Quality movies data ingestion

Project Description:

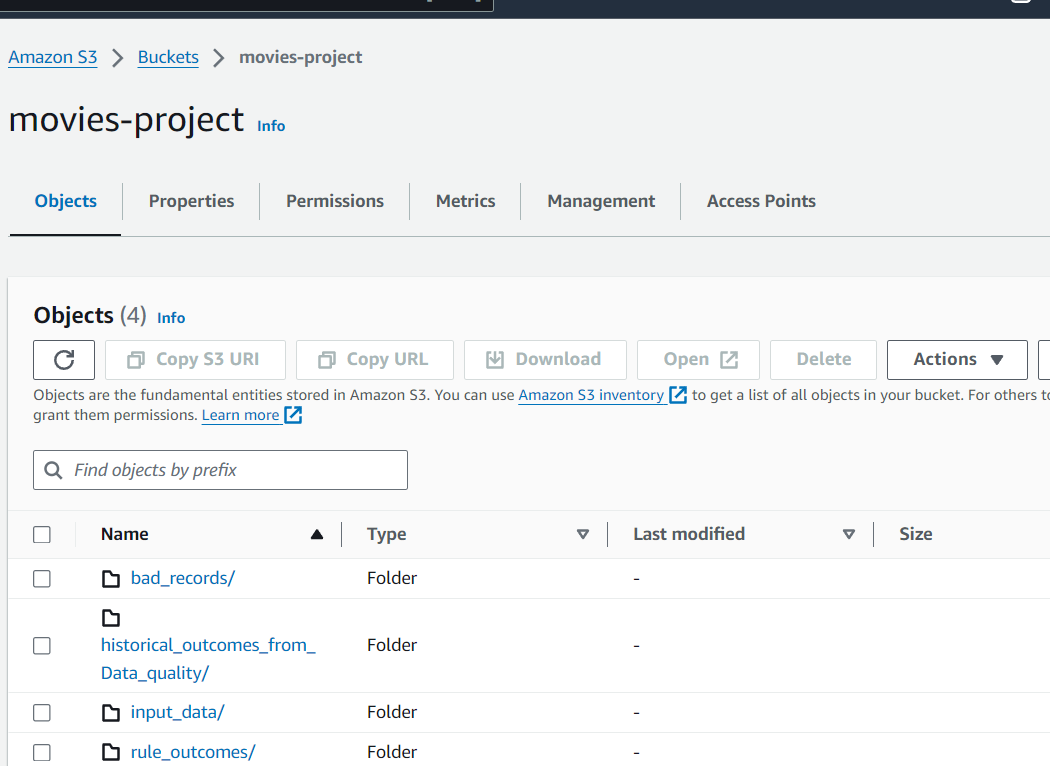
In this project, our focus is on ensuring the accuracy and reliability of our data through comprehensive data quality checks. Specifically, we'll be applying a predefined set of validation rules to assess the validity of the IMDb movies rating.csv file, which arrives daily. Upon the file's arrival to S3 , an event bridge rule will trigger a notification to a step function. Subsequently, this step function will kick off a crawler to start and execute its task. Once the crawler completes its operation, the step function will then initiate a Glue job. Within this job, thorough data quality checks will be conducted, allowing us to segregate the data based on its quality. This process enables us to determine the appropriate destination for records of both high and low quality. Finally, after the Glue job reads the data from the crawler, the refined dataset will be ingested into the Redshift warehouse for further analysis and utilization.

**Architecture:**



**S3**

🡪Create s3 Bucket i.e movies-project and inside that we create folder, so here in input\_data folder we are placing the imbd\_movies.csv file .



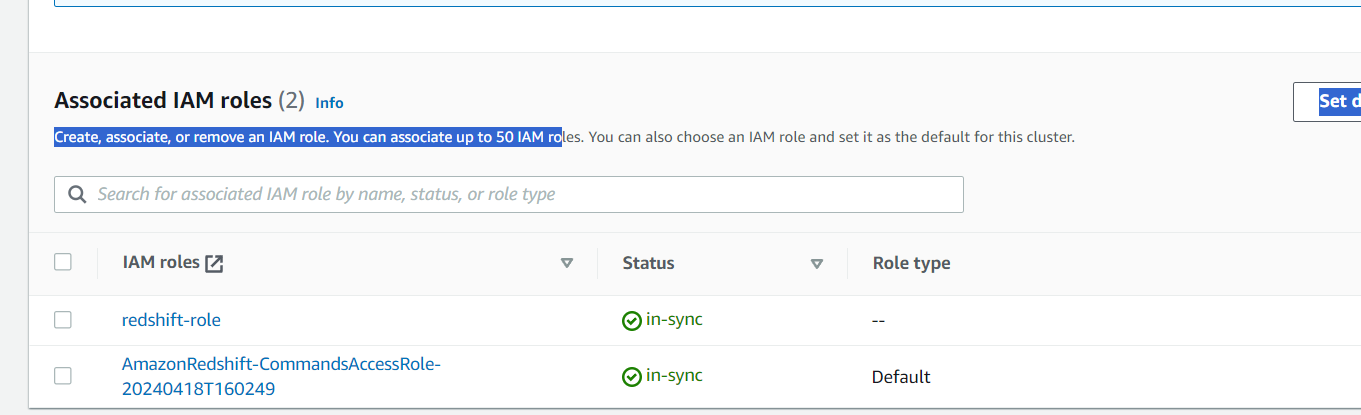
**🡪Redshift**

1.Next we need to create target table in redshift to ingest the data. So for that create schema and table in redshift data warehouse

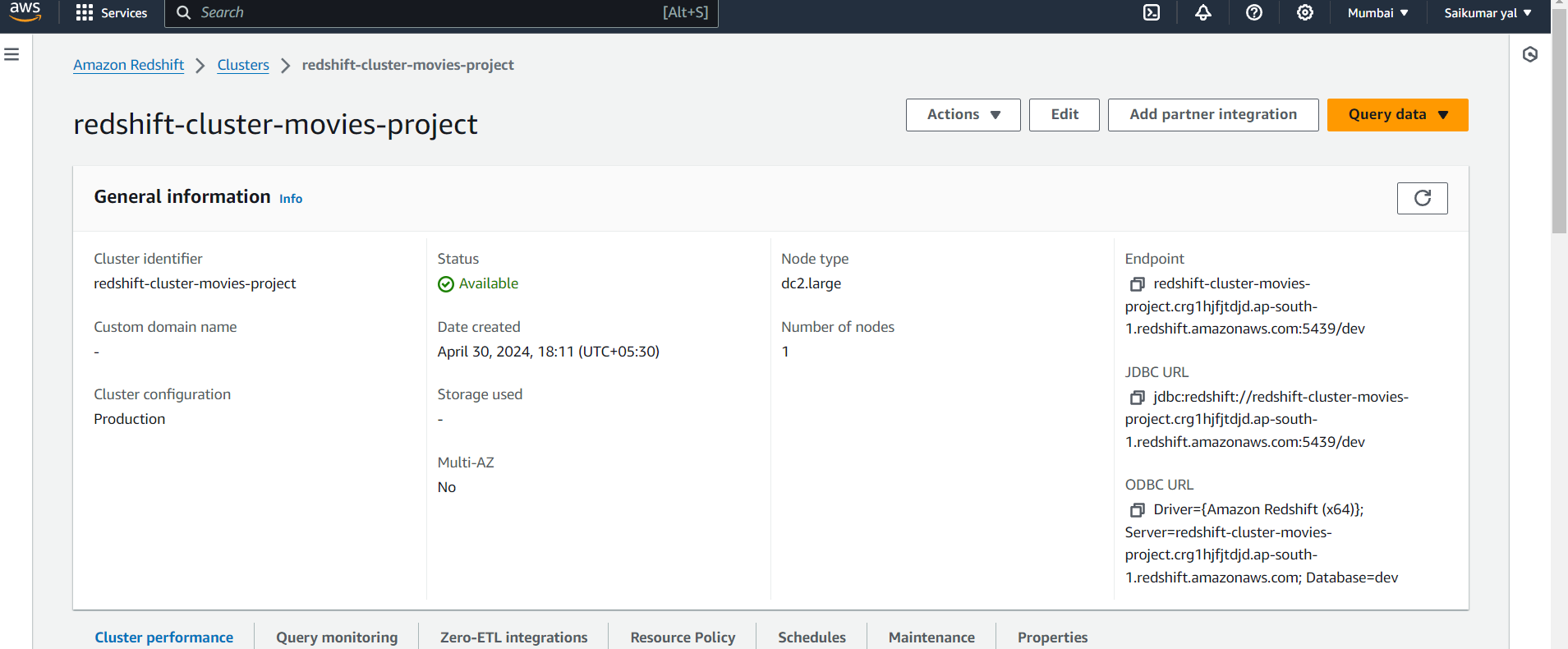
2.Firstly redshift cluster is created below u can see

3.While creating add IAM role with required permissions to create cluster

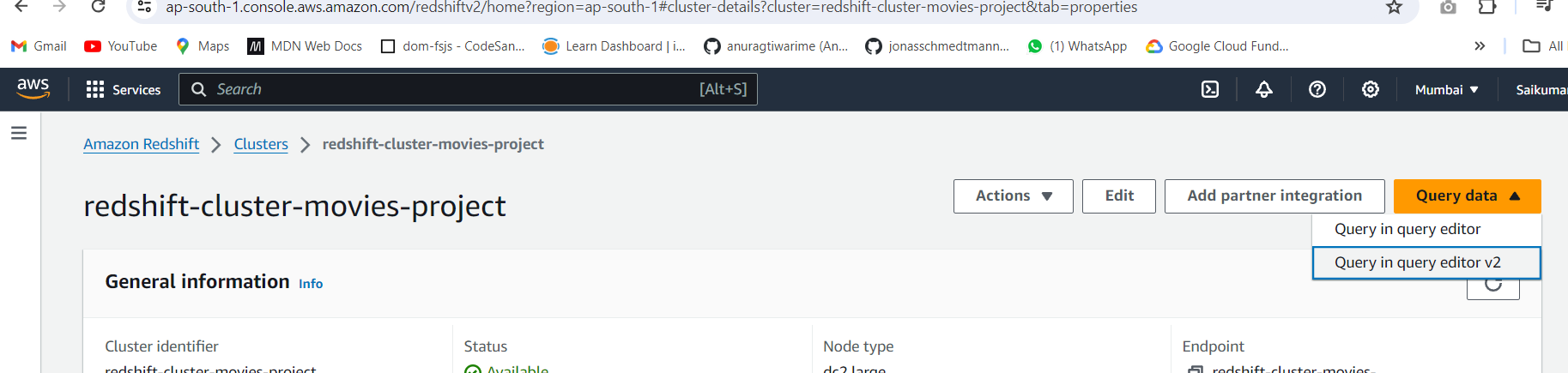
4. the permissions we added for role are



