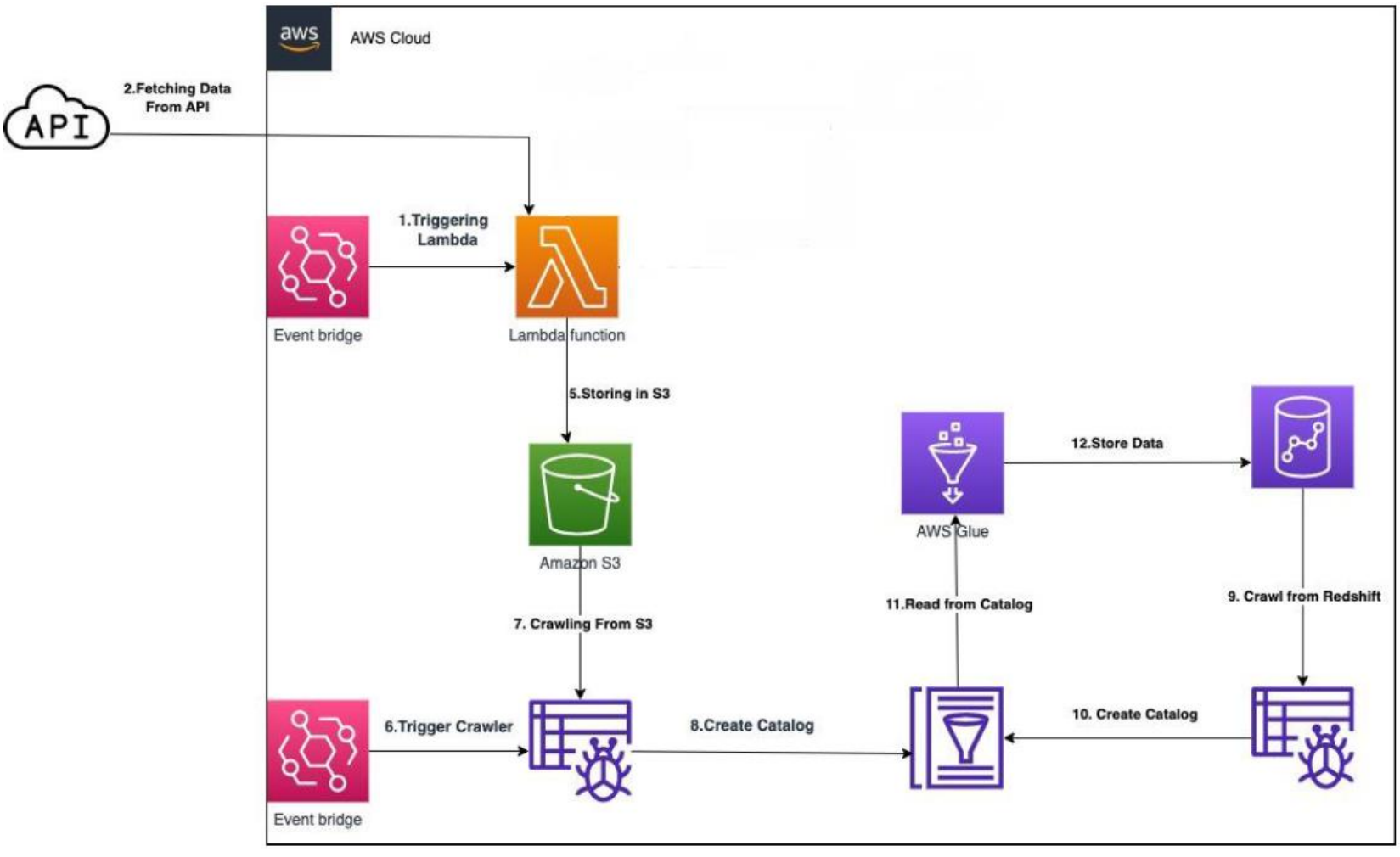
ETL PIPELINE IN AWS CLOUD

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.

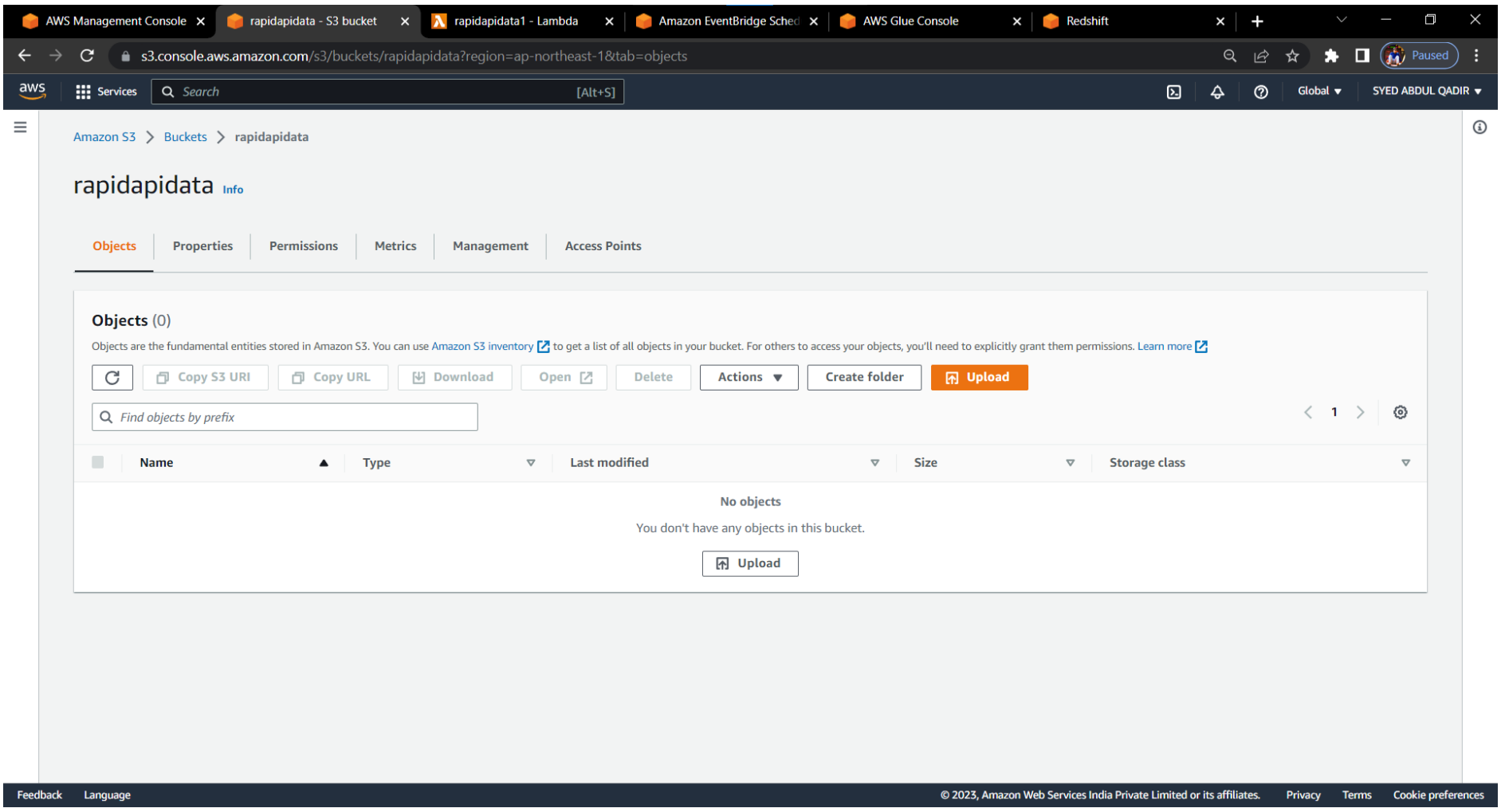
Now we need to create a database in AWS redshift and create a table (the table should have equal columns and schema similar to the file in the AWS S3 bucket),We now need to use the crawler to crawl the schema of the table, which was created in AWS Redshift, Now we run crawler to get schema catalogue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

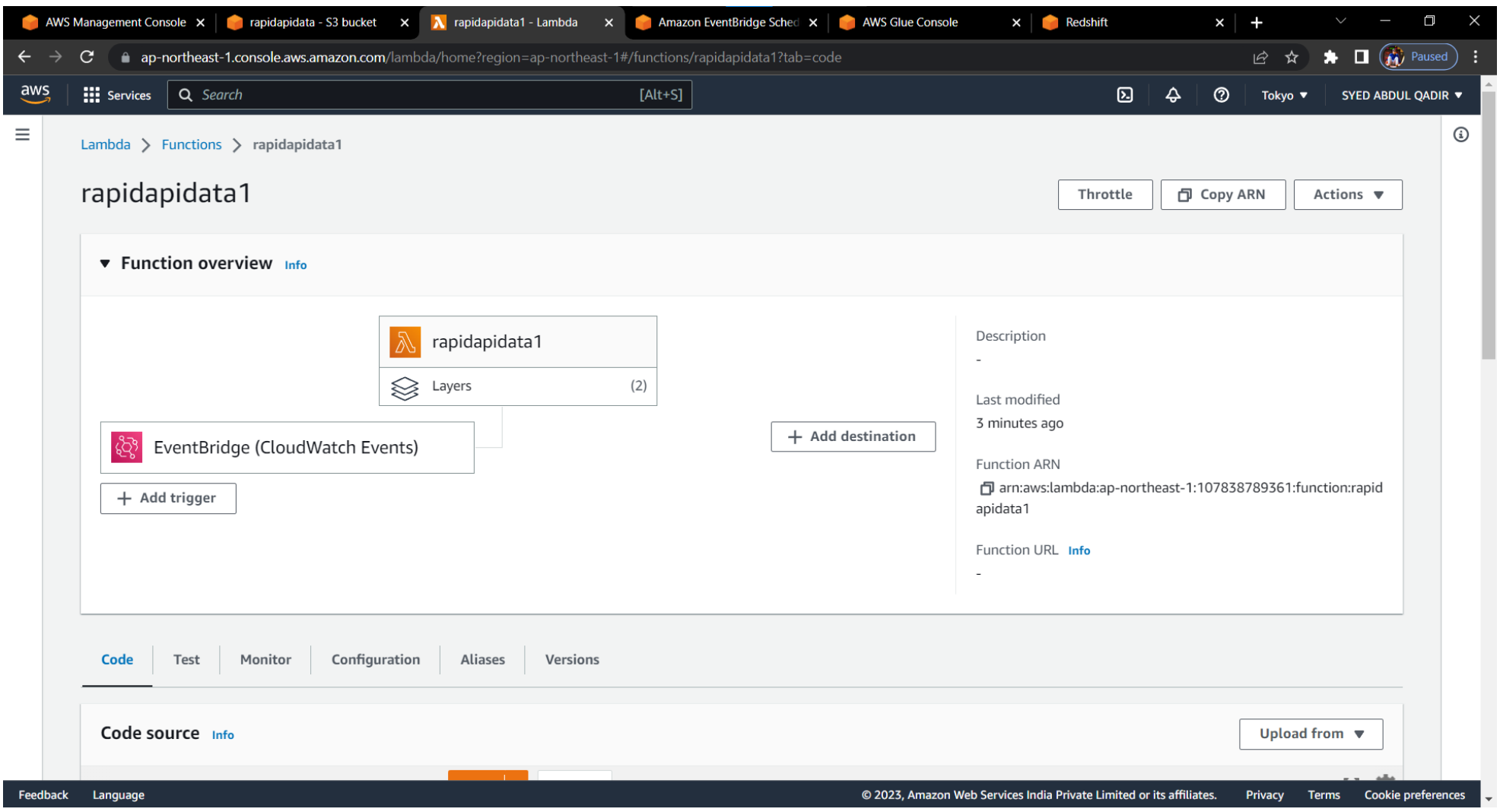
Finally, we create workflow in glue, we will be using AWS event bridge.

