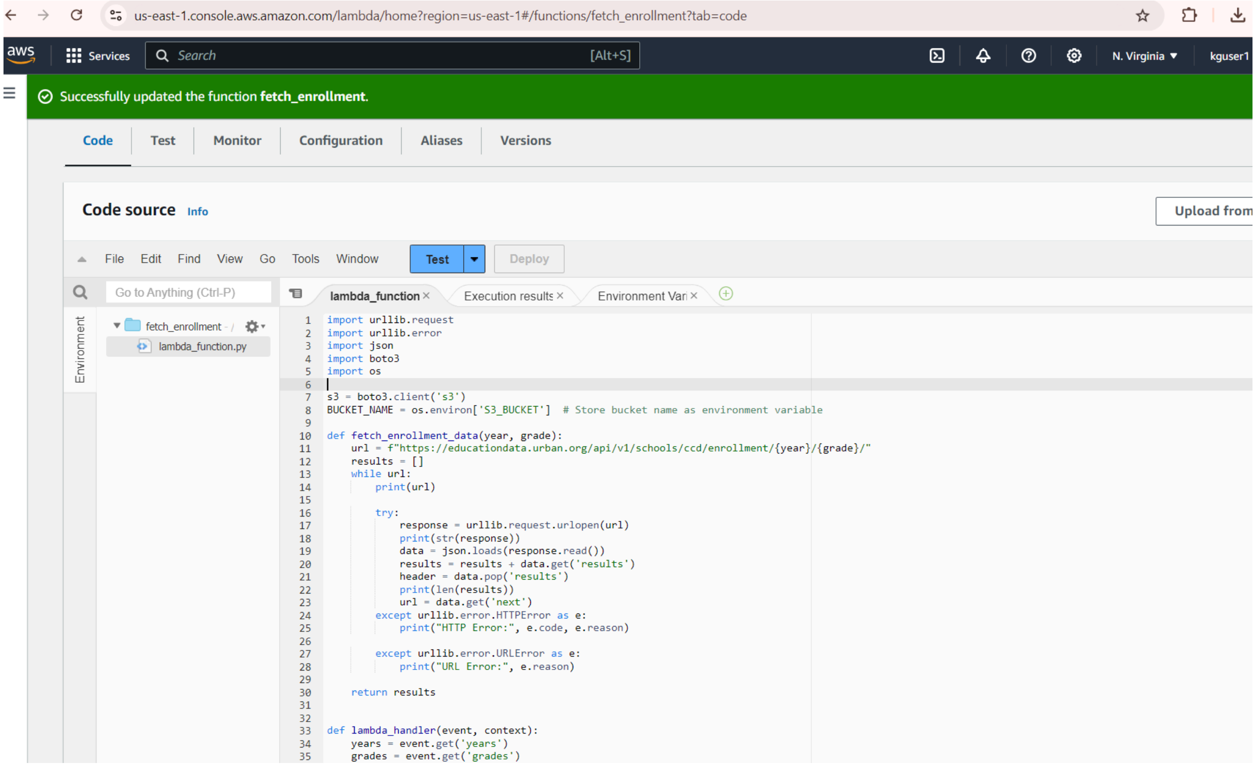
Approach

Fetch The Data

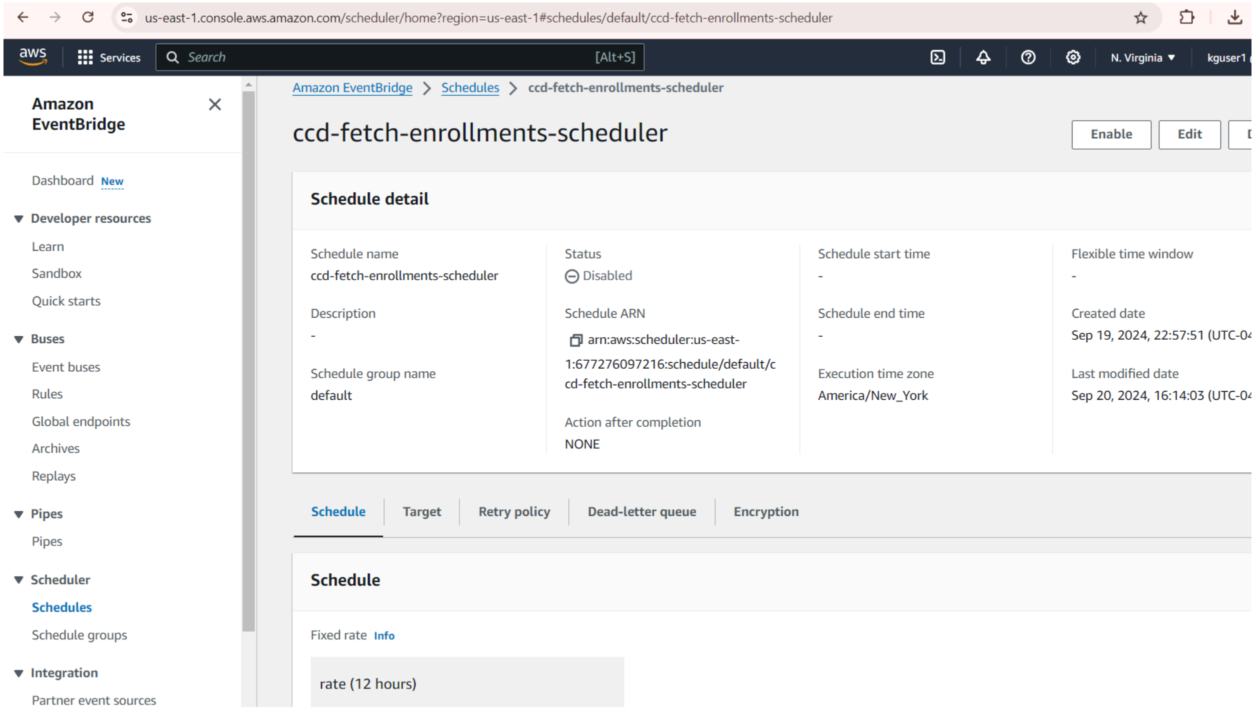
1. AWS Lambda suits this use case better as the source for the data is a REST endpoint.
2. This API doesn't need any credentials, if it needs any credentials AWS Secret Manager would be useful to store and retrieve the credentials.
3. There is no incremental parameter on the REST API endpoint like date or incremental id, so the incremental run would pull all the available data for the year and grade on a regular schedule. If there is an incremental way of pulling the data, we could use the db table like dynamo db to store the field value to be used for incremental data extract.



Successfully updated the function **fetch_enrollment**.

```python
import urllib.request
import urllib.error
import json
import boto3
import os

s3 = boto3.client('s3')
BUCKET_NAME = os.environ['S3_BUCKET']  # Store bucket name as environment variable

def fetch_enrollment_data(year, grade):
    url = f"https://educationdata.urban.org/api/v1/schools/ccd/enrollment/{year}/{grade}/"
    results = []
    while url:
        print(url)

        try:
            response = urllib.request.urlopen(url)
            print(str(response))
            data = json.loads(response.read())
            results = results + data.get('results')
            header = data.pop('results')
            print(len(results))
            url = data.get('next')
        except urllib.error.HTTPError as e:
            print("HTTP Error:", e.code, e.reason)

        except urllib.error.URLError as e:
            print("URL Error:", e.reason)

    return results


def lambda_handler(event, context):
    years = event.get('years')
    grades = event.get('grades')
```