5.Cluster is Created and in below Properties tab we can see all details



6.Go to query editor to create Queries i.e Tables

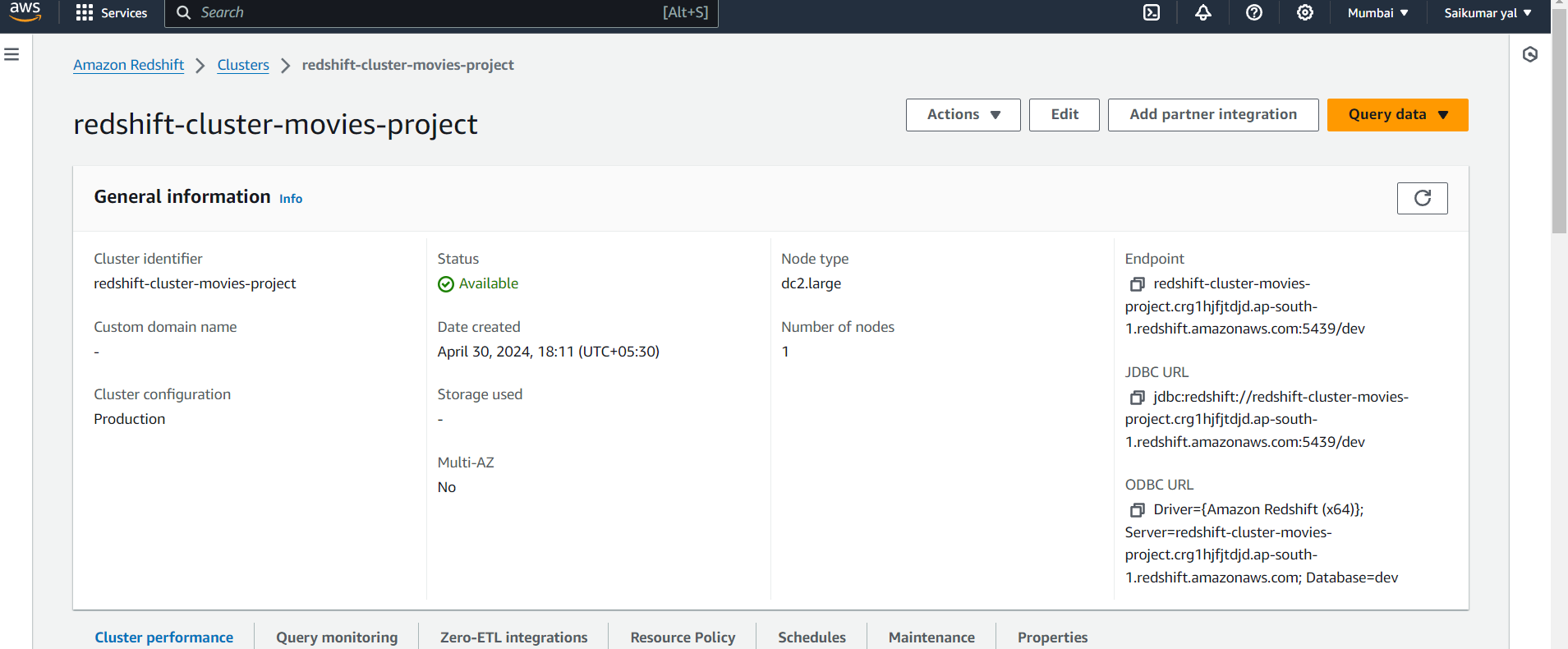


7.Open query editor and create Schema and Table and Run the query so Table and Schema will be created In redshift in that .

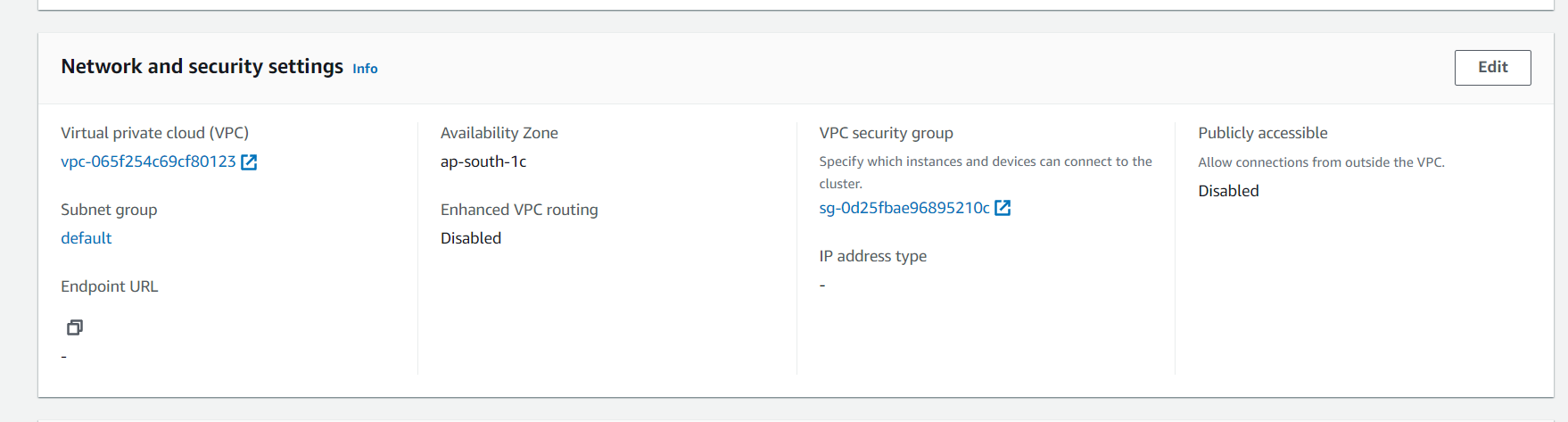
We have used this Query below



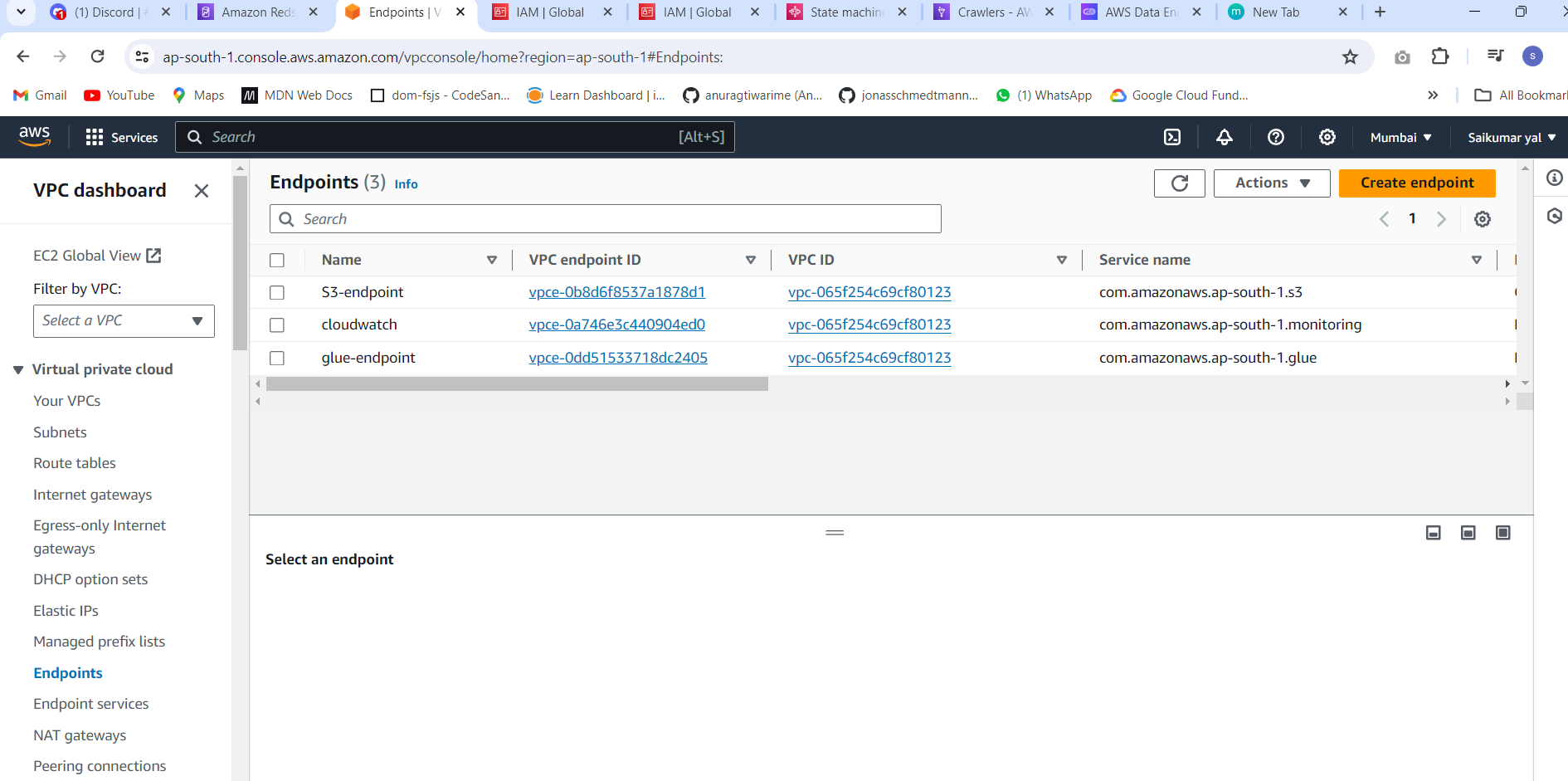
Now go back to cluster we need some permissions and rules we will add



8.Go to properties and scroll down 🡪 Network and security settings here we will add some endpoints and inbound rules needed for the project