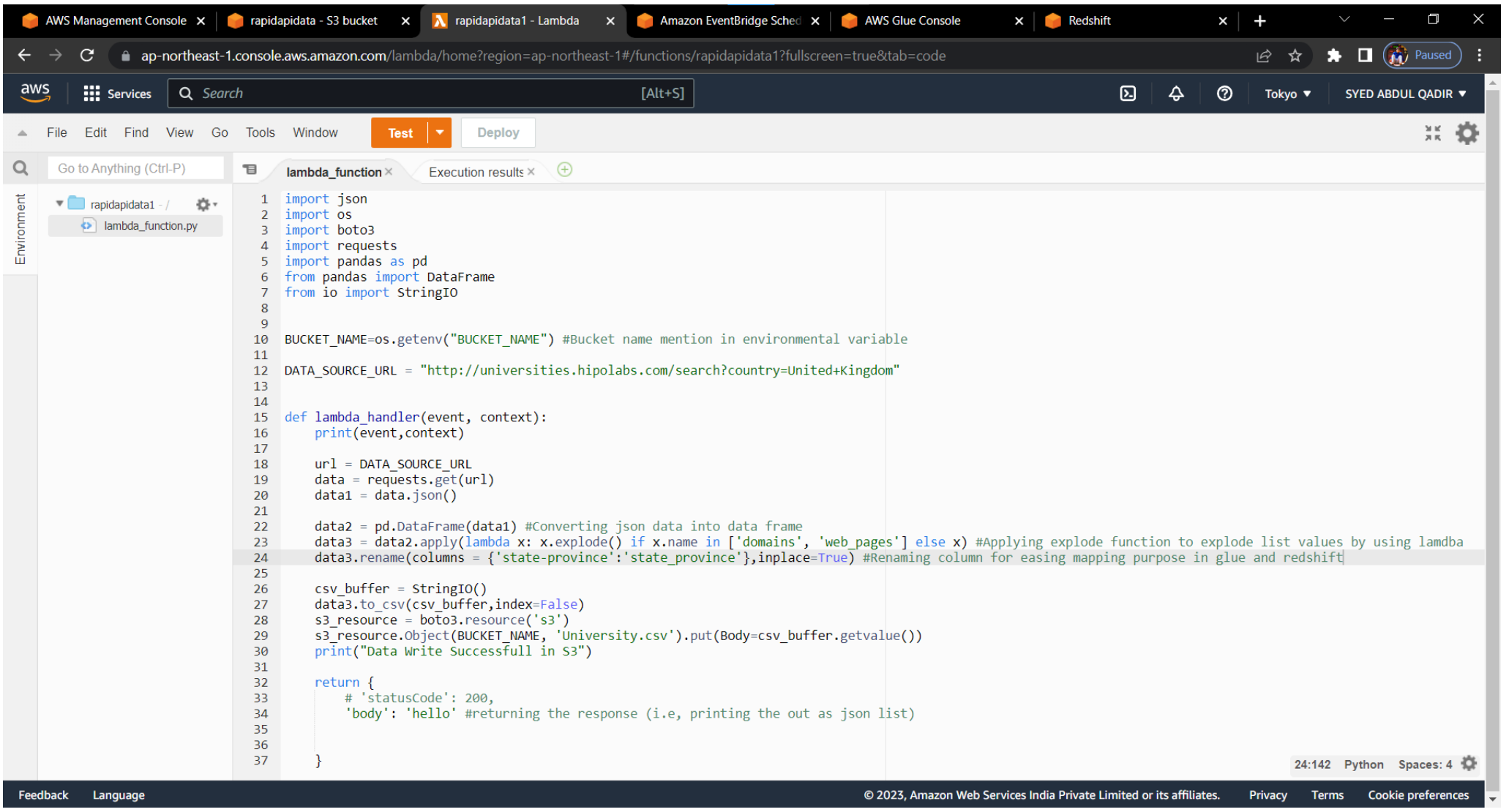
Implementation

Created bucket with name **rapidapidata**

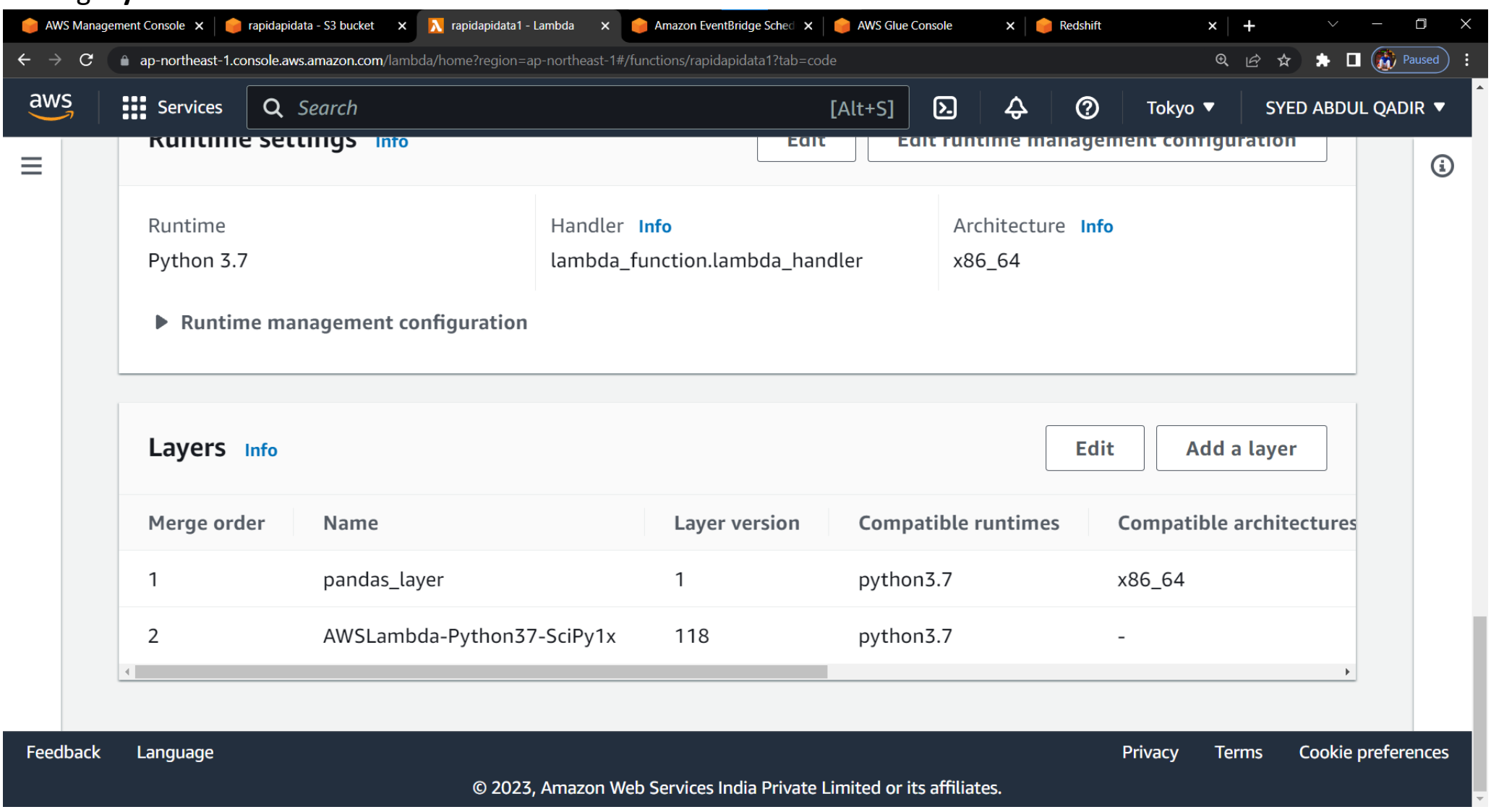


Create lambda function with name **rapidapidata1**

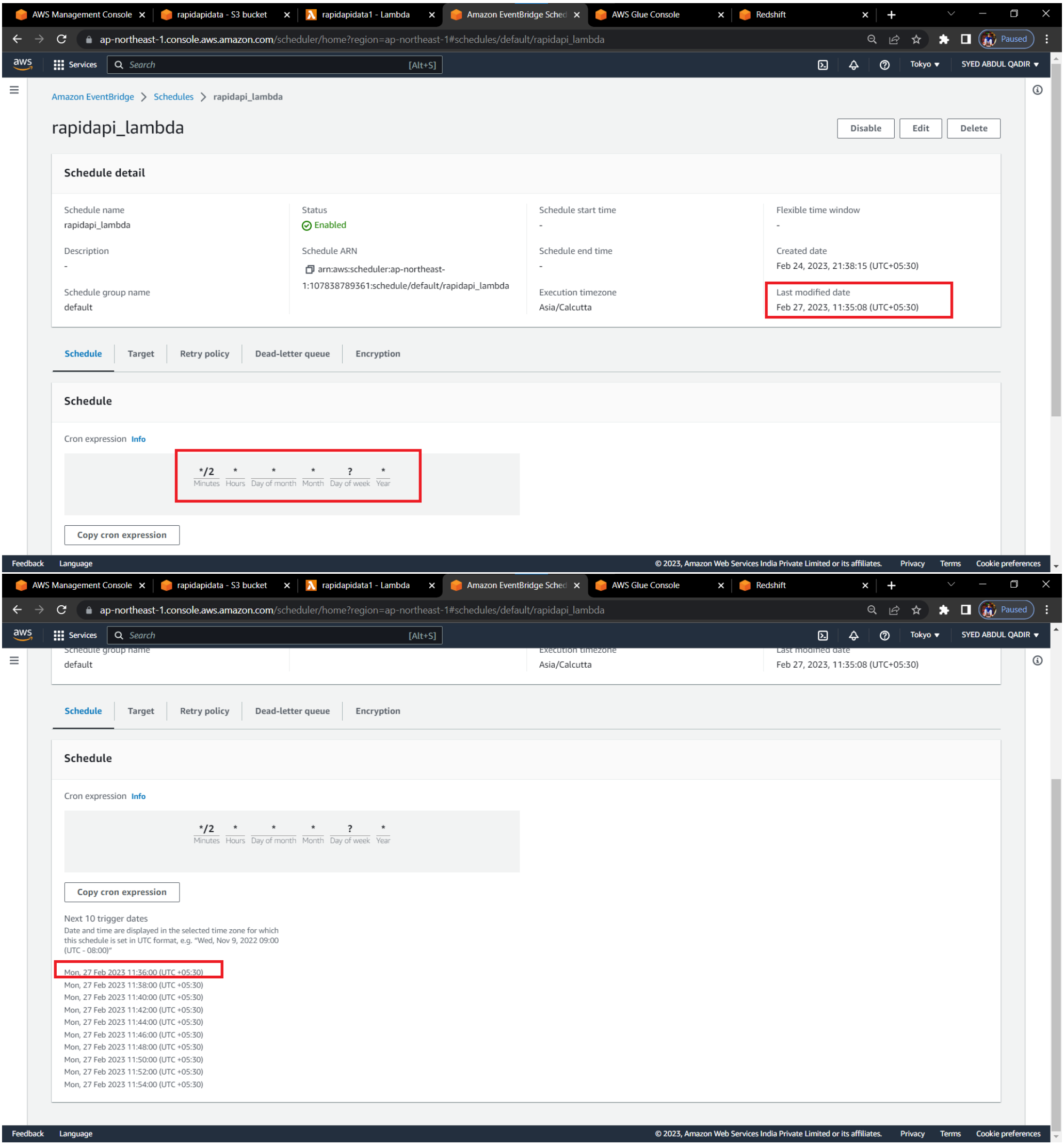




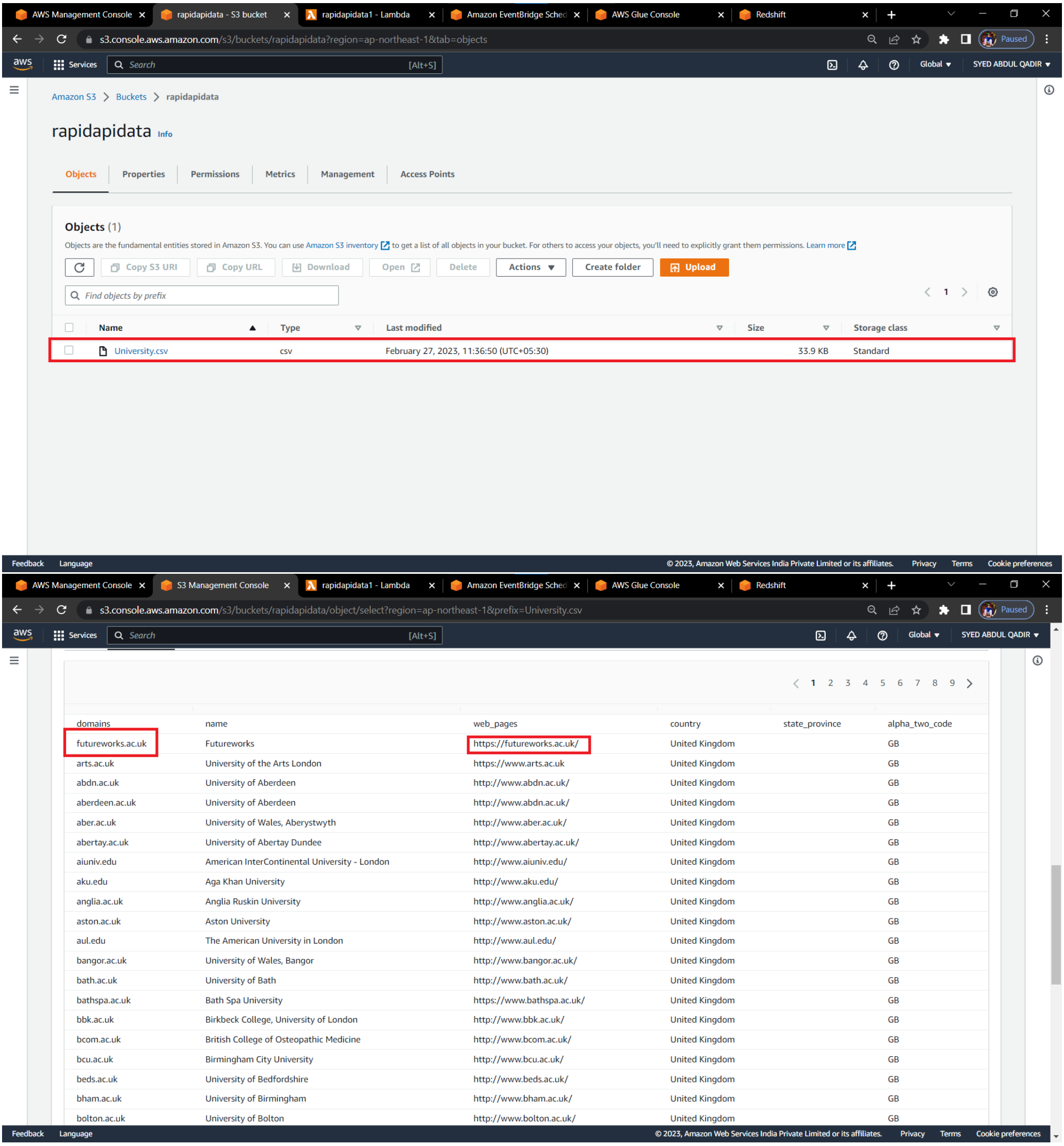
Adding **layers** to lambda function



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

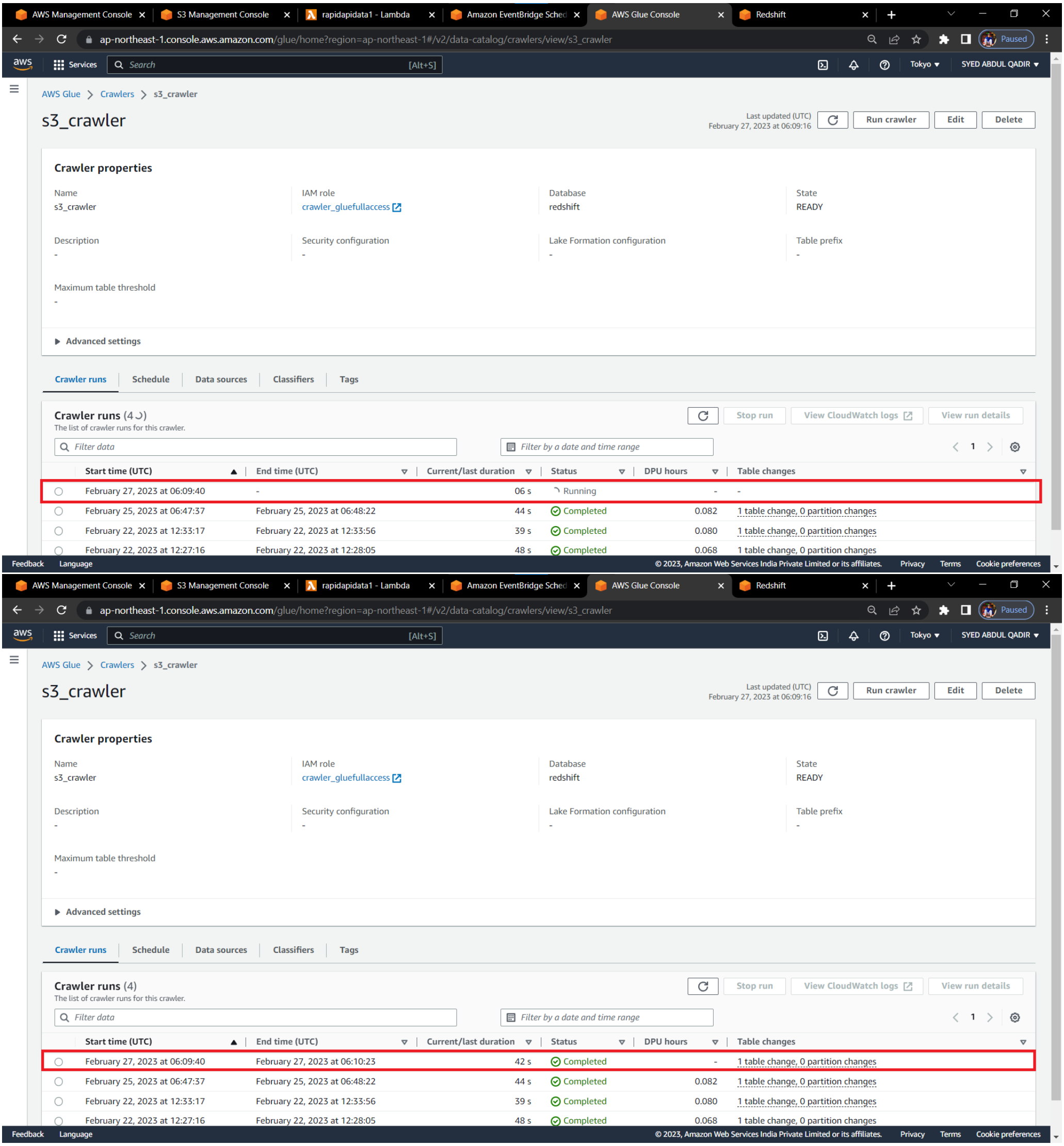


Fetch data from **Open source API** to AWS s3 bucket as **university.csv**

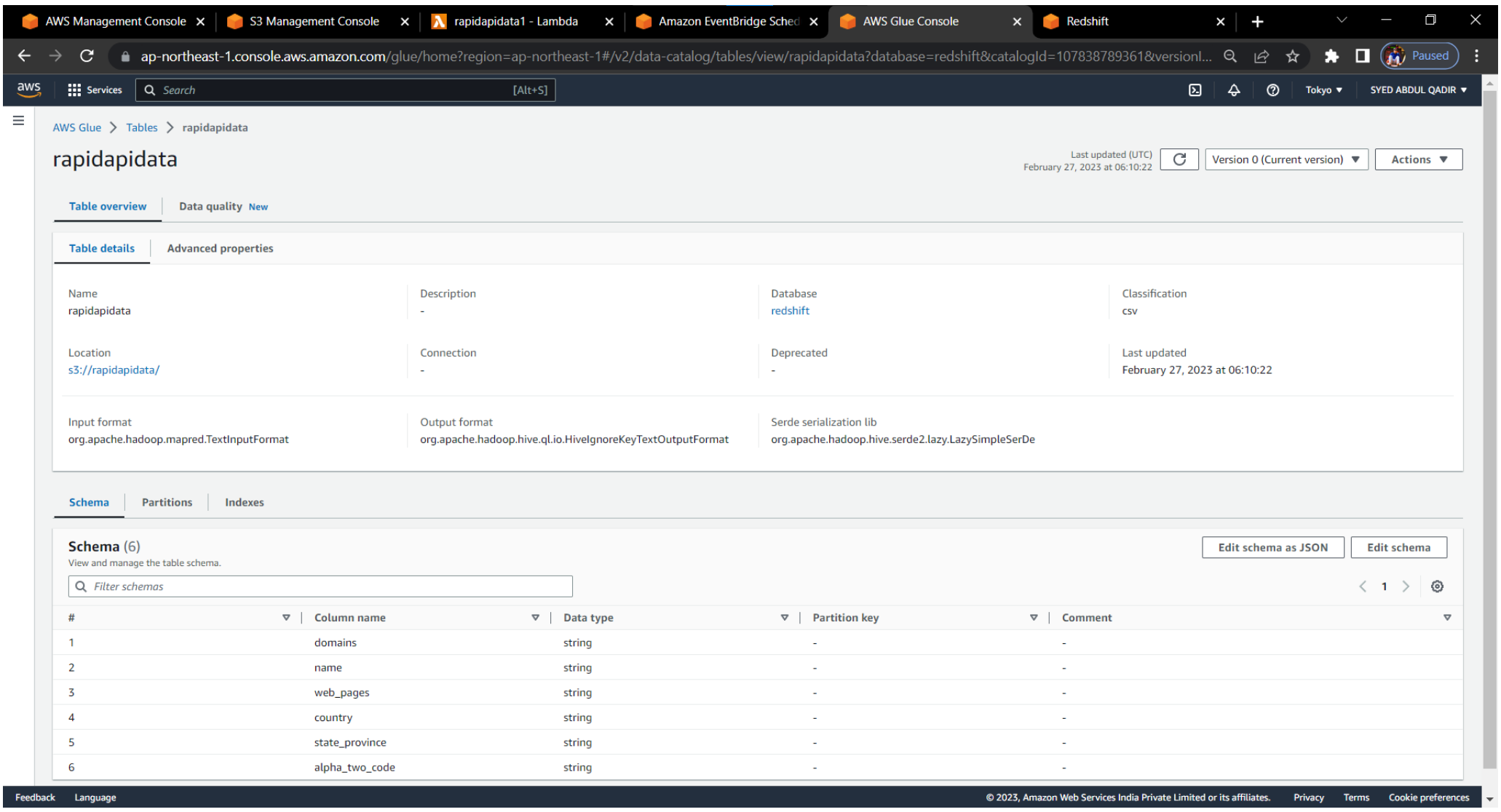


domains	name	web_pages	country	state_province	alpha_two_code