4. Scheduled the Lambda using AWS EventBridge scheduler to run every 12 hours. The scheduler passes the event with year and grade parameters to the Lambda function, this gives flexibility to expand the dataset with changing the code.
    a. `{ "years": [2020, 2021], "grades": ["grade-pk"] }`



ccd-fetch-enrollments-scheduler

Schedule detail

Schedule name
ccd-fetch-enrollments-scheduler

Description
-

Schedule group name
default

Status
⊖ Disabled

Schedule ARN
arn:aws:scheduler:us-east-1:677276097216:schedule/default/ccd-fetch-enrollments-scheduler

Action after completion
NONE

Schedule start time
-

Schedule end time
-

Execution time zone
America/New_York

Flexible time window
-

Created date
Sep 19, 2024, 22:57:51 (UTC-04

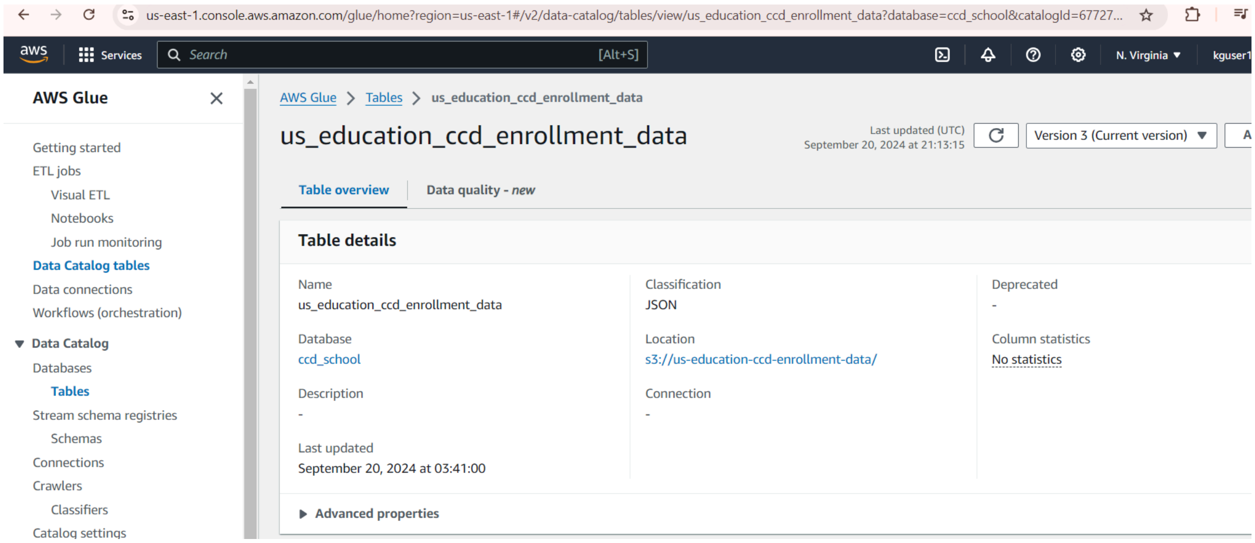
Last modified date
Sep 20, 2024, 16:14:03 (UTC-04

Schedule

Fixed rate Info

rate (12 hours)