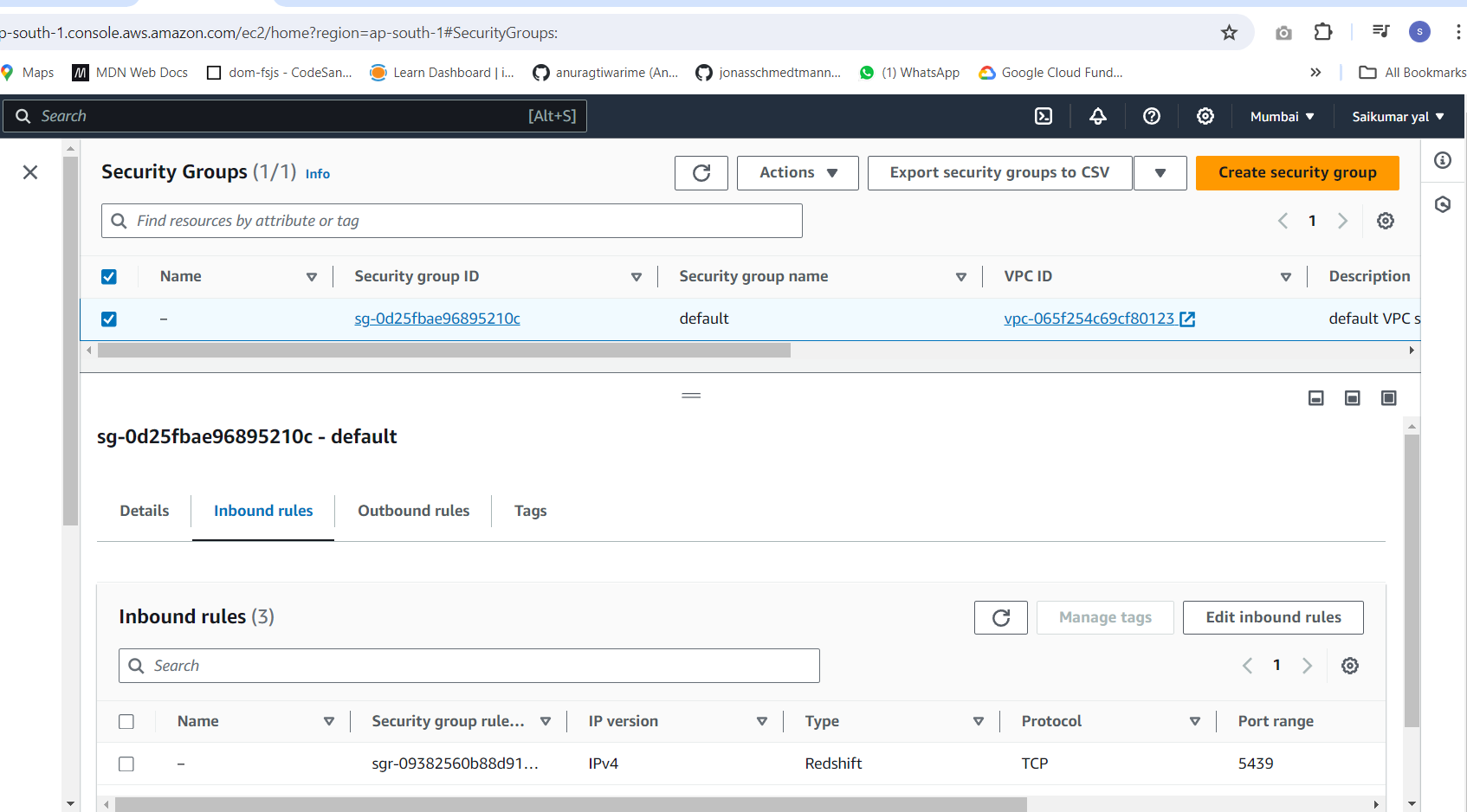


9. Click VPC and go to End points and add these endpoints

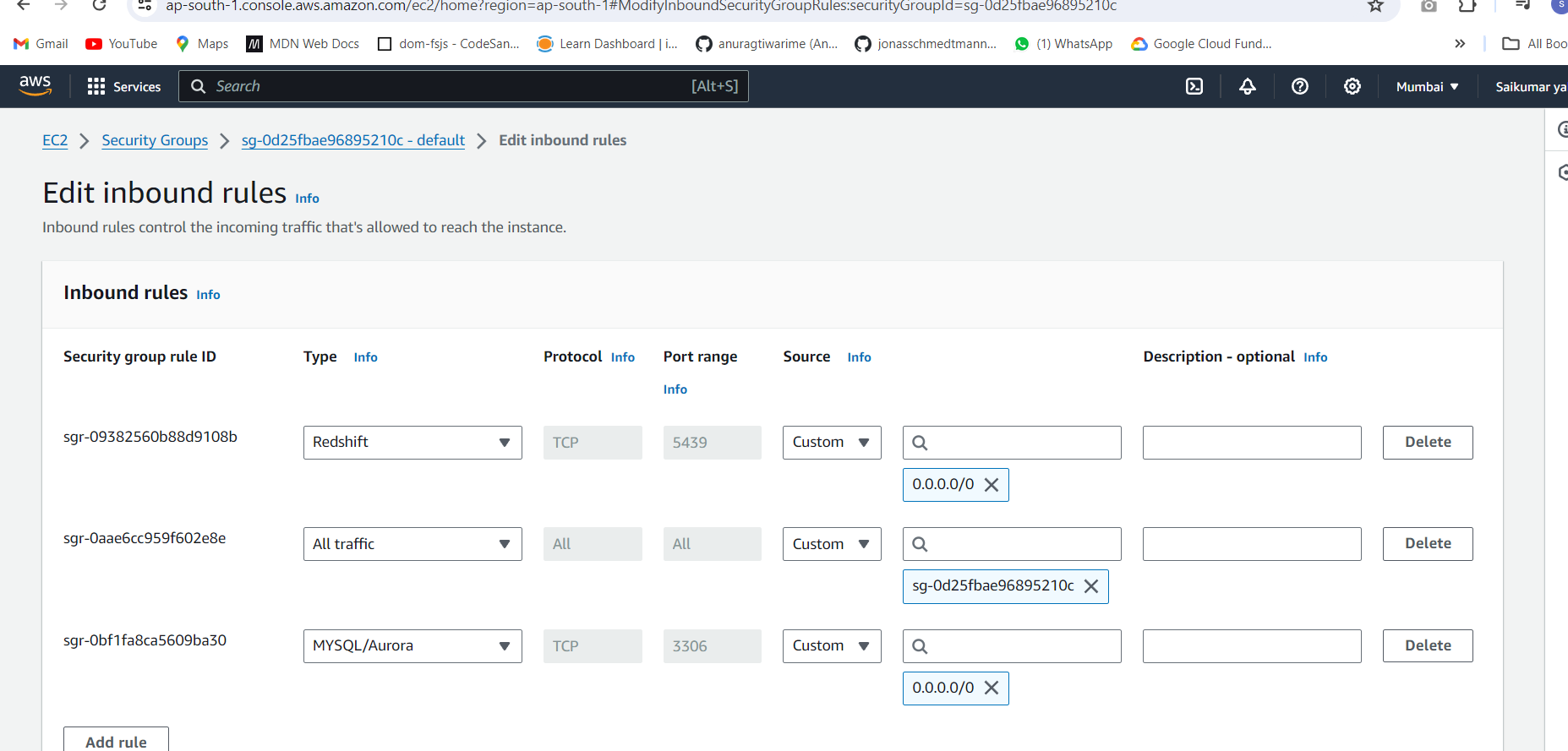


With the same subnets and security groups of glue job

10.Next in properties 🡪 open Security group and select that and go to edit Inbound rules and add following rules