futureworks.ac.uk	Futureworks	https://futureworks.ac.uk/	United Kingdom		GB
arts.ac.uk	University of the Arts London	https://www.arts.ac.uk	United Kingdom		GB
abdn.ac.uk	University of Aberdeen	http://www.abdn.ac.uk/	United Kingdom		GB
aberdeen.ac.uk	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

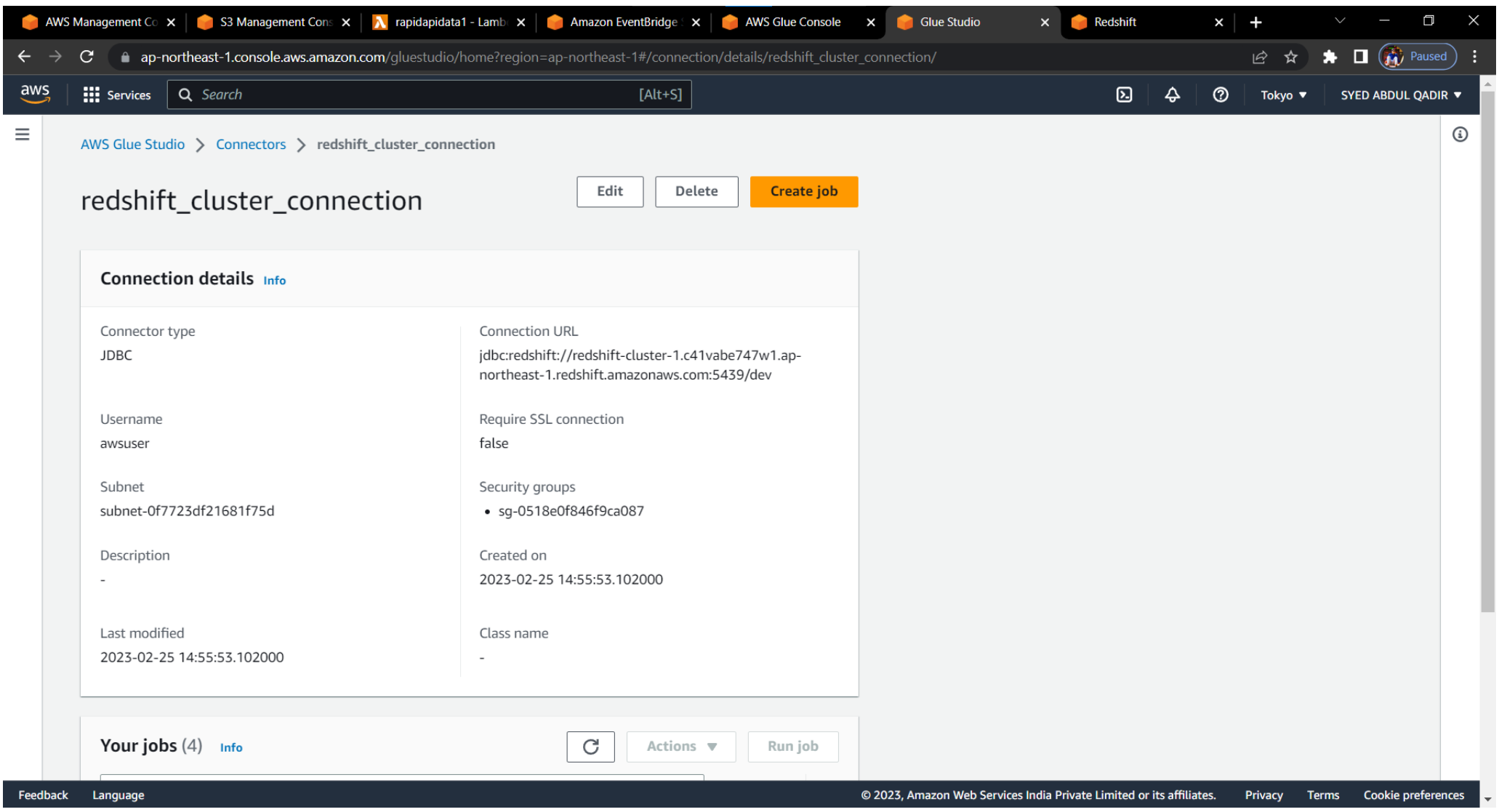
Created **crawler** for s3 bucket to get **schema metadata catalogue** for university.csv, started crawler to run.



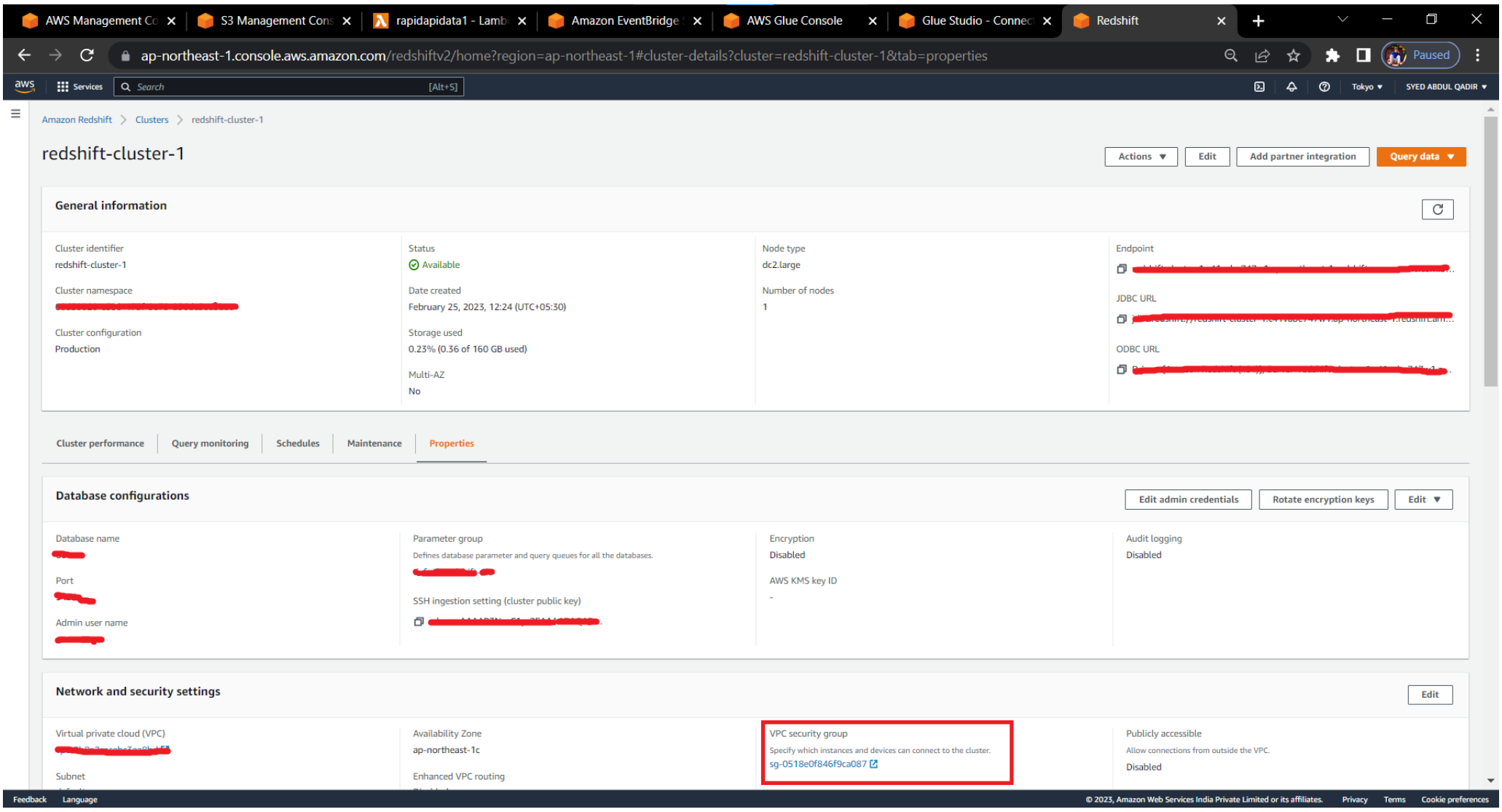
Able to verify **schema metadata** in database section Tables.



