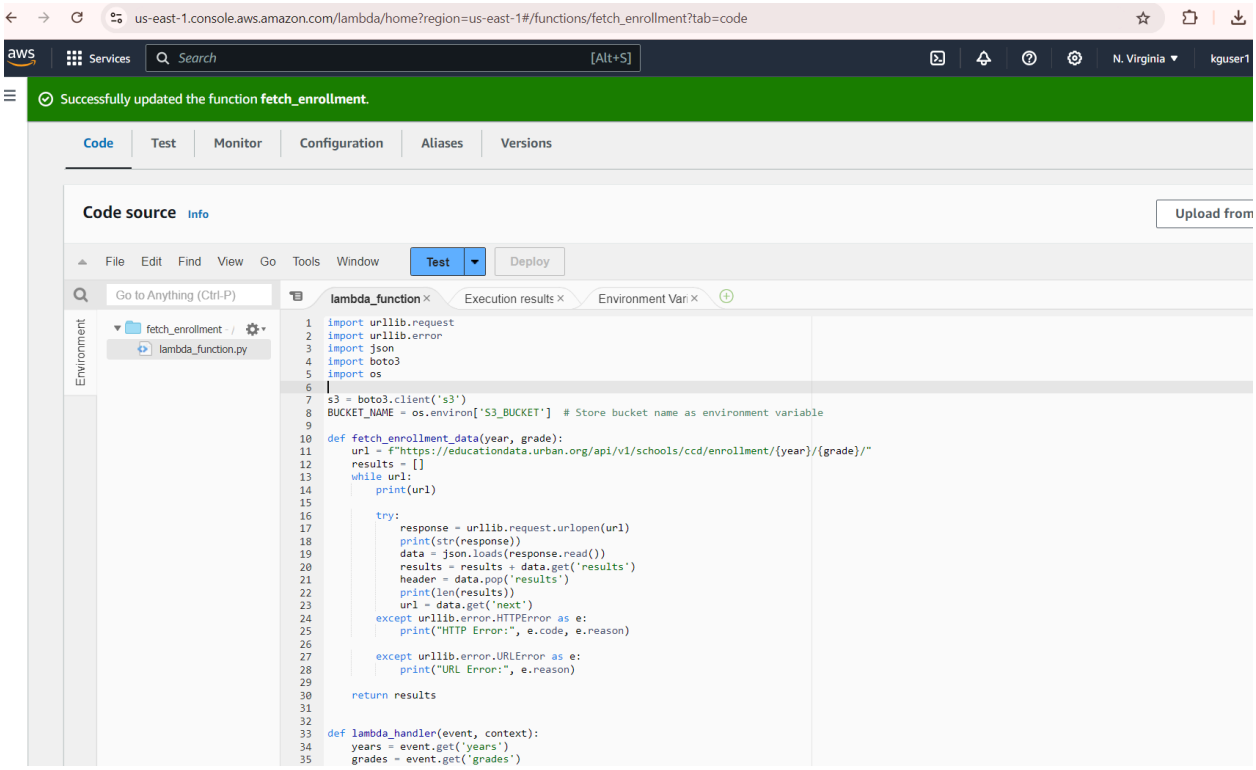
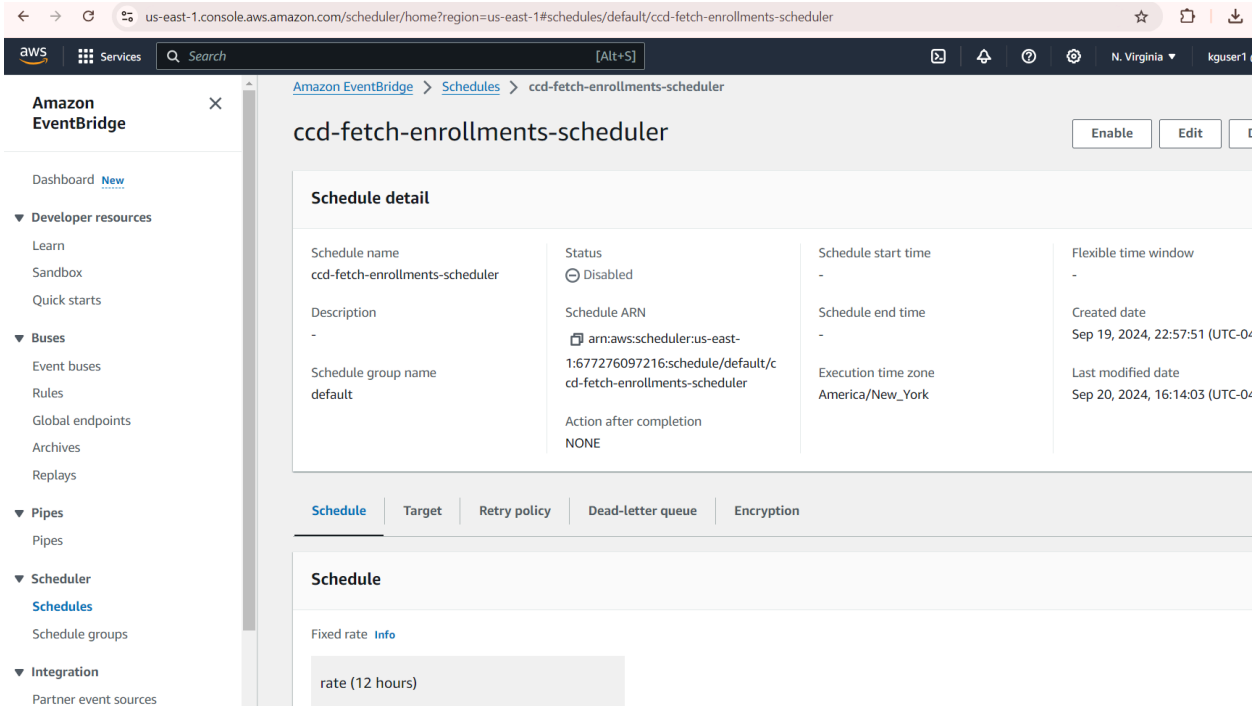
Approach