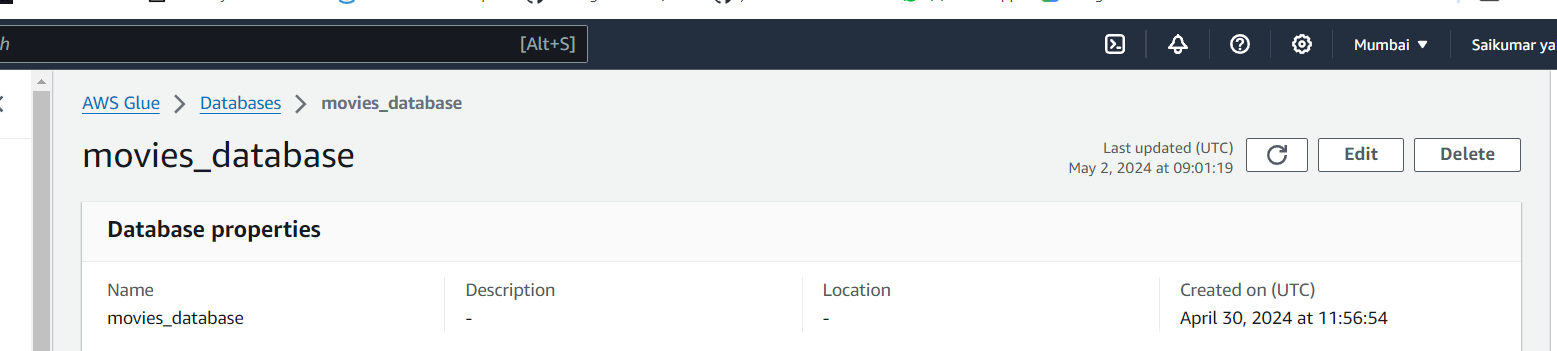


We added these rules



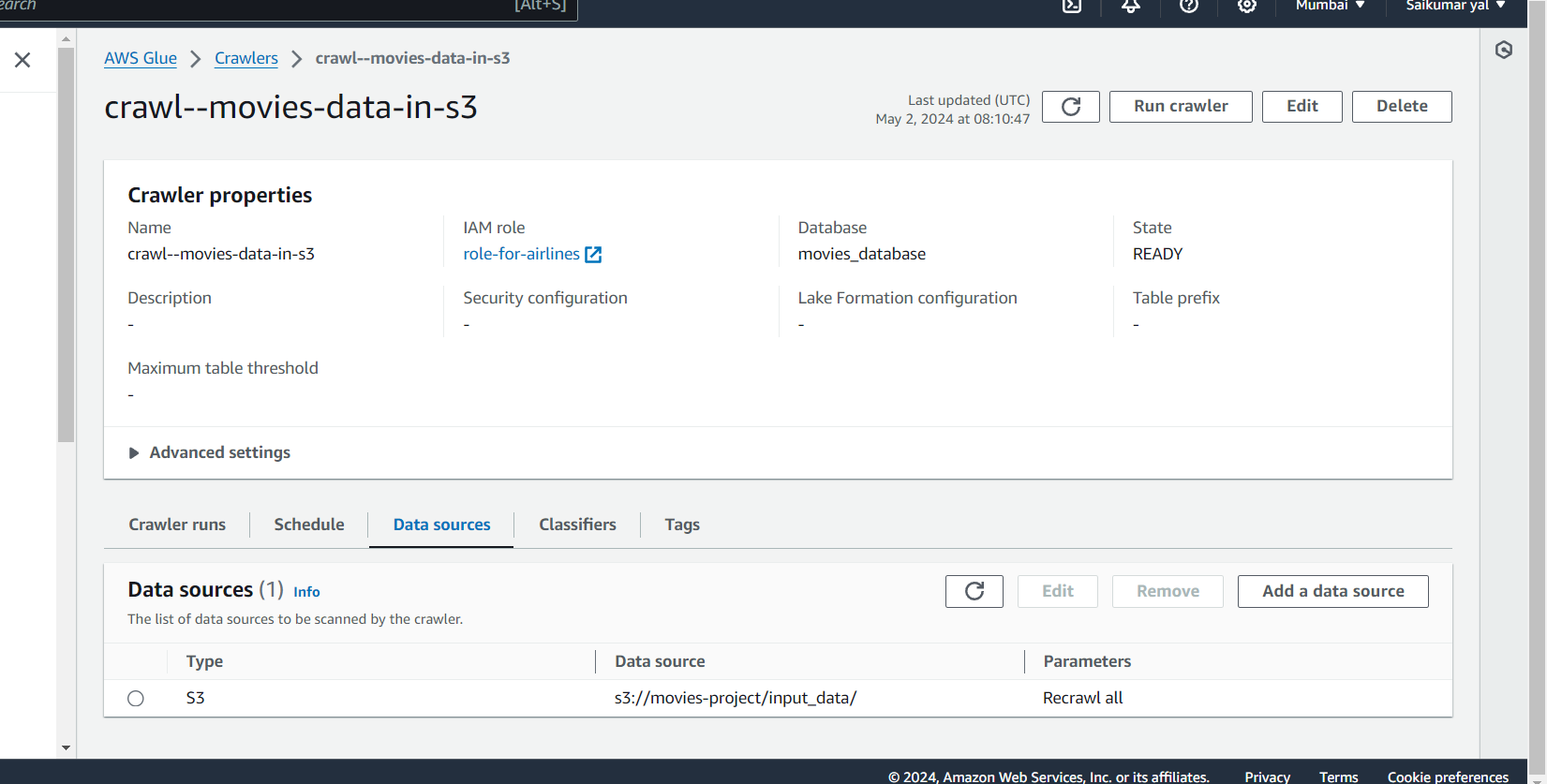
**GLUE**

1.In glue first we create database to store metadata i.e below we

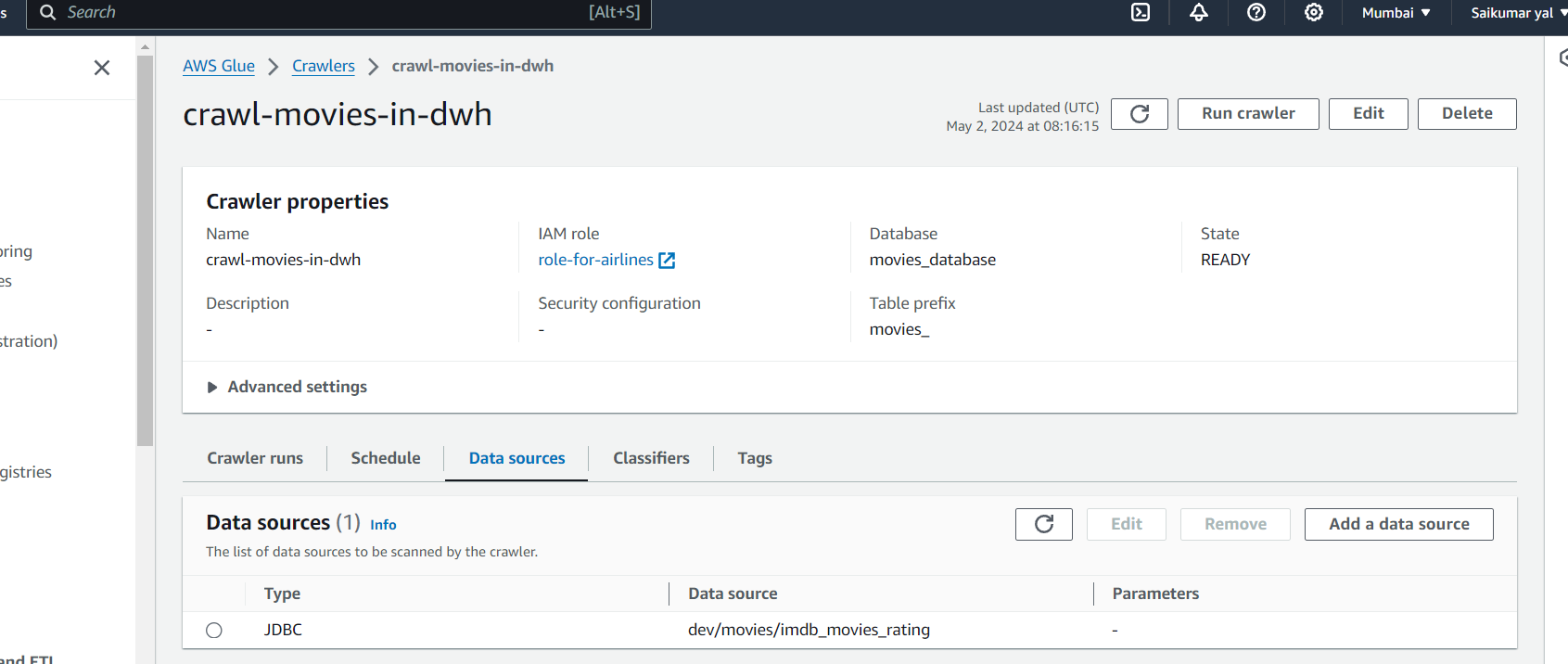


2.Next in Glue part we create crawlers we use 2 crawlers in this project i.e 1 crawler will crawl the file in S3 and another crawler to crawl redshift

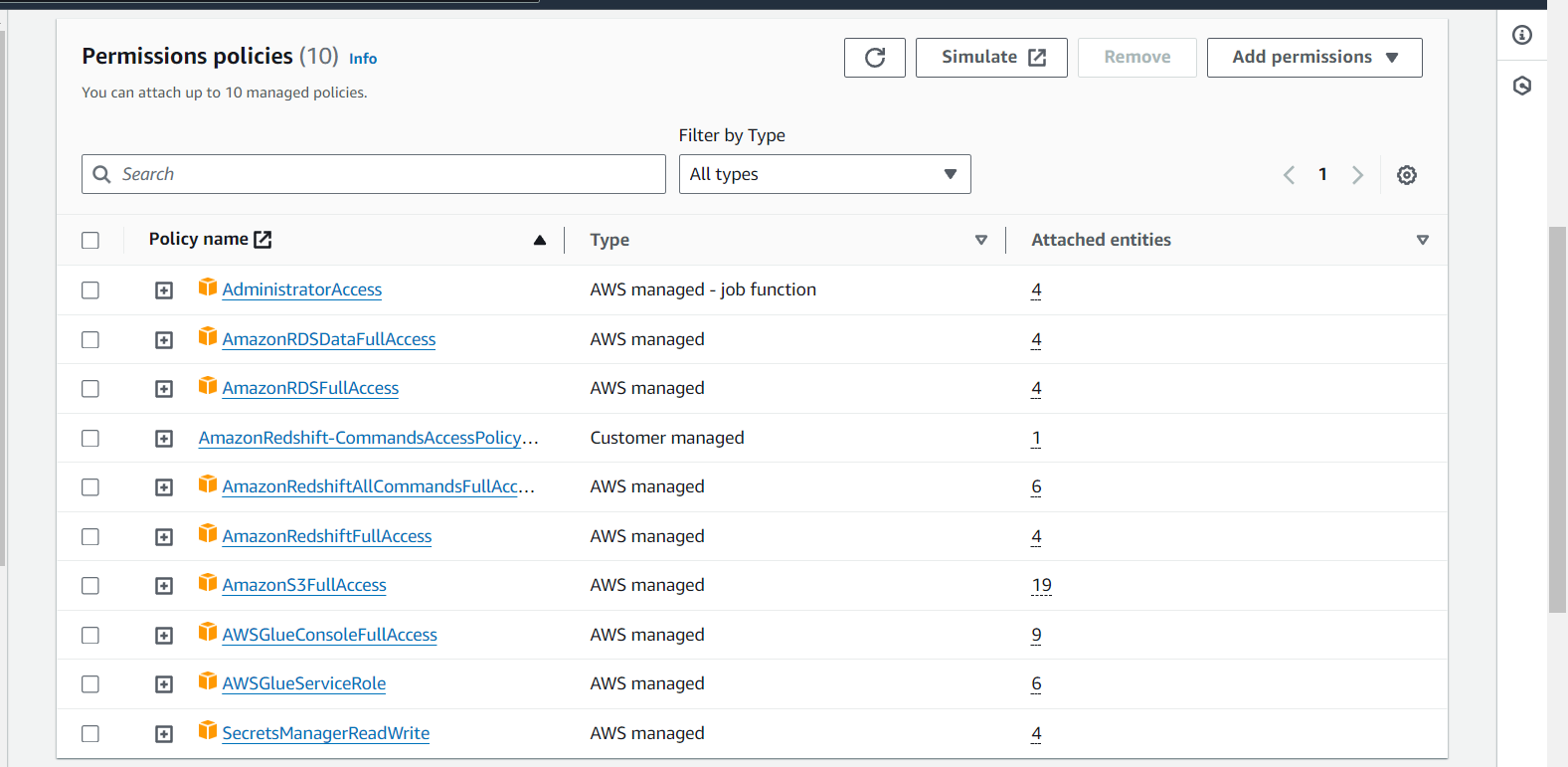
🡪Below Crawler is created and ran which crawls s3 and collects metadata and stores here in glue

🡪Data source is added i.e movie-project bucket inside that input\_data folder where we place imdb.movies.csv file 

🡪Next Another crawler is created below which crawls redshift data warehouse and datasource is added i.e JDBC connection is used because of redshift, and in that data source is given as dev/Schema/Table name u can see