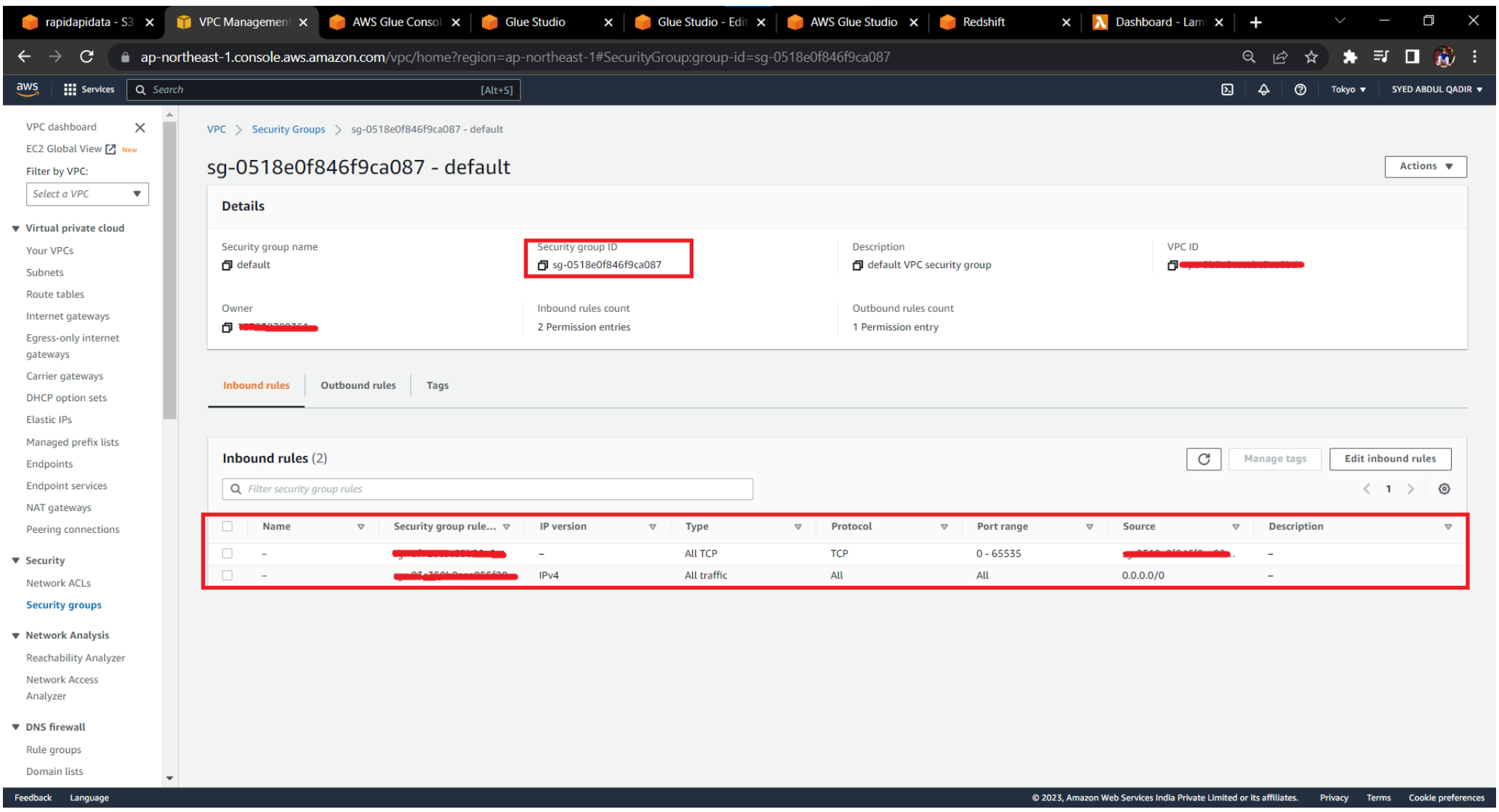
Created **connector** to make connection for **AWS redshift cluster.**



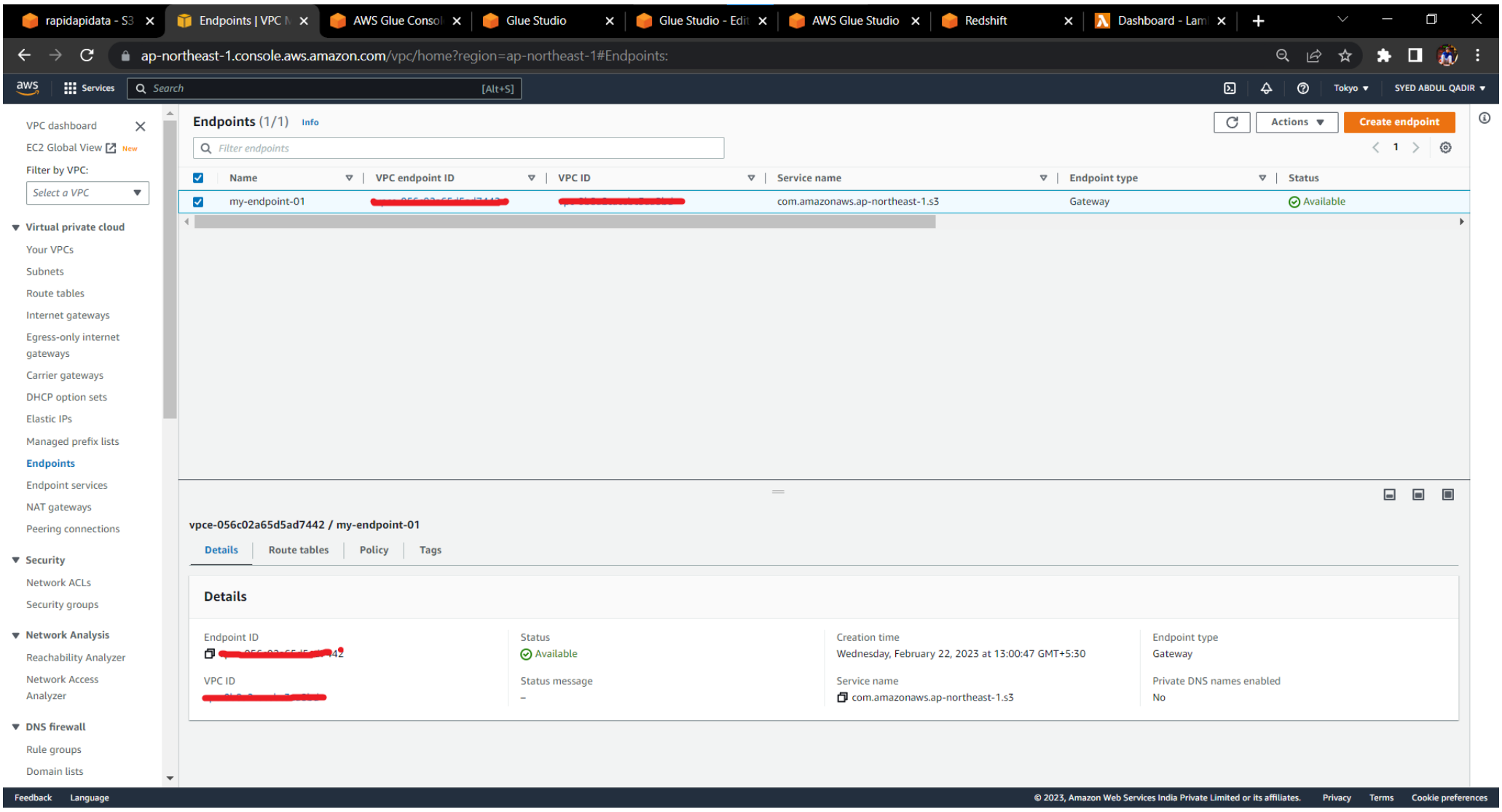
Created **AWS Redshift cluster** with free trial version configuration as **redshift-cluster-1.**



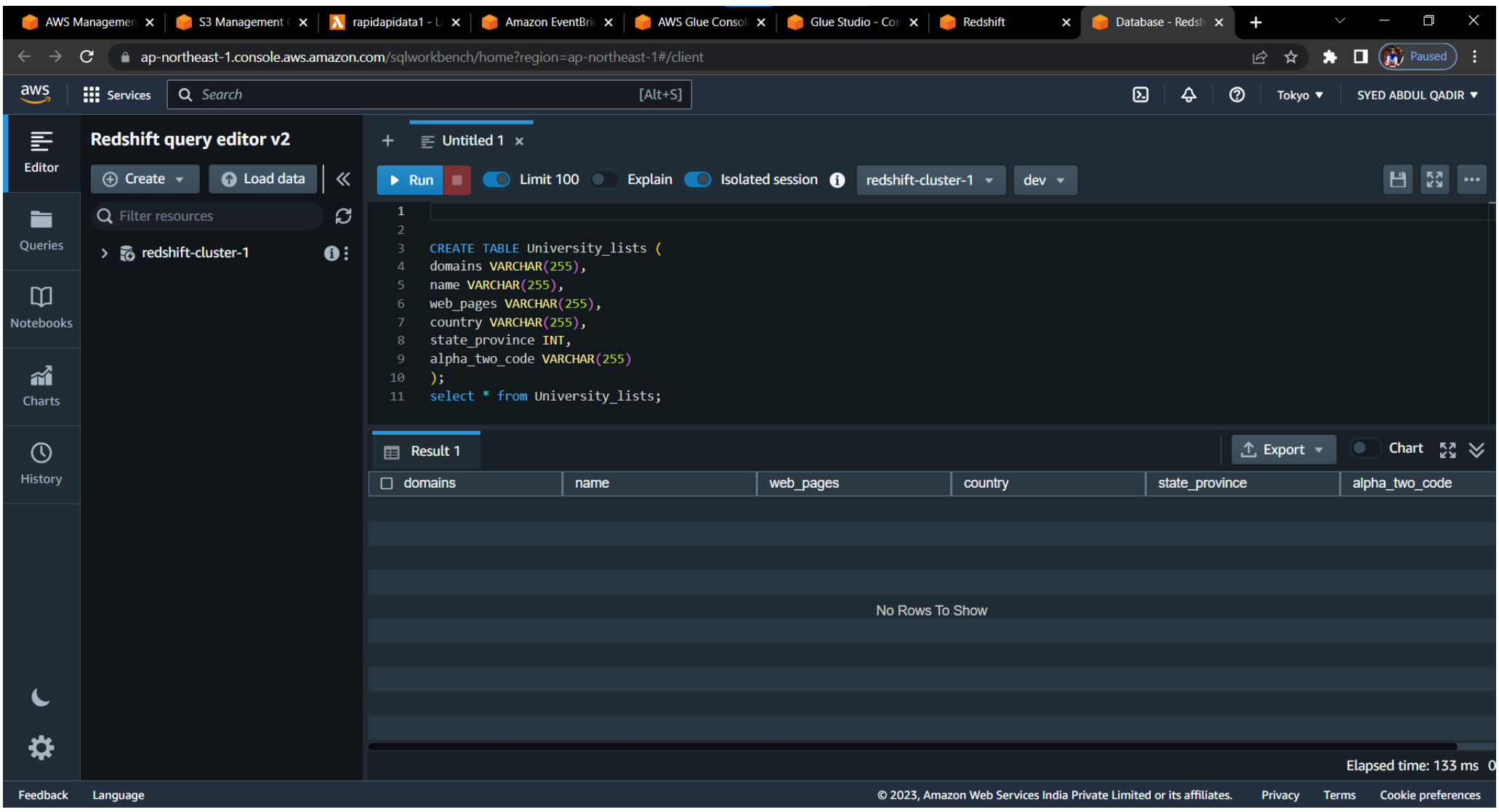
For **VPC security group** redshift cluster attaching Type **All TCP, All traffic.**