5. AWS Glue is used to create the catalog table for the S3 data



us_education_ccd_enrollment_data

Last updated (UTC)
September 20, 2024 at 21:13:15

Table details

Name
us_education_ccd_enrollment_data

Database
ccd_school

Description
-

Last updated
September 20, 2024 at 03:41:00

Classification
JSON

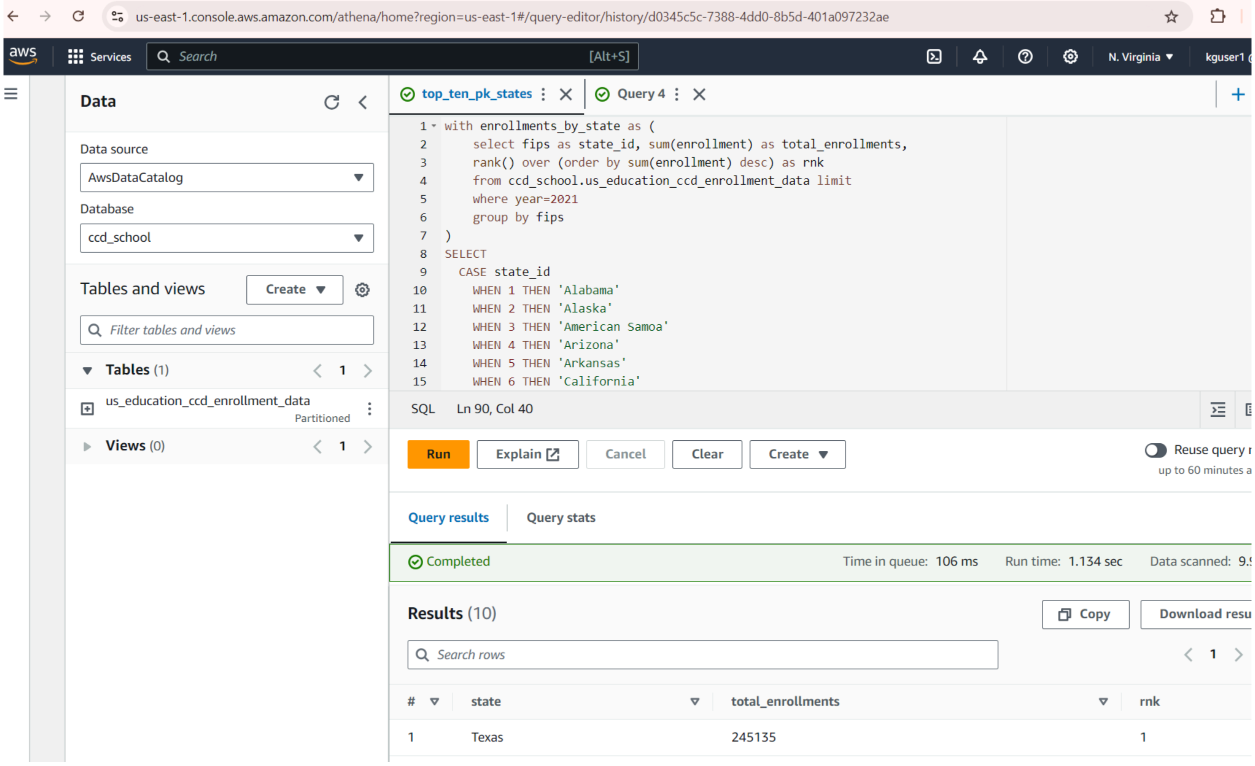
Location
s3://us-education-ccd-enrollment-data/

Connection
-

Deprecated
-

Column statistics
No statistics

6. AWS Athena is used to query the data



Data source: AwsDataCatalog

Database: ccd_school

Tables and views

Tables (1)

us_education_ccd_enrollment_data — Partitioned

Views (0)

```sql
with enrollments_by_state as (
    select fips as state_id, sum(enrollment) as total_enrollments,
    rank() over (order by sum(enrollment) desc) as rnk
    from ccd_school.us_education_ccd_enrollment_data limit
    where year=2021
    group by fips
)
SELECT
  CASE state_id
    WHEN 1 THEN 'Alabama'
    WHEN 2 THEN 'Alaska'
    WHEN 3 THEN 'American Samoa'
    WHEN 4 THEN 'Arizona'
    WHEN 5 THEN 'Arkansas'
    WHEN 6 THEN 'California'
```

SQL Ln 90, Col 40

Completed — Time in queue: 106 ms — Run time: 1.134 sec — Data scanned: 9.

Results (10)

#	state	total_enrollments	rnk
1	Texas	245135	1

Possible Improvements that would need more time:

1. Cloud Formation templates could be used for creating infrastructure components
2. Could build CI/CD pipeline using tools like Github Actions or Jenkins to automate the build and deployment