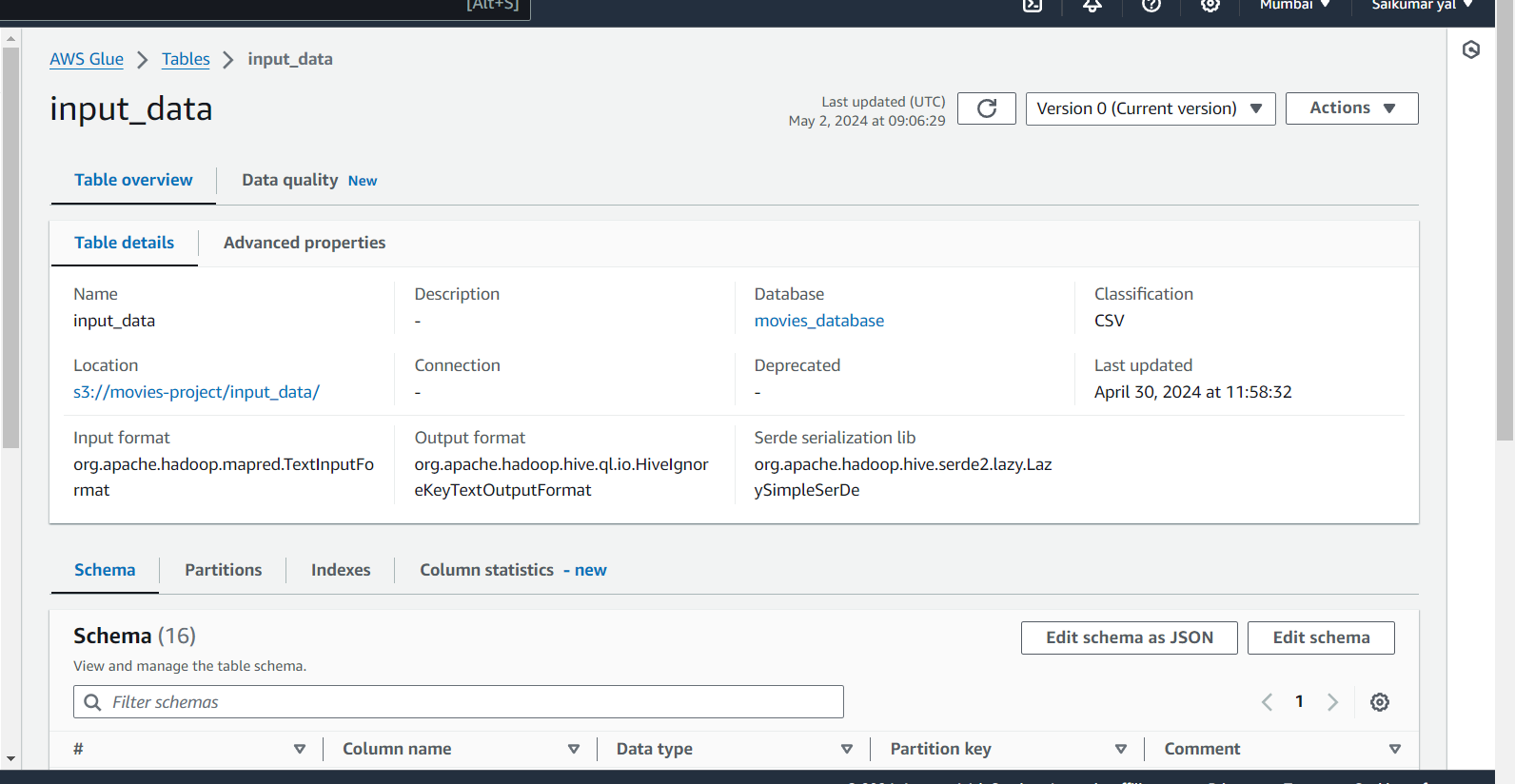


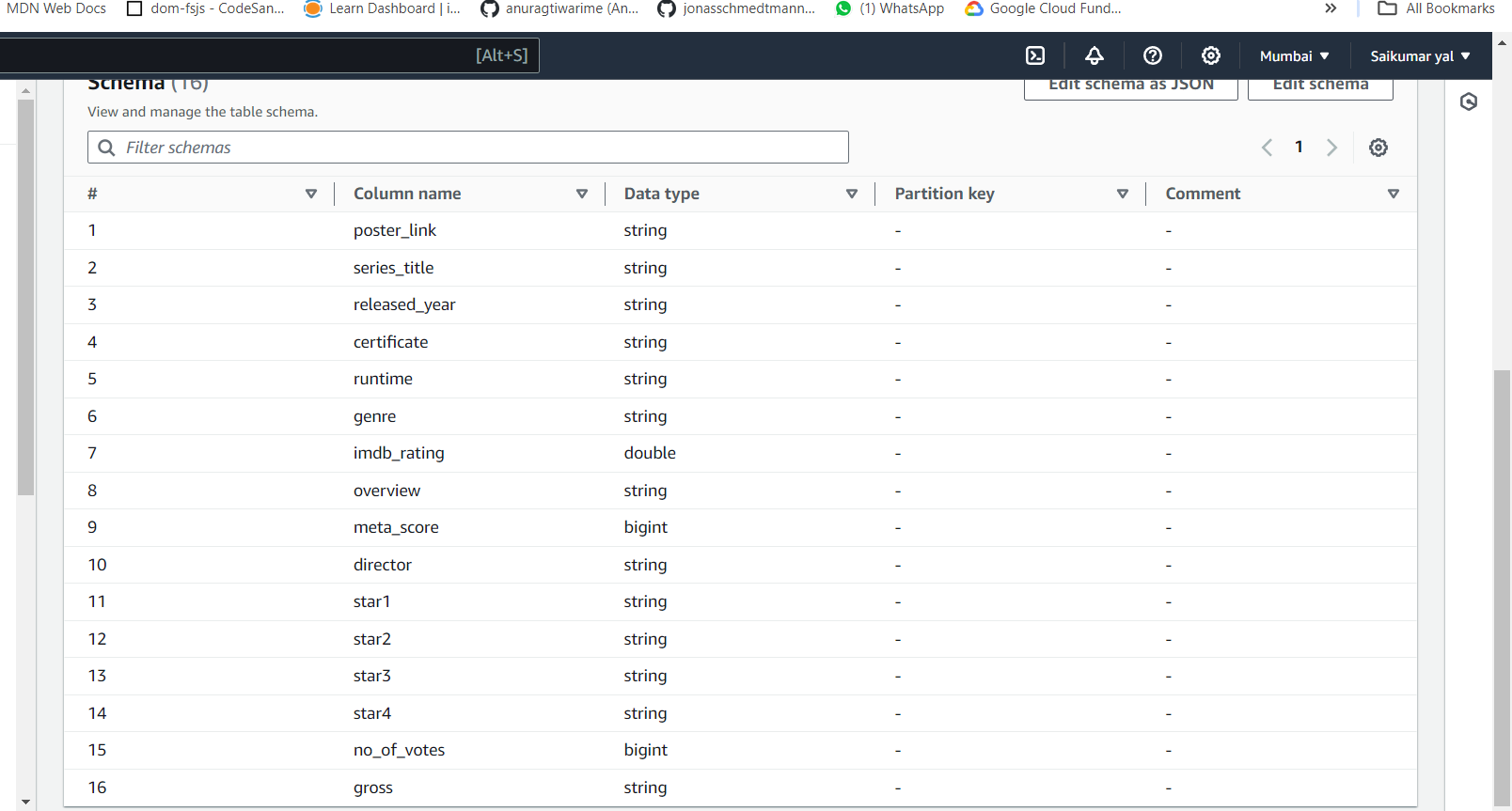
🡪 While creating add IAM role that has following permissions



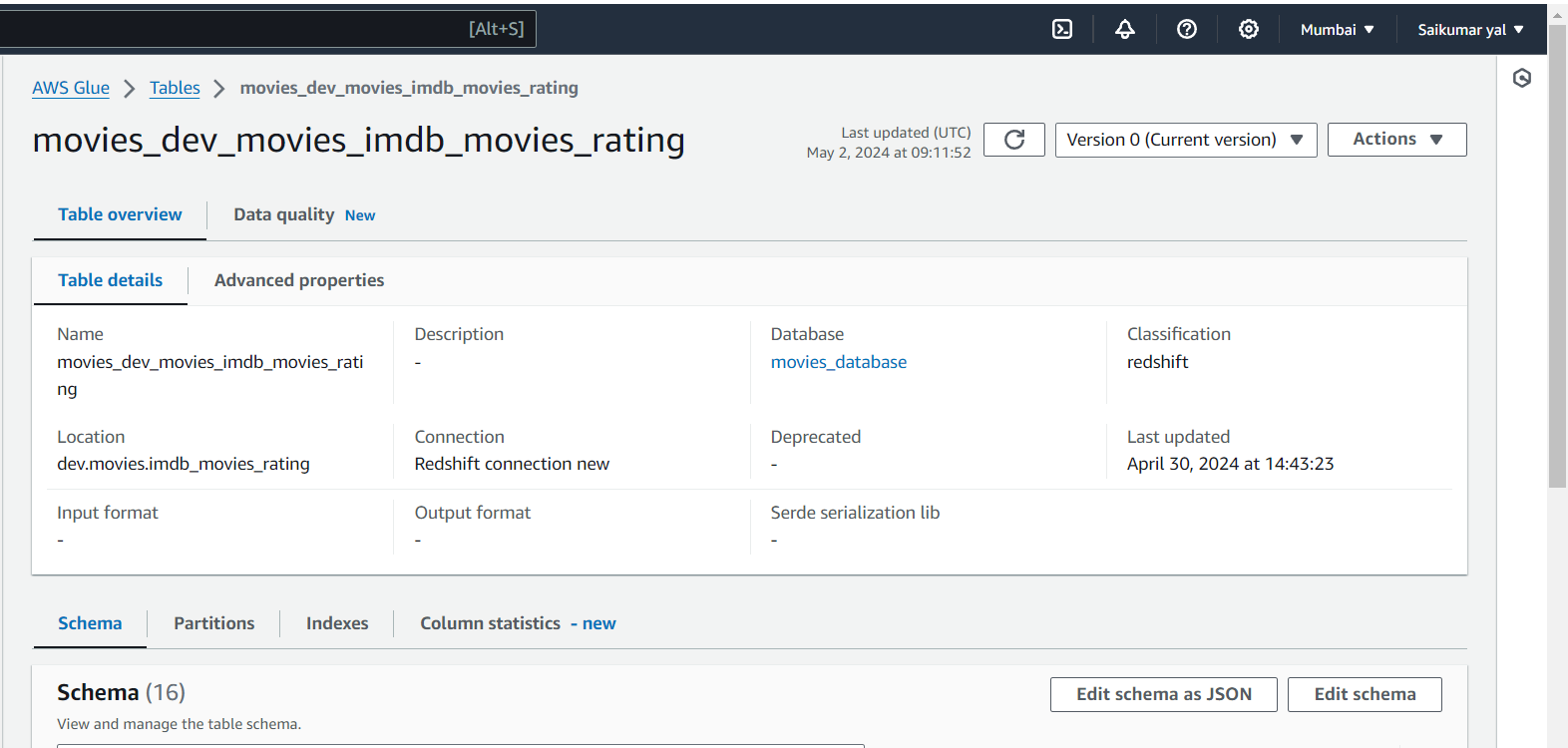
Next after crawling go to Tables and see whether tables are created from crawlers

🡪this table is created from S3 crawler we can see schema is created





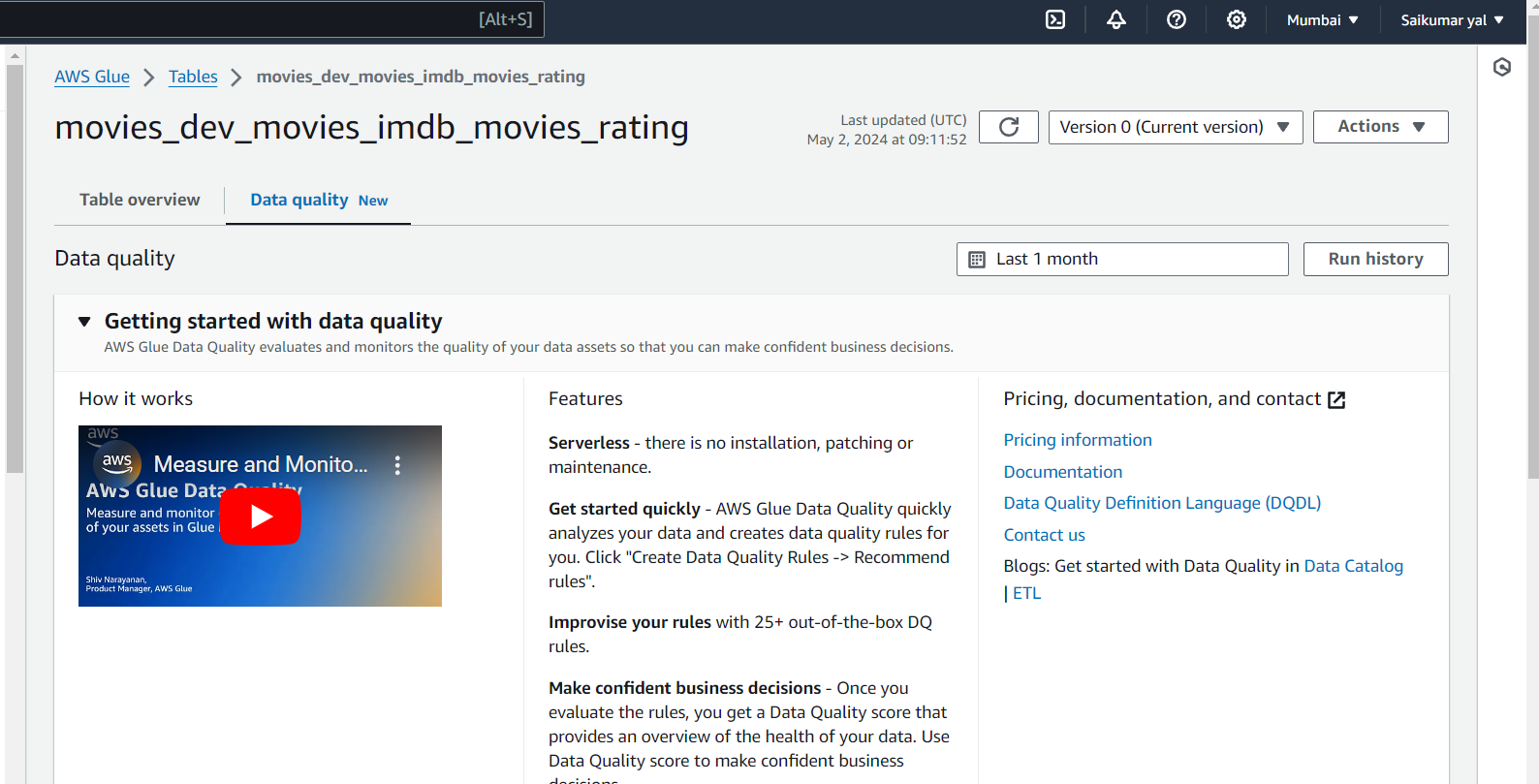
🡪And next this table is created from another crawler i.e by crawling redshift



**Data Quality**

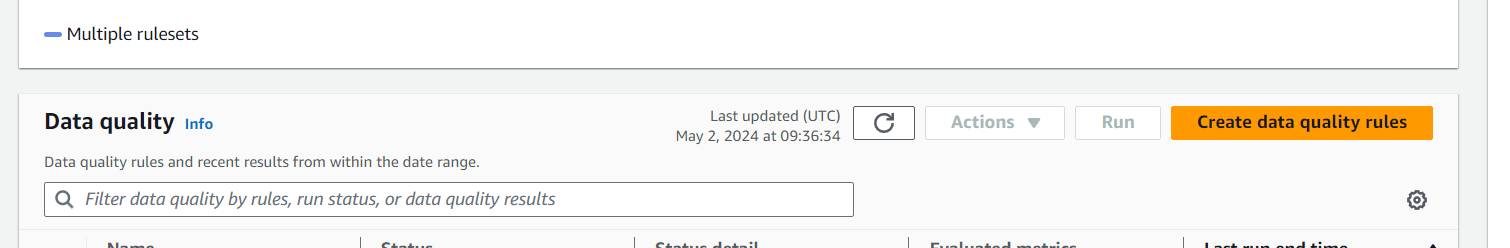
Here we can see in table Beside Data Quality

data quality can be used to analyse historical data, similar to this a particular transformation has given in glue Etl where we can make use of data quality execution and based on that we can take further decision like if some record is failing data quality check we determine whether to pass it to another process or redirect it elsewhere.Top of Form