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

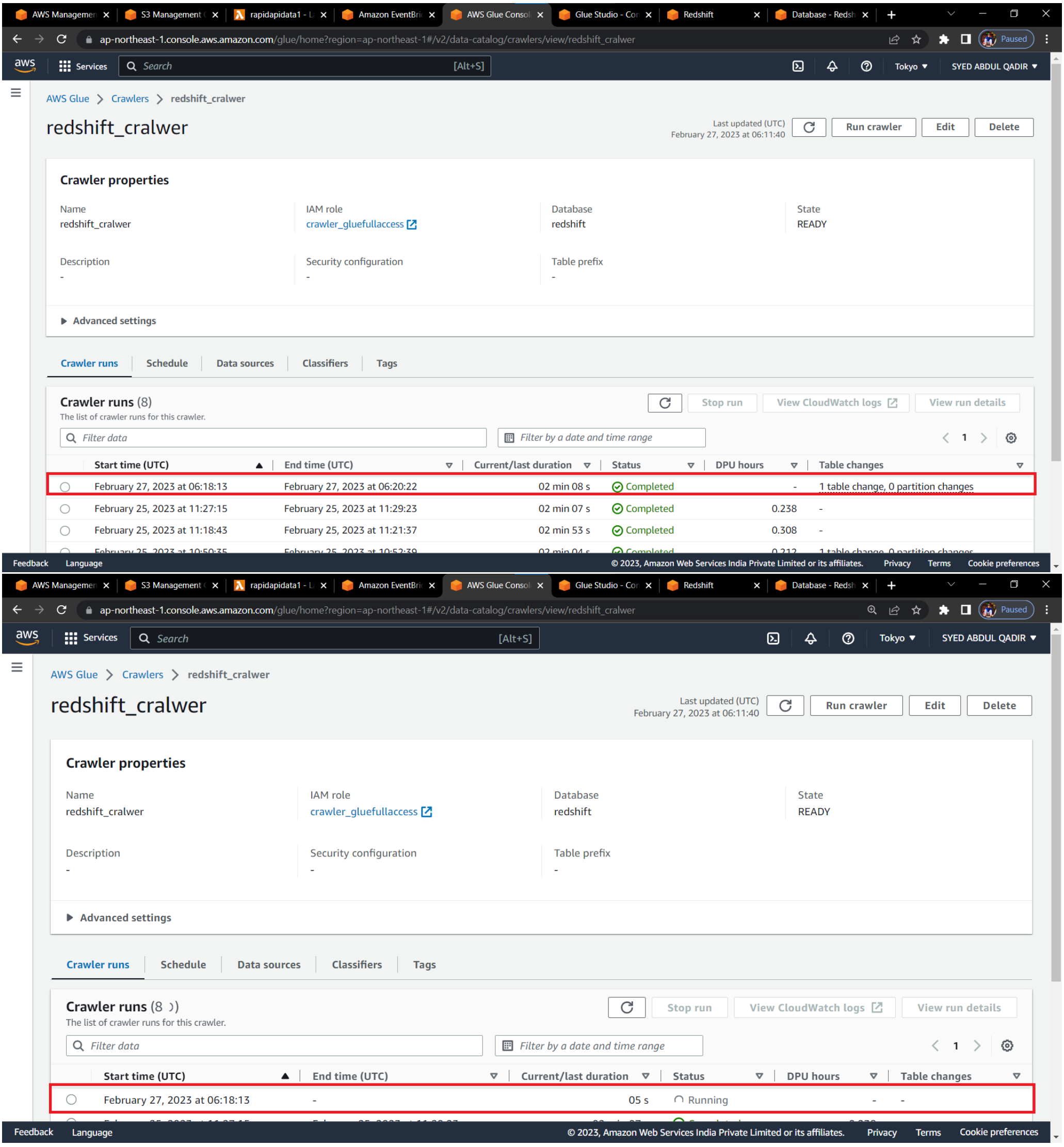


Created **Table** in redshift cluster as **University_lists.**

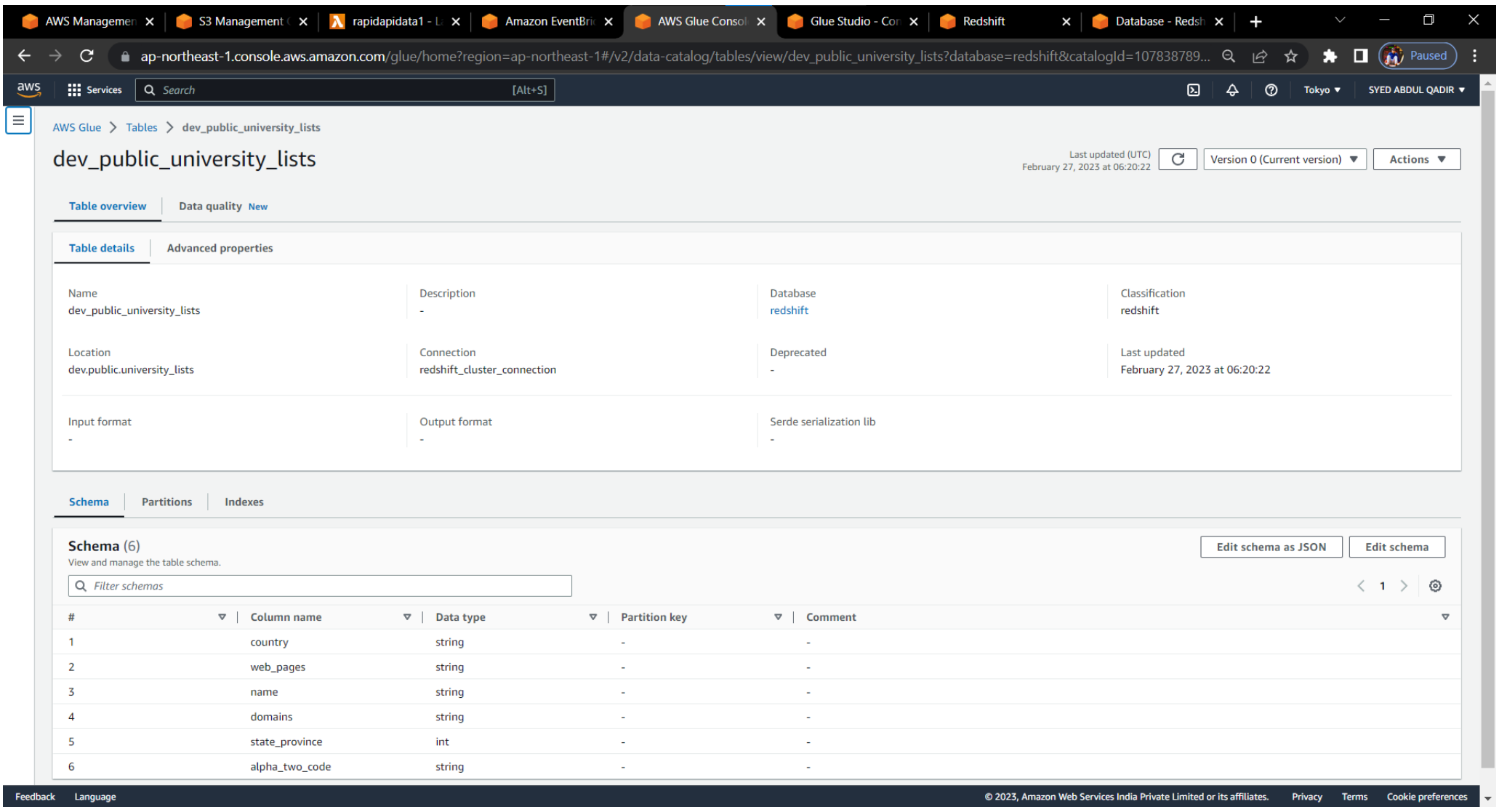


Created **crawler** pointing to redshift cluster for **University_lists** table to obtained schema catalogue in glue with help of connector of **redshift_cluster_connection,** with crawler name **redshift_crawler.**

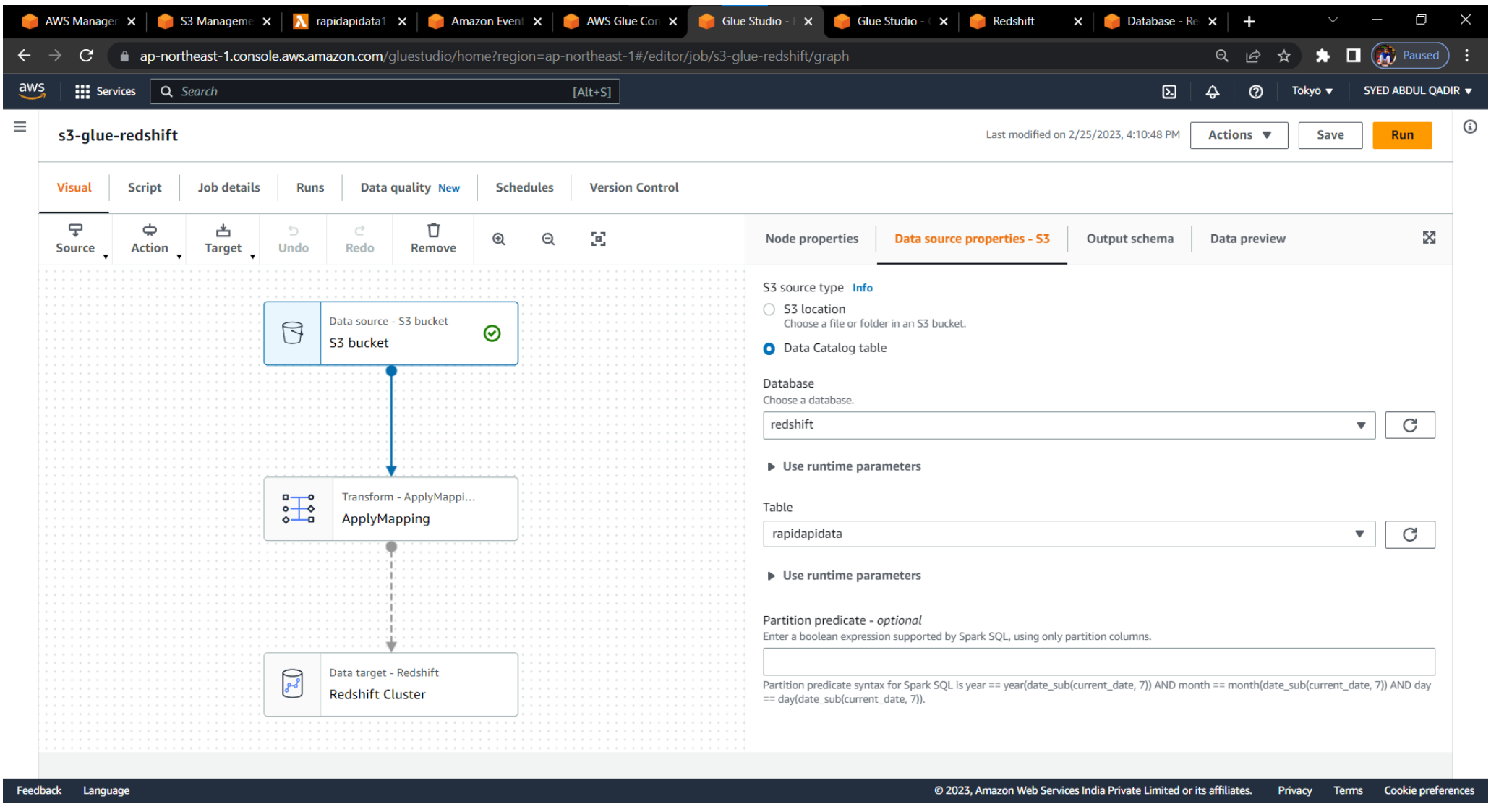
Crawler get successfully executed.

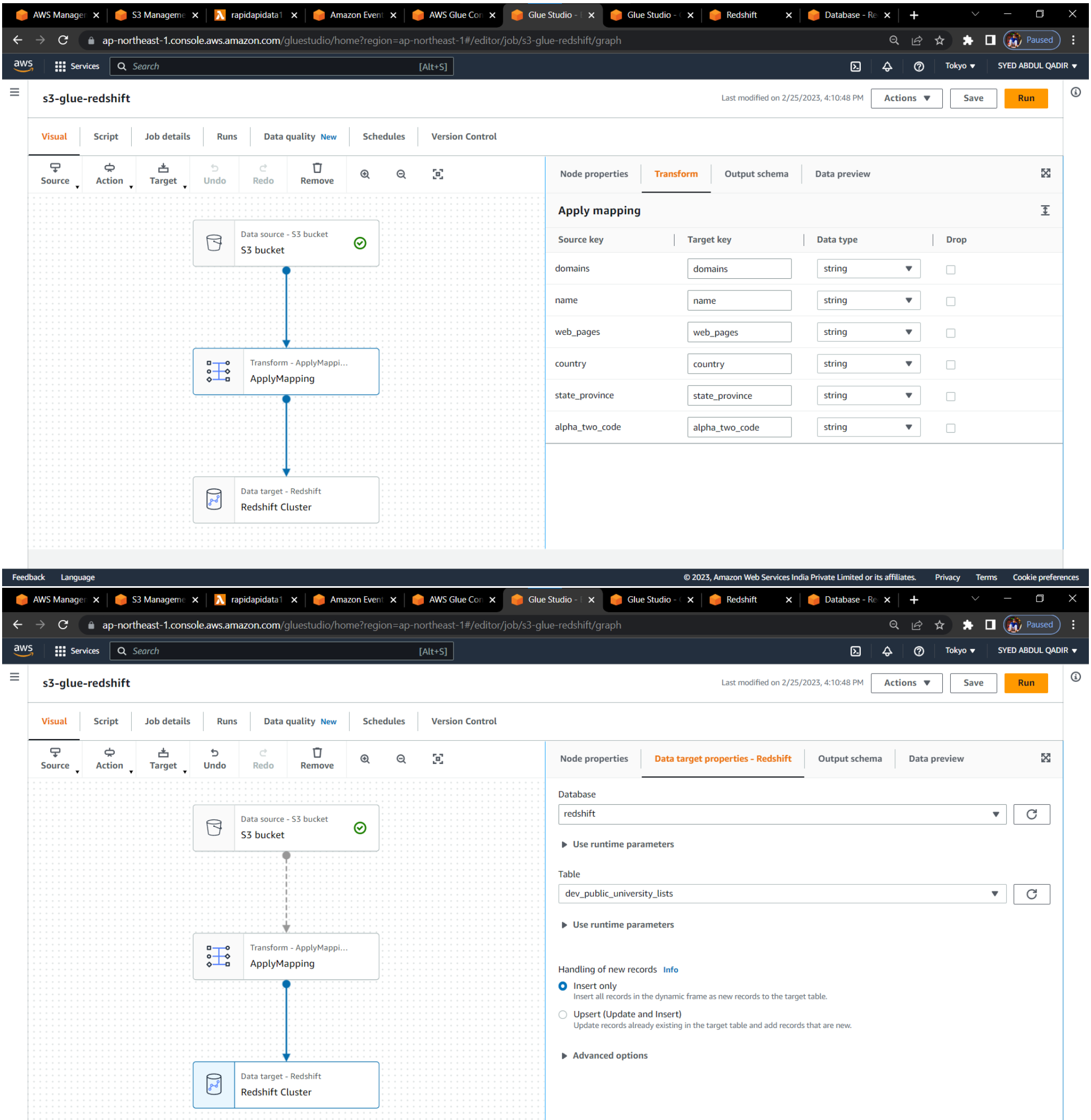


Able to fetch schema from redishift cluster.