Fetch The Data

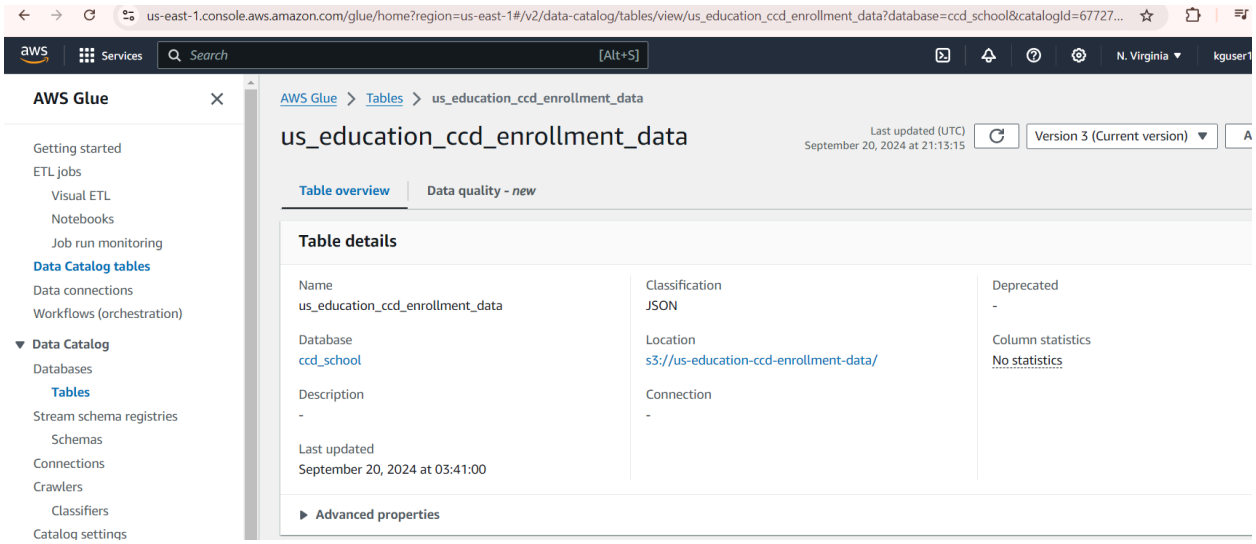
1. AWS Lambda suits this use case better as the source for the data is a REST endpoint.
2. This API doesn't need any credentials, if it needs any credentials AWS Secret Manager would be useful to store and retrieve the credentials.
3. There is no incremental parameter on the REST API endpoint like date 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.



