NULL	GB	
brookes.ac.uk	Oxford Brookes University	http://www.brookes.ac.uk/	United Kingdom	NULL	GB	
brunel.ac.uk	Brunel University Uxbridge	http://www.brunel.ac.uk/	United Kingdom	NULL	GB	
bton.ac.uk	University of Brighton	http://www.bton.ac.uk/	United Kingdom	NULL	GB	
buck.ac.uk	University of Buckingham	http://www.buck.ac.uk/	United Kingdom	NULL	GB	
bucks.ac.uk	Buckinghamshire New Un...	http://www.bucks.ac.uk/	United Kingdom	NULL	GB	
cam.ac.uk	University of Cambridge	http://www.cam.ac.uk/	United Kingdom	NULL	GB	
canterbury.kcl.ac.uk	University of Kent Canterbury	http://www.canterbury.kcl.ac.uk/	United Kingdom	NULL	GB	