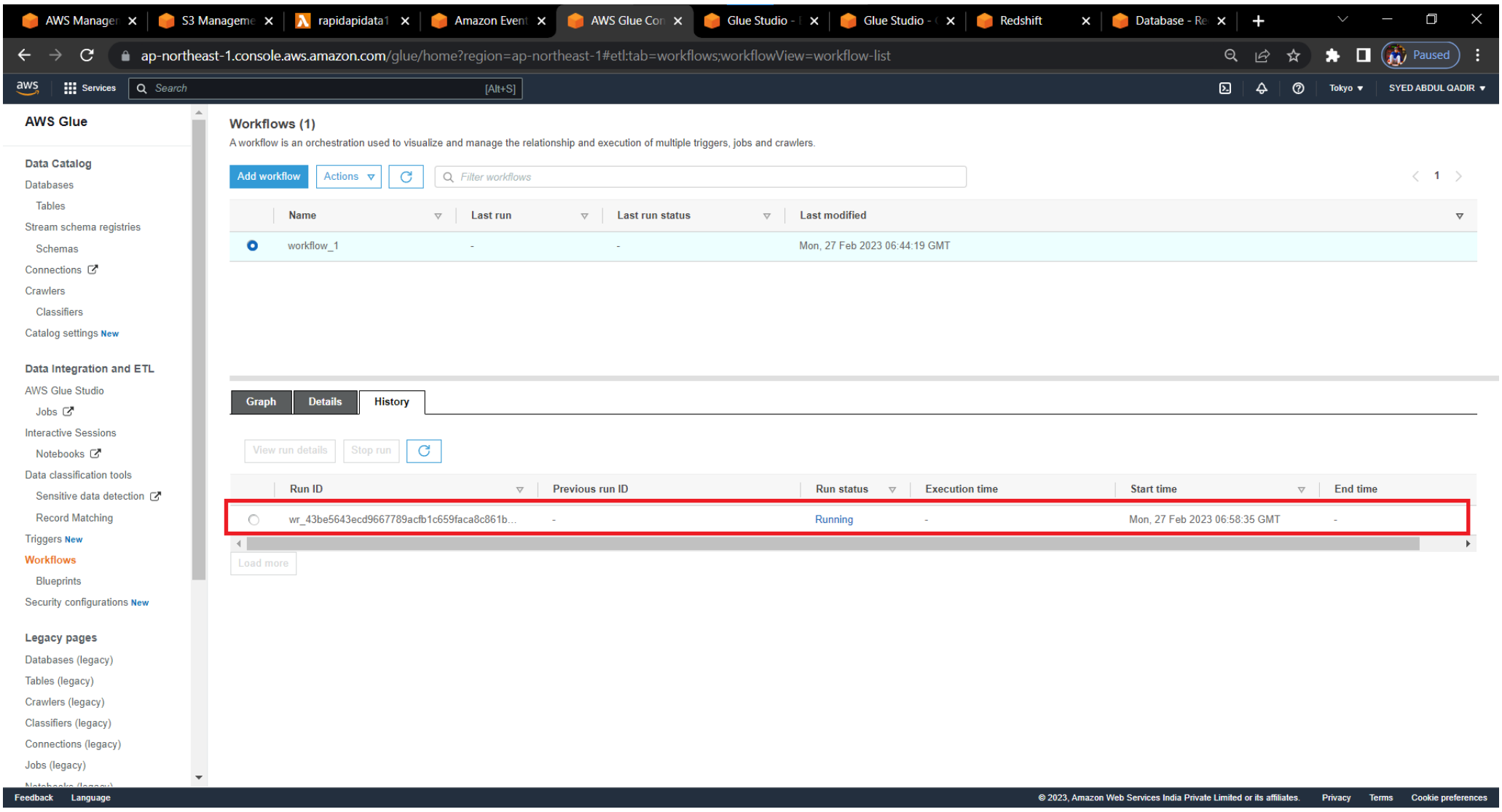


Now run **glue job** with help **visual source:s3** with data catalogue of s3 schema and **target** would be **Redshift cluster.**

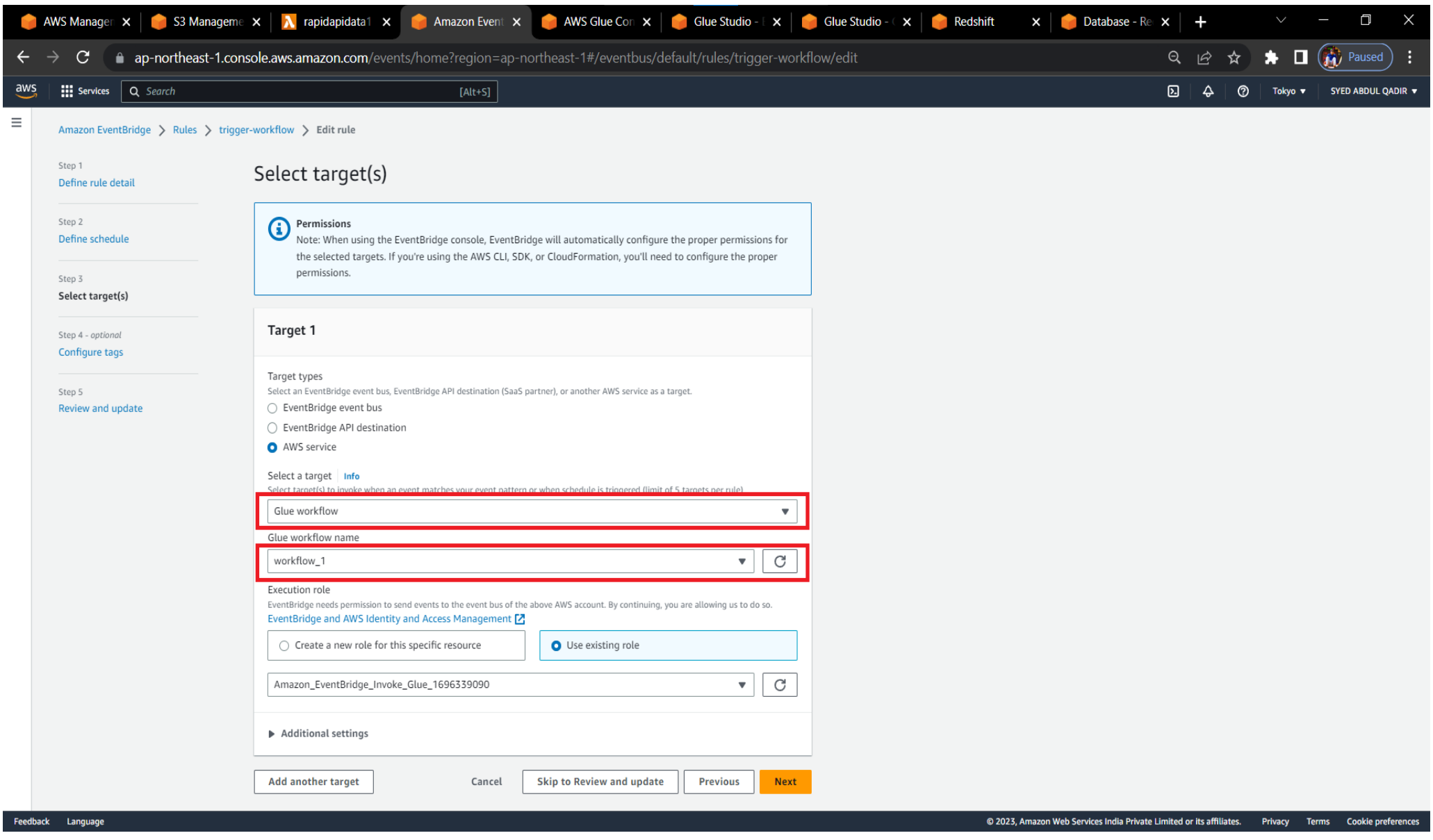


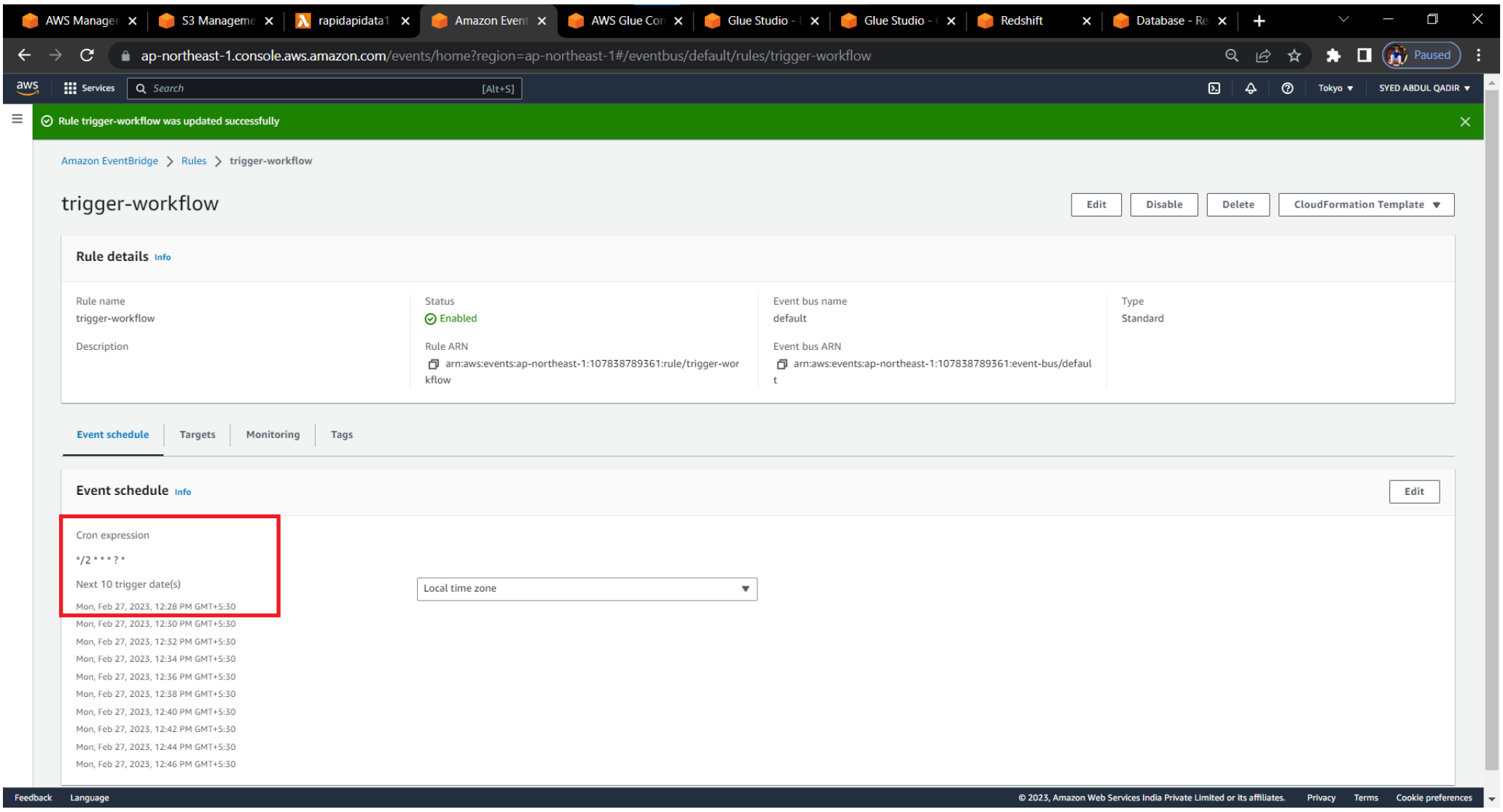

AWS Manager
S3 Manageme
rapidapidata1
Amazon Event
AWS Glue Cor
Glue Studio -
Glue Studio -
Redshift
Database - Re
ap-northeast-1.console.aws.amazon.com/gluestudio/home?region=ap-northeast-1#/editor/job/s3-glue-redshift/graph
Paused
aws
Services
Search
[Alt+S]
Tokyo
SYED ABDUL QADIR
s3-glue-redshift
Last modified on 2/25/2023, 4:10:48 PM
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality New
Schedules
Version Control
Source
Action
Target
Undo
Redo
Remove
Data source - S3 bucket
S3 bucket
Transform - ApplyMappi...
ApplyMapping
Data target - Redshift
Redshift Cluster
Node properties
Transform
Output schema
Data preview
Apply mapping
Source key
Target key
Data type
Drop
domains
domains
string
name
name
string
web_pages
web_pages
string
country
country
string
state_province
state_province
string
alpha_two_code
alpha_two_code
string
Feedback
Language
© 2023, Amazon Web Services India Private Limited or its affiliates.
Privacy
Terms
Cookie preferences
Data target properties - Redshift
Database
redshift
Use runtime parameters
Table
dev_public_university_lists
Use runtime parameters
Handling of new records Info
Insert only
Insert all records in the dynamic frame as new records to the target table.
Upsert (Update and Insert)
Update records already existing in the target table and add records that are new.
Advanced options

Created **workflow** in glue job to trigger job.