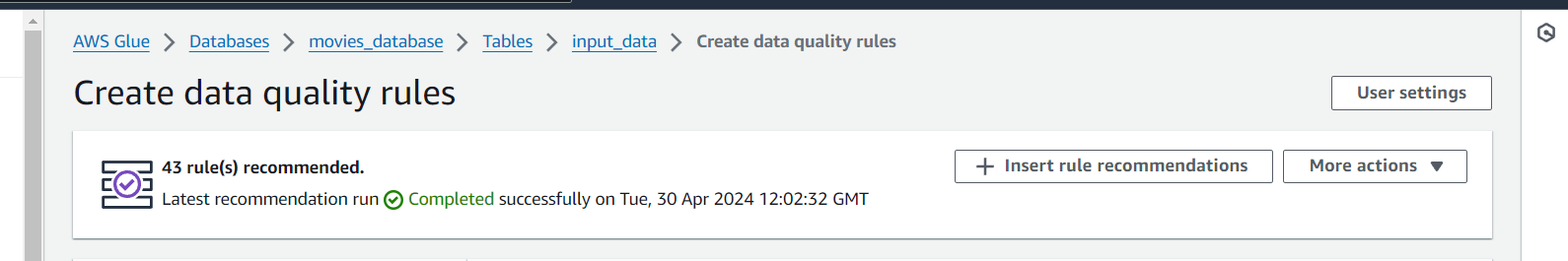
Now we will do data quality check of historical and incremental data so that we can know the percentage of data quality

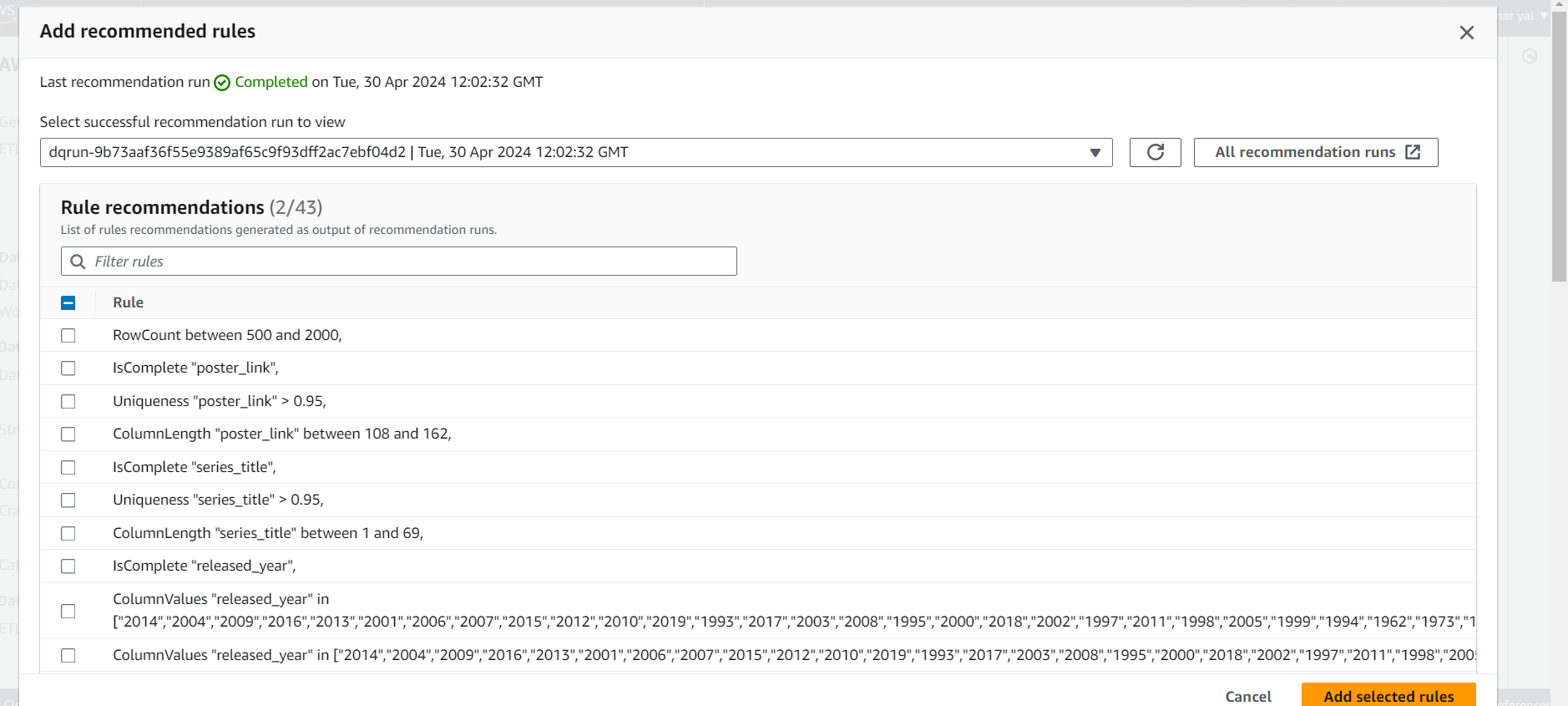
* Create data quality rule



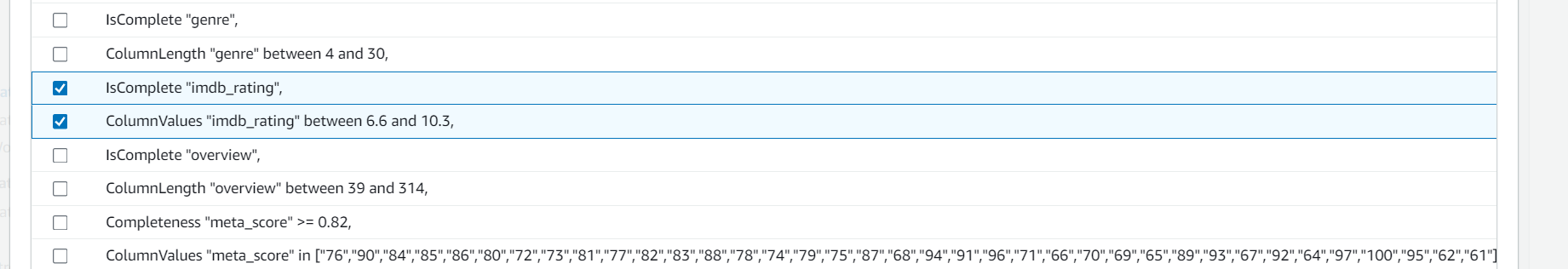
For first time we should click recommended rules and add IAM role this scans the our data set and gives rules ,So after scanning it gives rules recommended and from insert rule recommendations we can add required roles