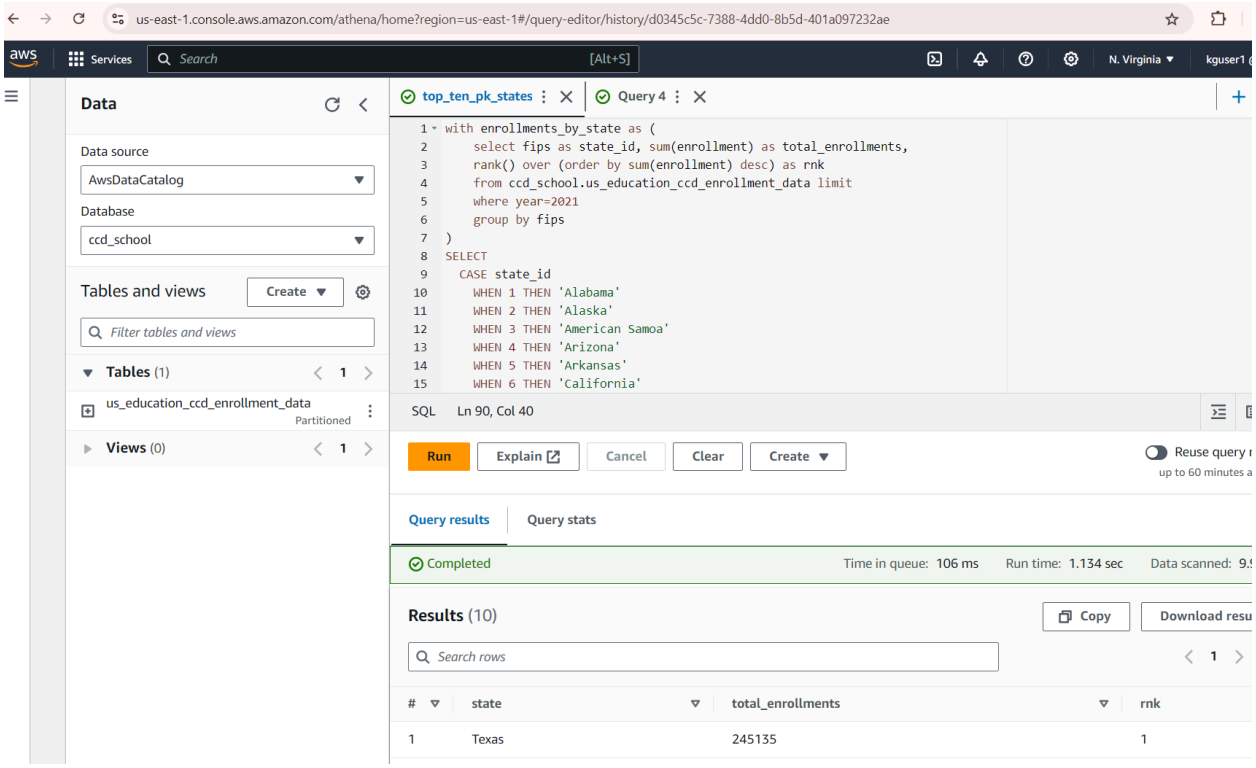
4. Scheduled the Lambda using AWS EventBridge scheduler to run every 12 hours. The scheduler passes the event with year and grade parameters to the Lambda function, this gives flexibility to expand the dataset with changing the code.
    a. `{ "years": [2020, 2021], "grades": ["grade-pk"] }`



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



6. AWS Athena is used to query the data



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
3. Look up tables could simplify querying the data for ids like states returned in the results.