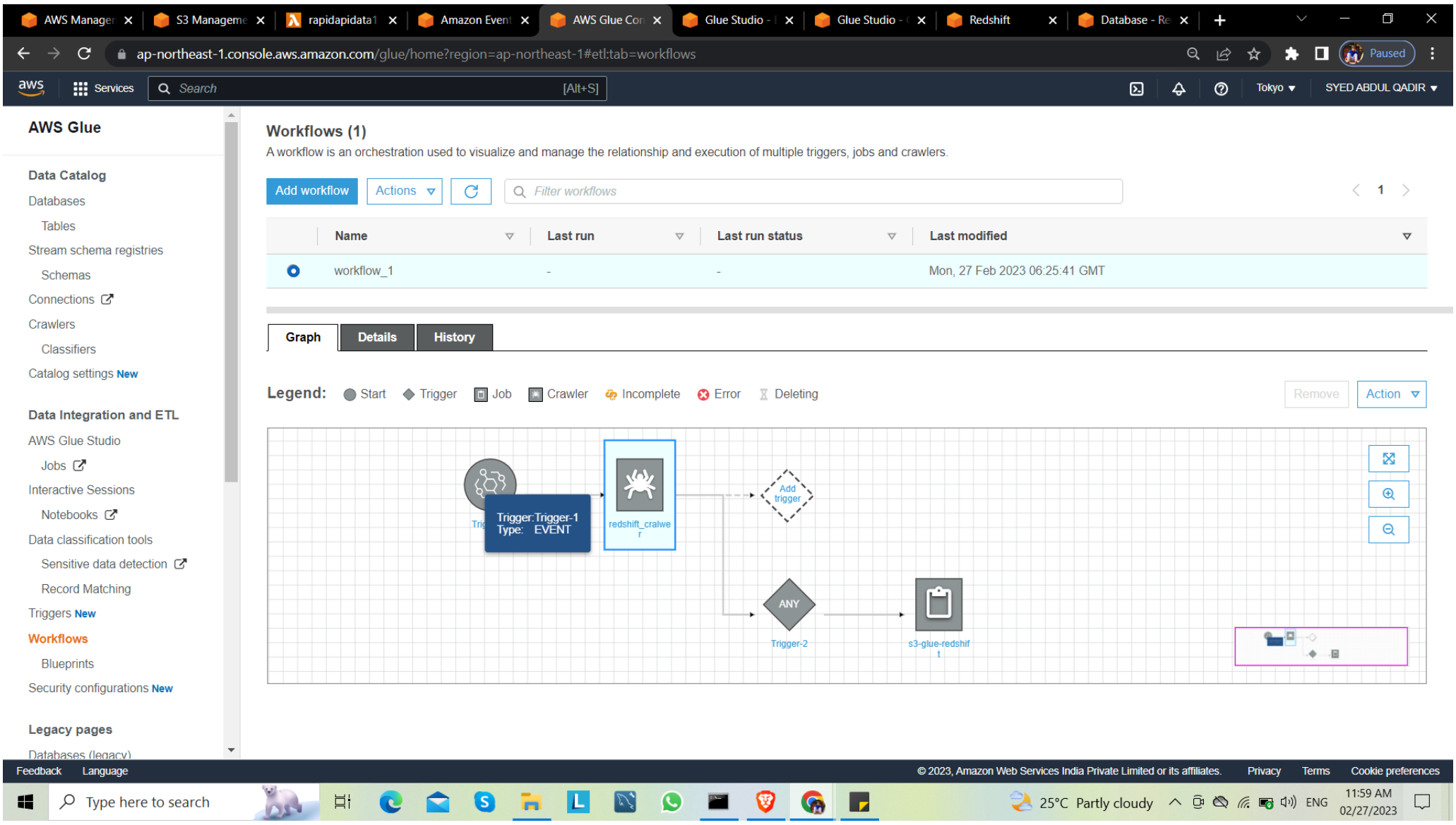


Created **AWS Event Bridge** of **rule** to trigger the **workflow** created in glue to run it by every two minutes.