In that click insert rule recommendation and select 2 rules to see data quality

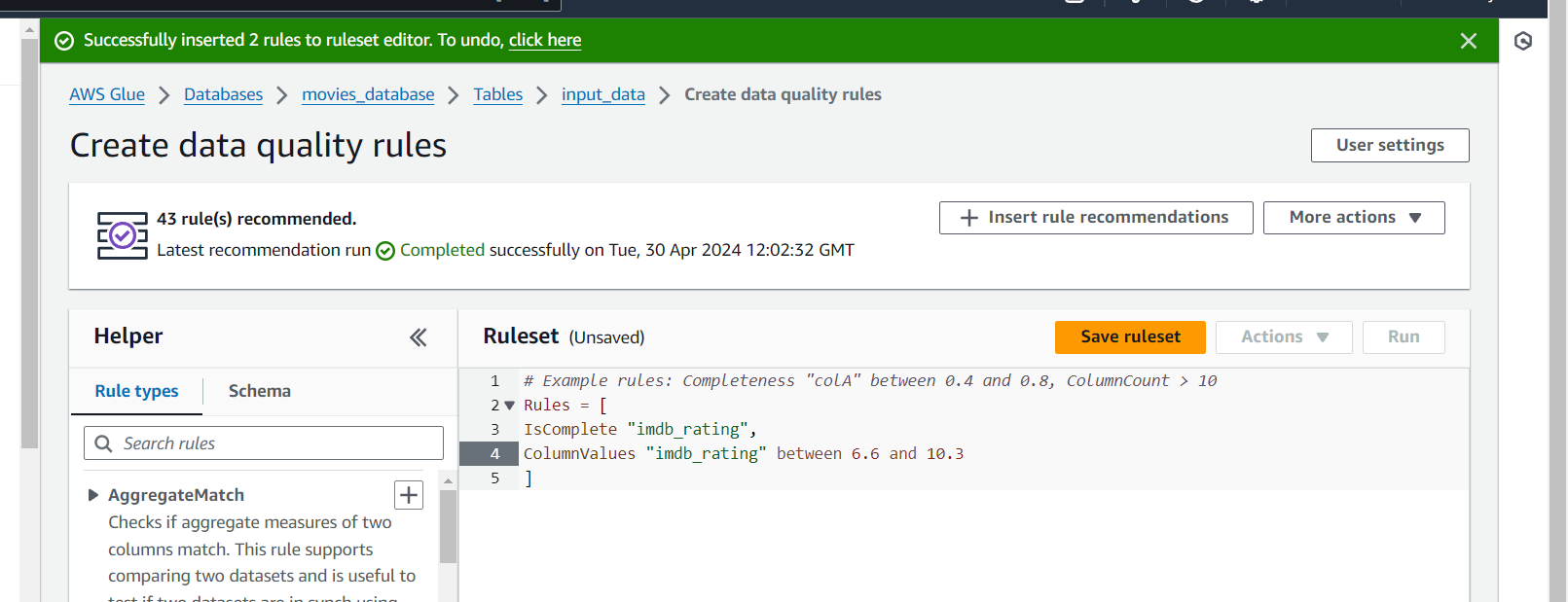




From list of rules we are adding these 2 rules



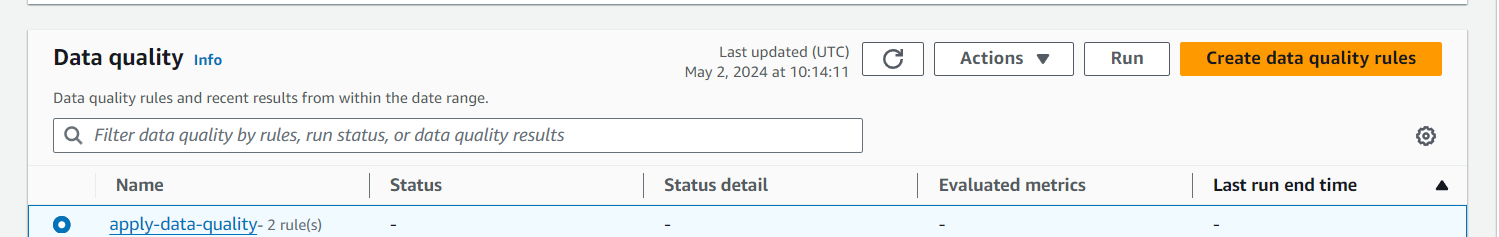
And click save ruleset

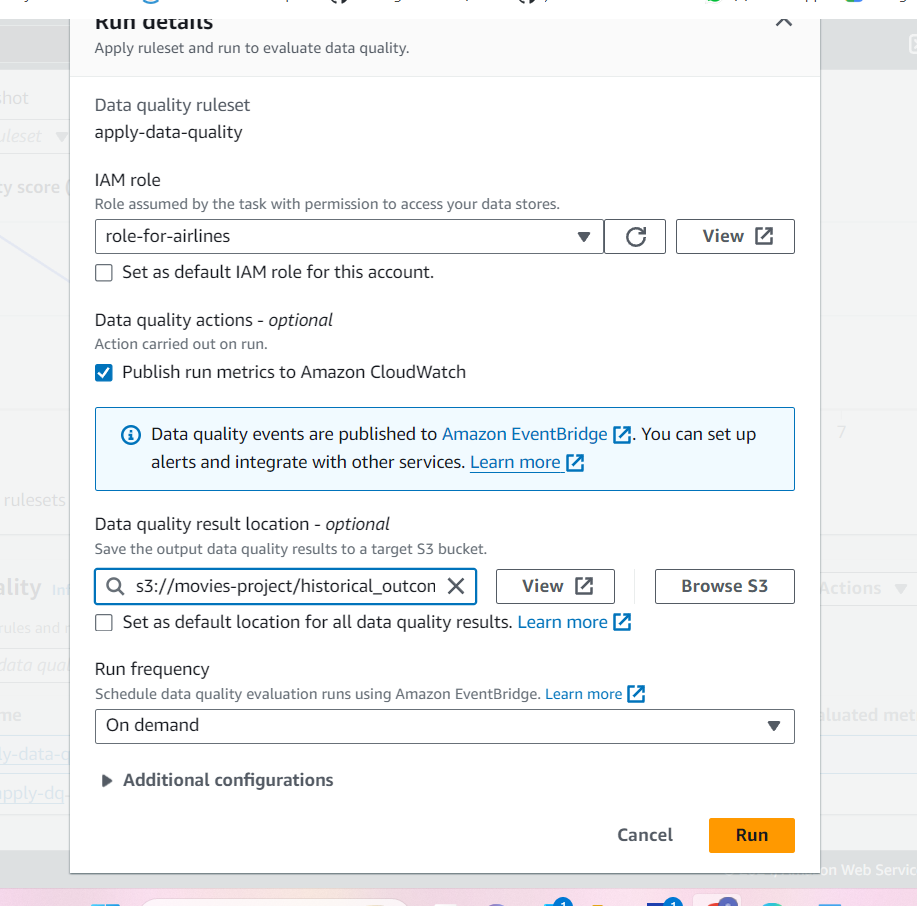


And now select and click run it will ask IAM role attach that and for the outputs from this ruleset we are placing below S3 path { s3://movies-project/historical\_outcomes\_from\_Data\_quality/}

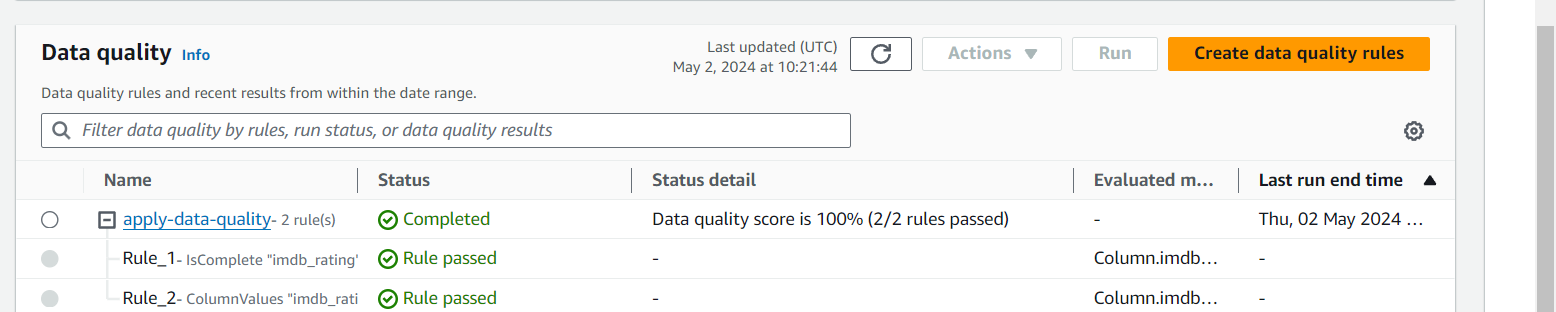
Here in this movies-project bucket the historical\_outcomes\_from\_Data\_quality folder is placed to receive outputs from this rule

And run



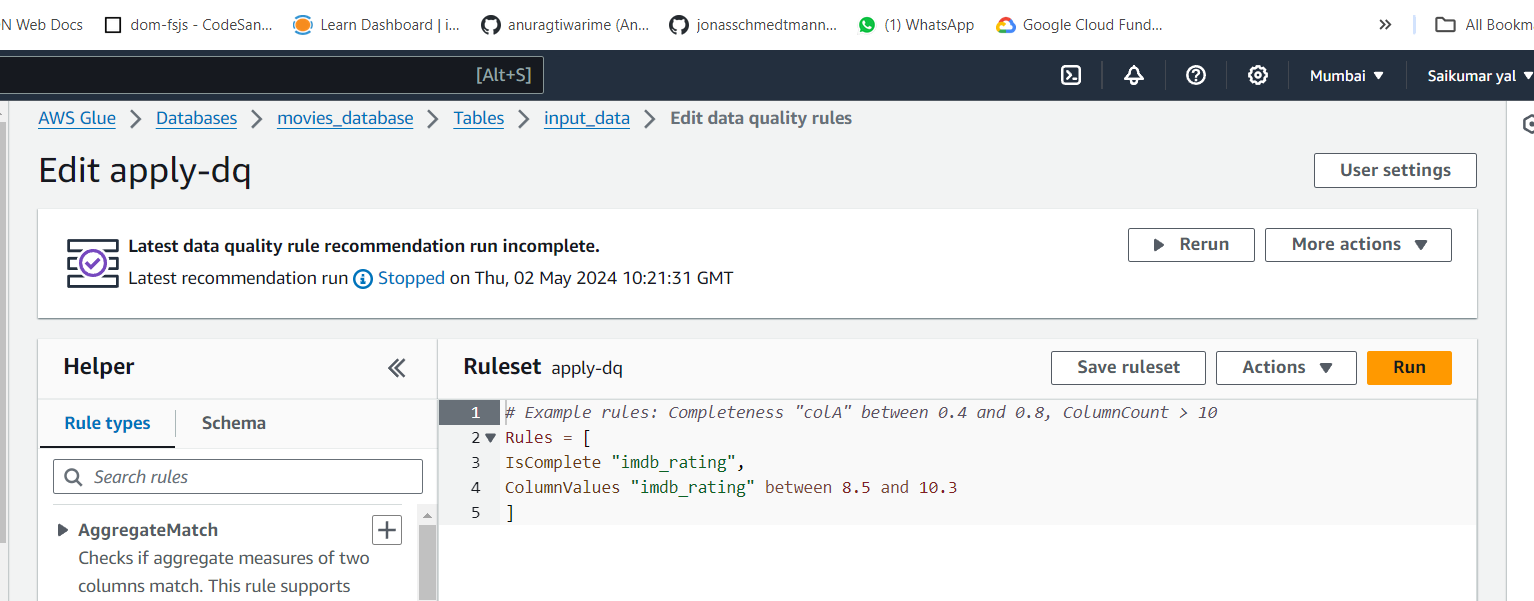


We can see both rules are passed and data quality score is 100%

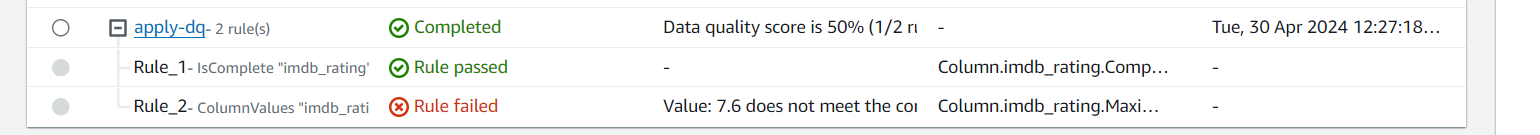


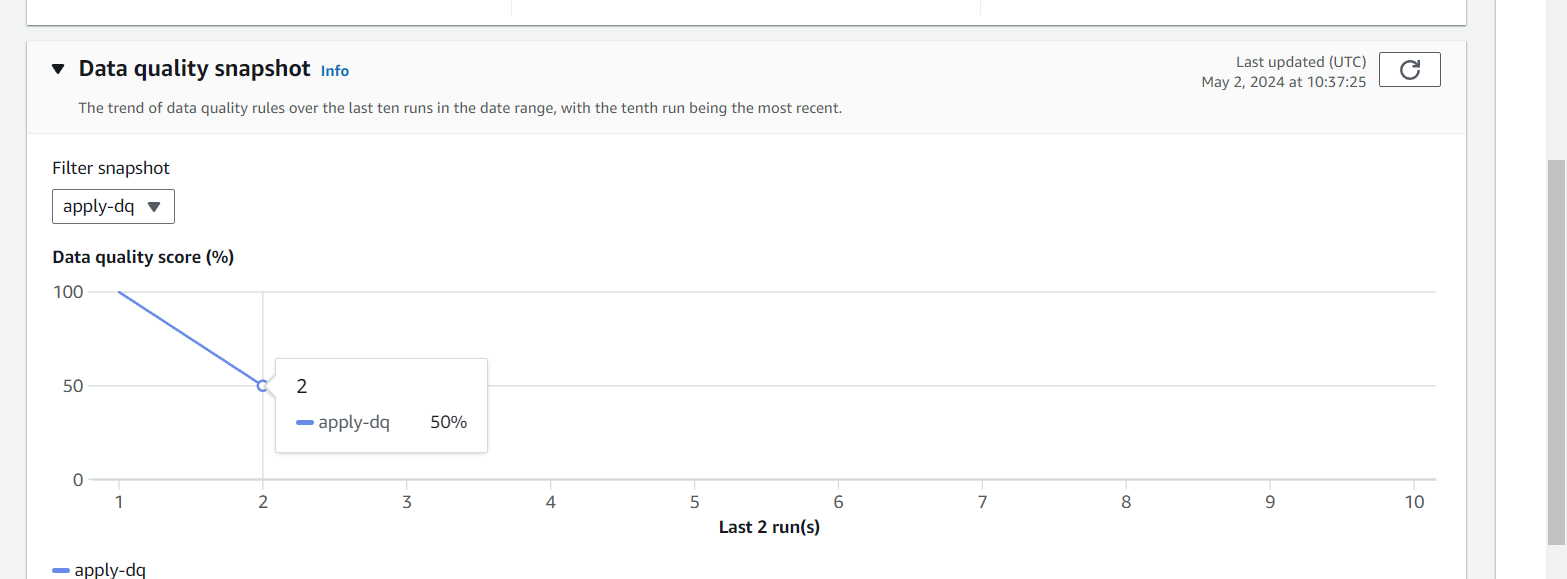
🡪Now we edit some changes in 1 rule and we will see is that rule failing or passing and data quality

We are editing the Rule\_2 that we are changing imdb\_rating here we increased to 8.5



After ran we can see Rule1 is passed and Rule2 is failed and reason is Rule 2 has some records below 8.5 so that this rule does’nt satisfy the criteria and failed and data quality decreased to 50%