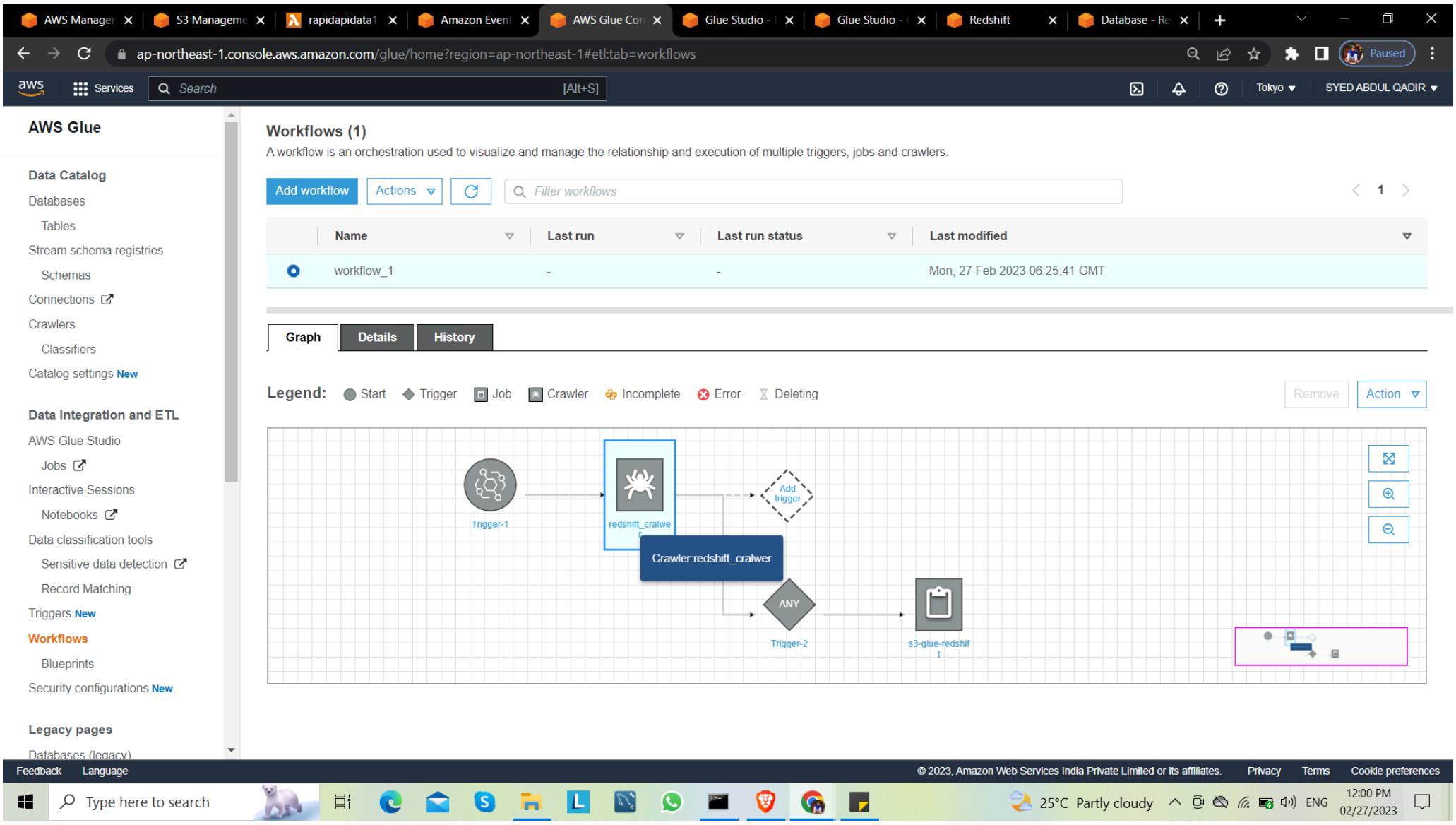




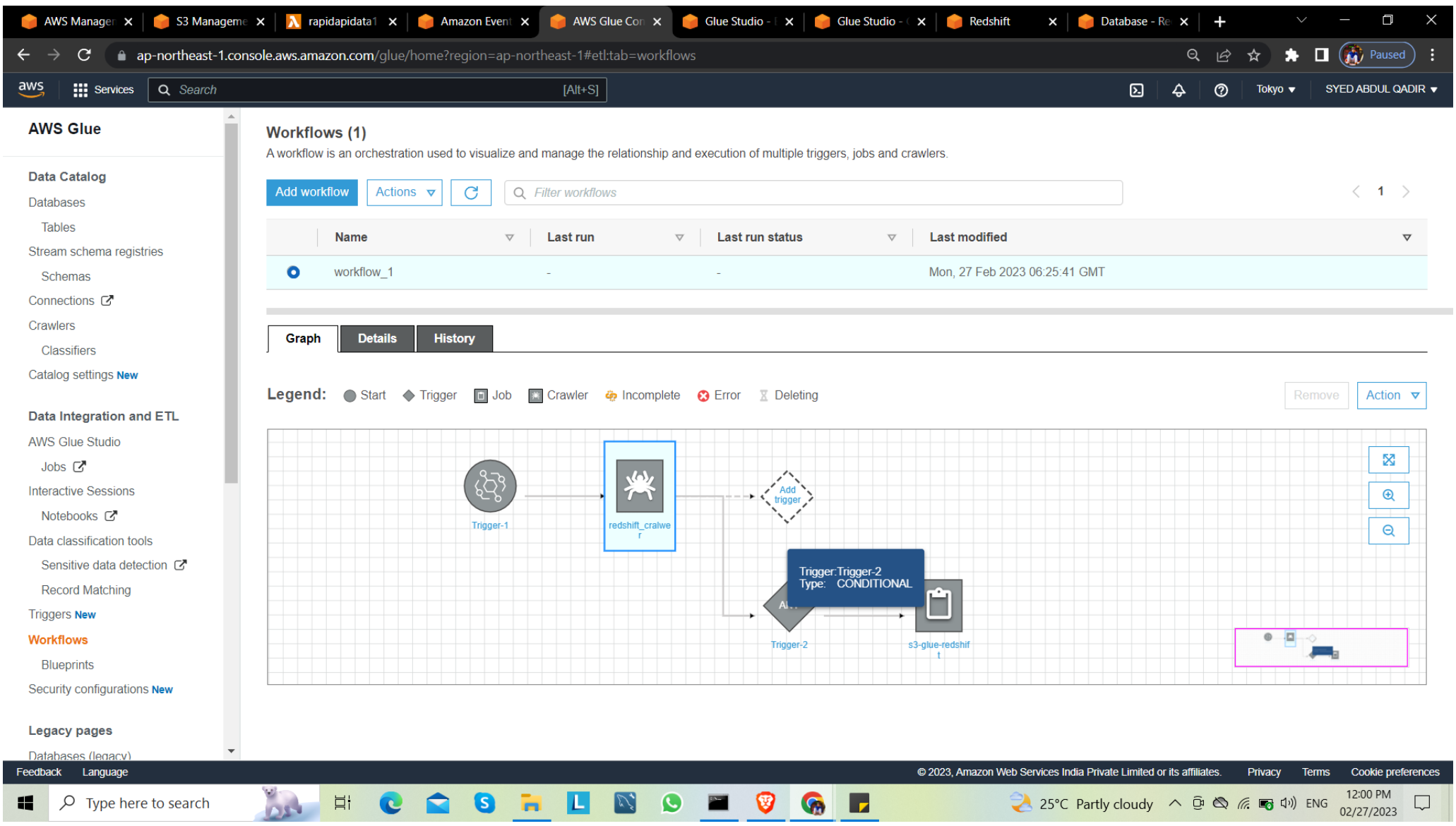
Trigger-1 with type: **event bridge.**



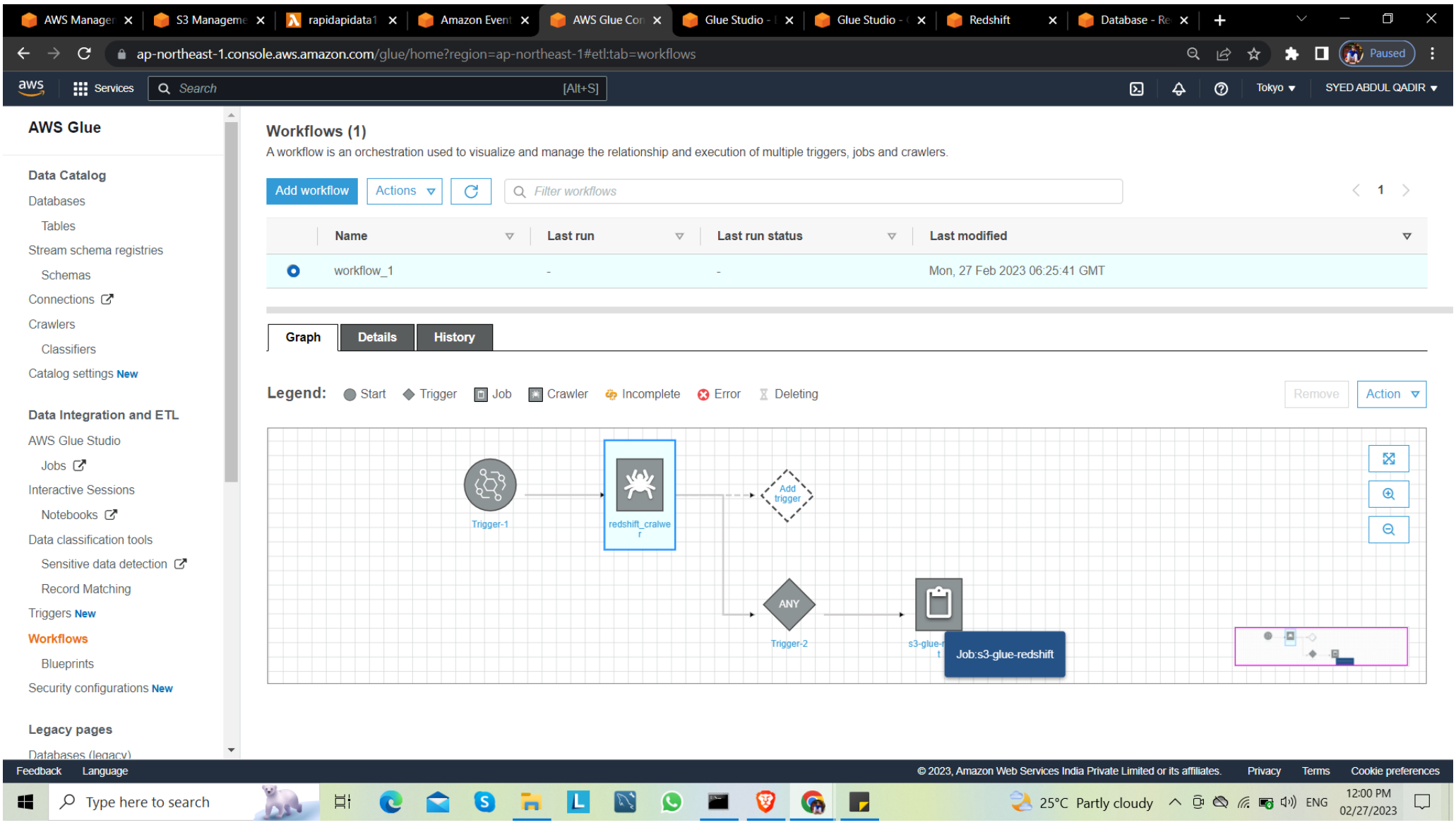
For **Trigger-1** attaching **crawler** of **redshift cluster.**



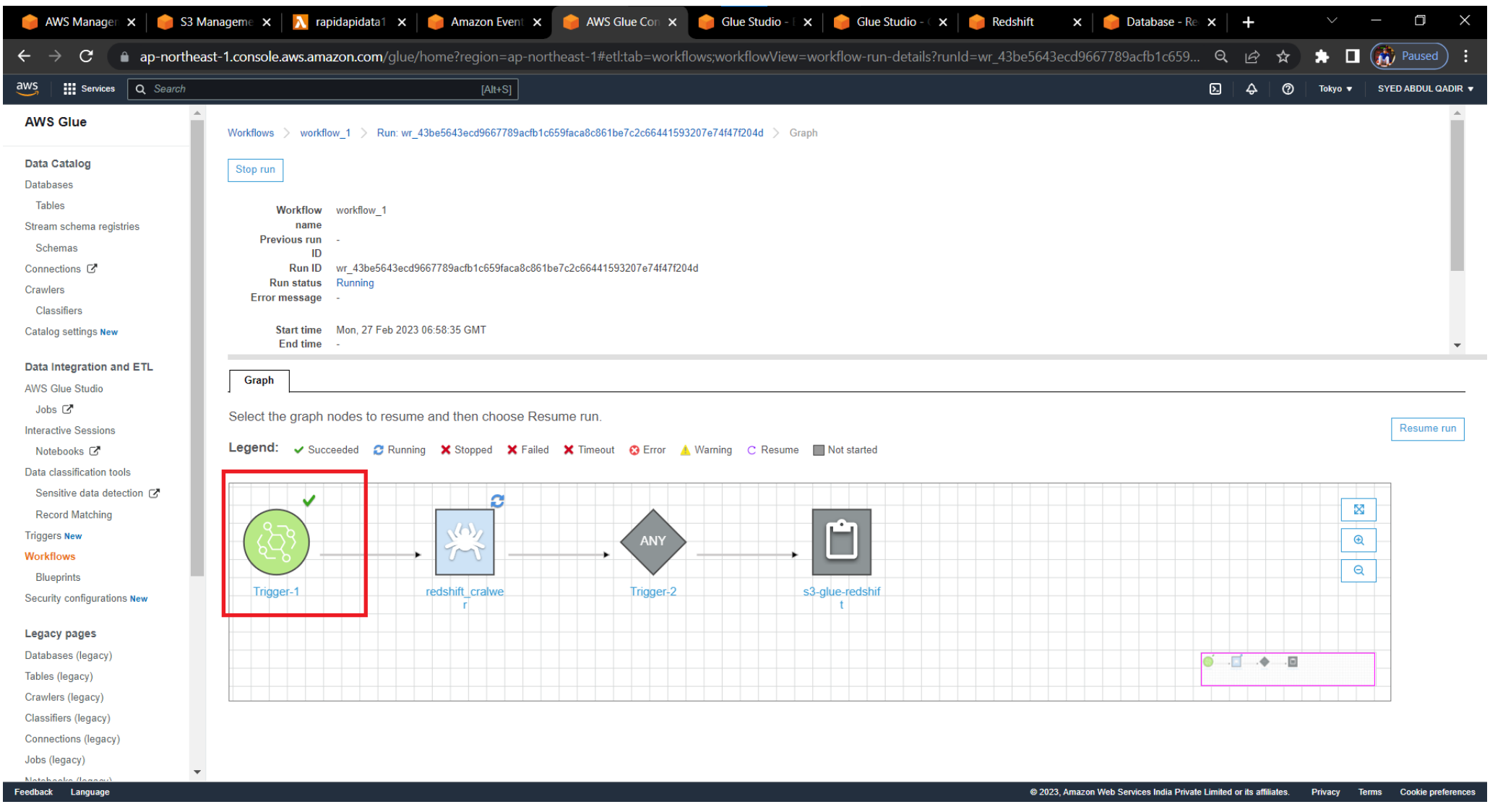
Trigger-2 with type: by default **event.**



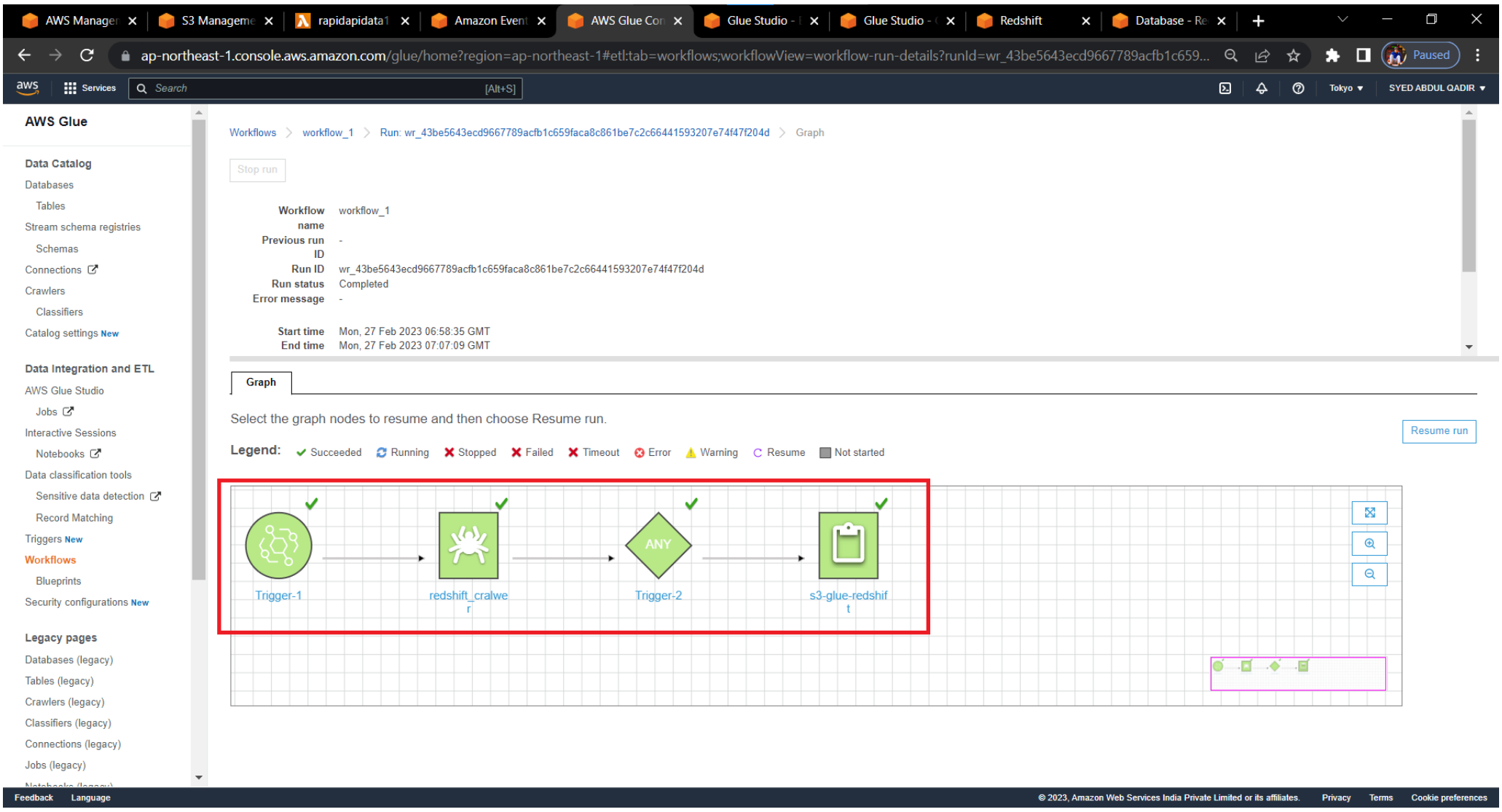
Attaching **glue job** with name **s3-glue-redshift** to **Trigger-2.**



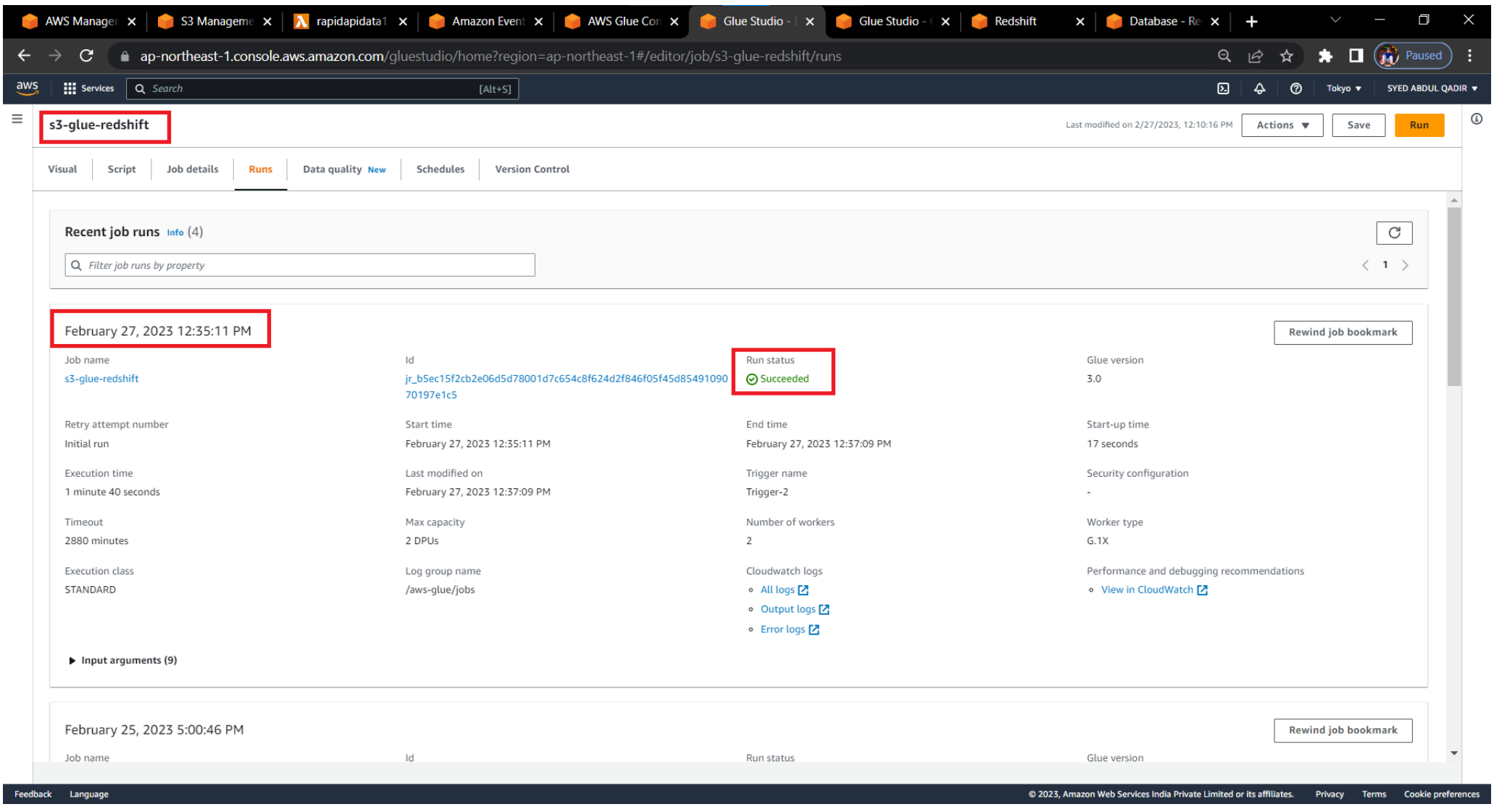
After **Trigger-1** succeeded disenable, the rule created at event bridge.



Overall **workflow** will get executed **succeeded**.



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



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