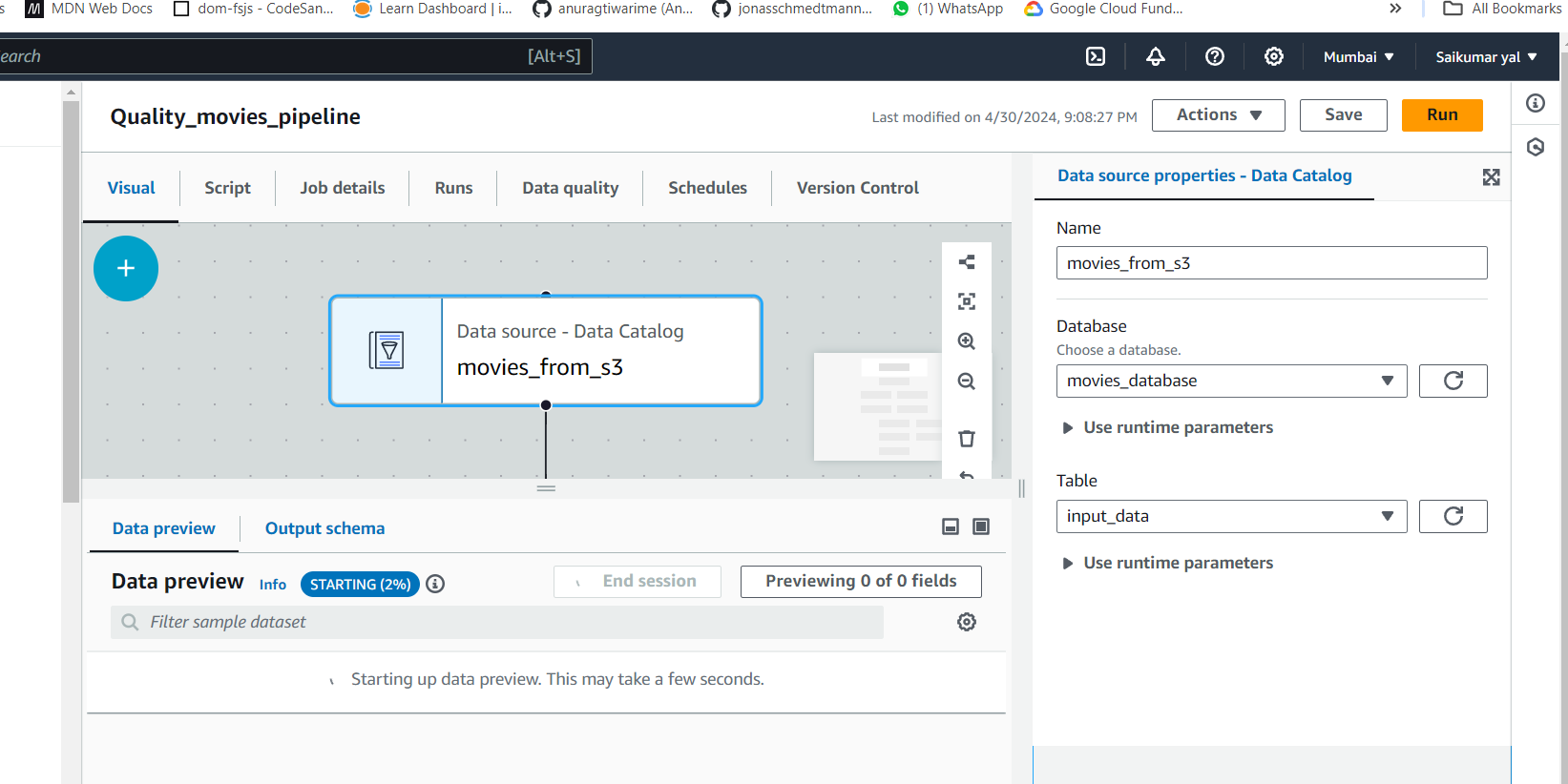




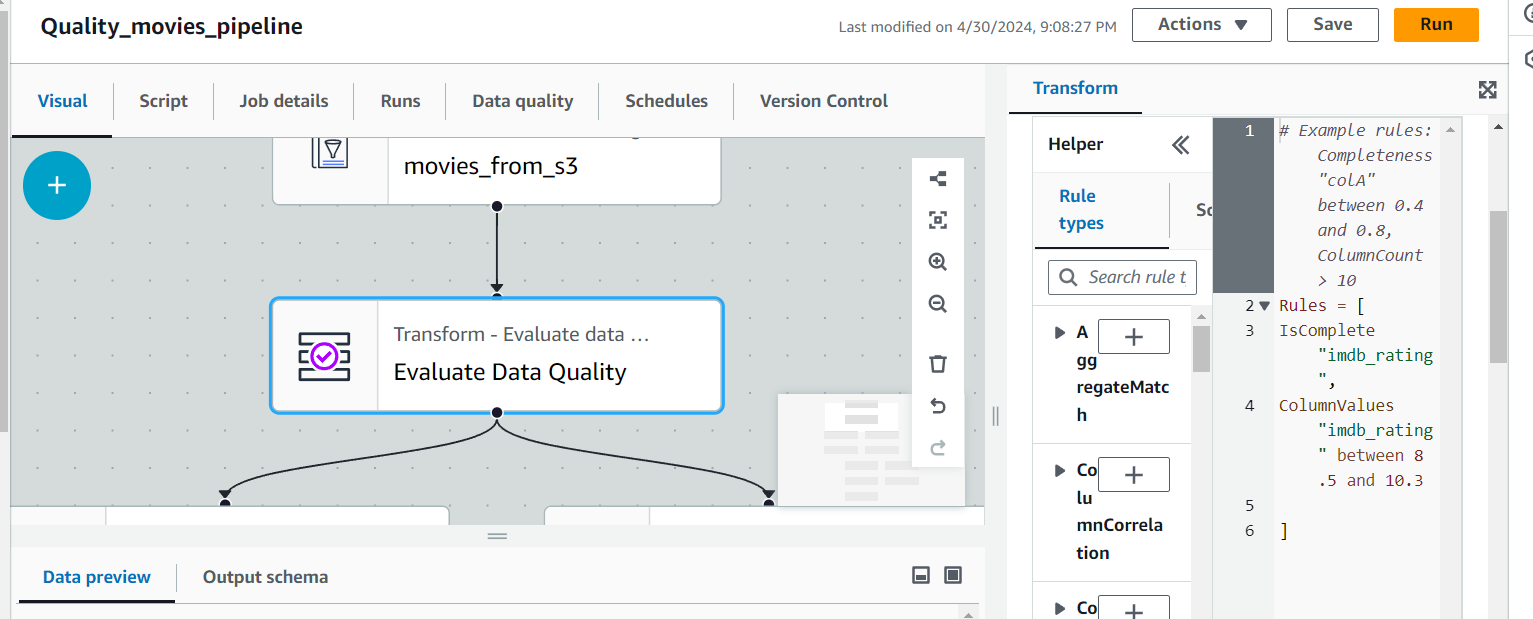
**🡪🡪Glue**

Here go to Visual Etl and we will create Etl pipeline here

First we will add data source and name it and choose database and select the table that crawled S3



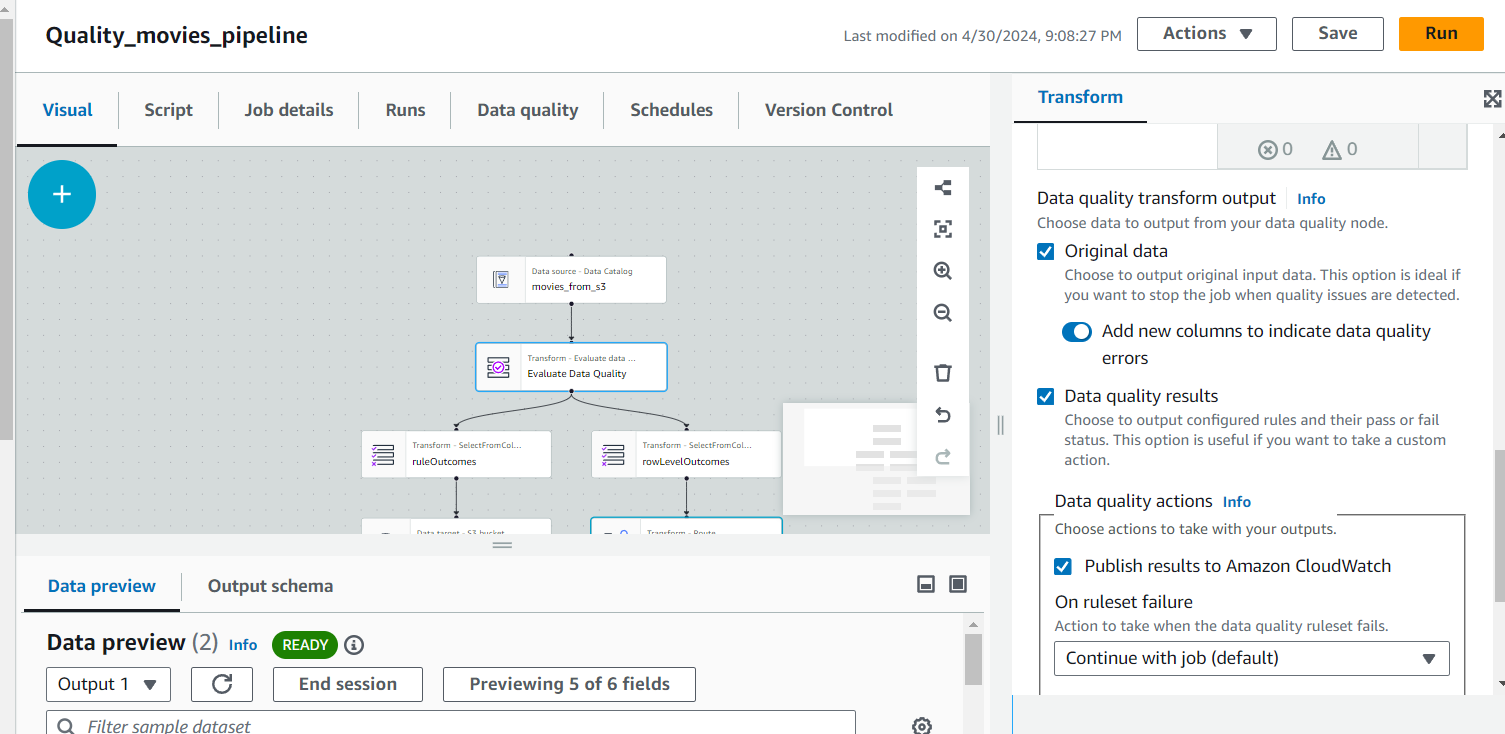
Next in Transform we will add Evaluate Data Quality and we will copy paste the ruleset from data quality rules here



**And next step**

Select the publish results to CloudWatch and Data quality results with their status i.e pass or fail that come in rule Outcomes

And will select Original data and also select add new columns that are failing and that will come in rowLevelOutcomes

****

And for Data quality results that i.e ruleOutcomes results we will send to S3 Bucket to store them so we will add bucket path s3://movies-project/rule\_outcomes/ --

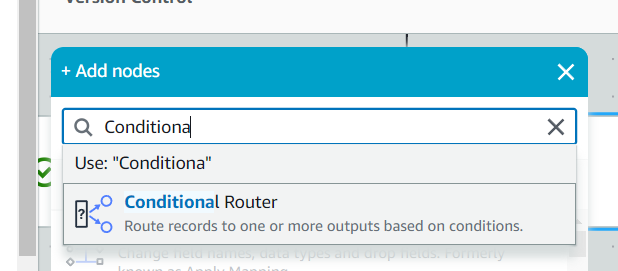
Bucket – movies\_project, folder – rule\_outcomes



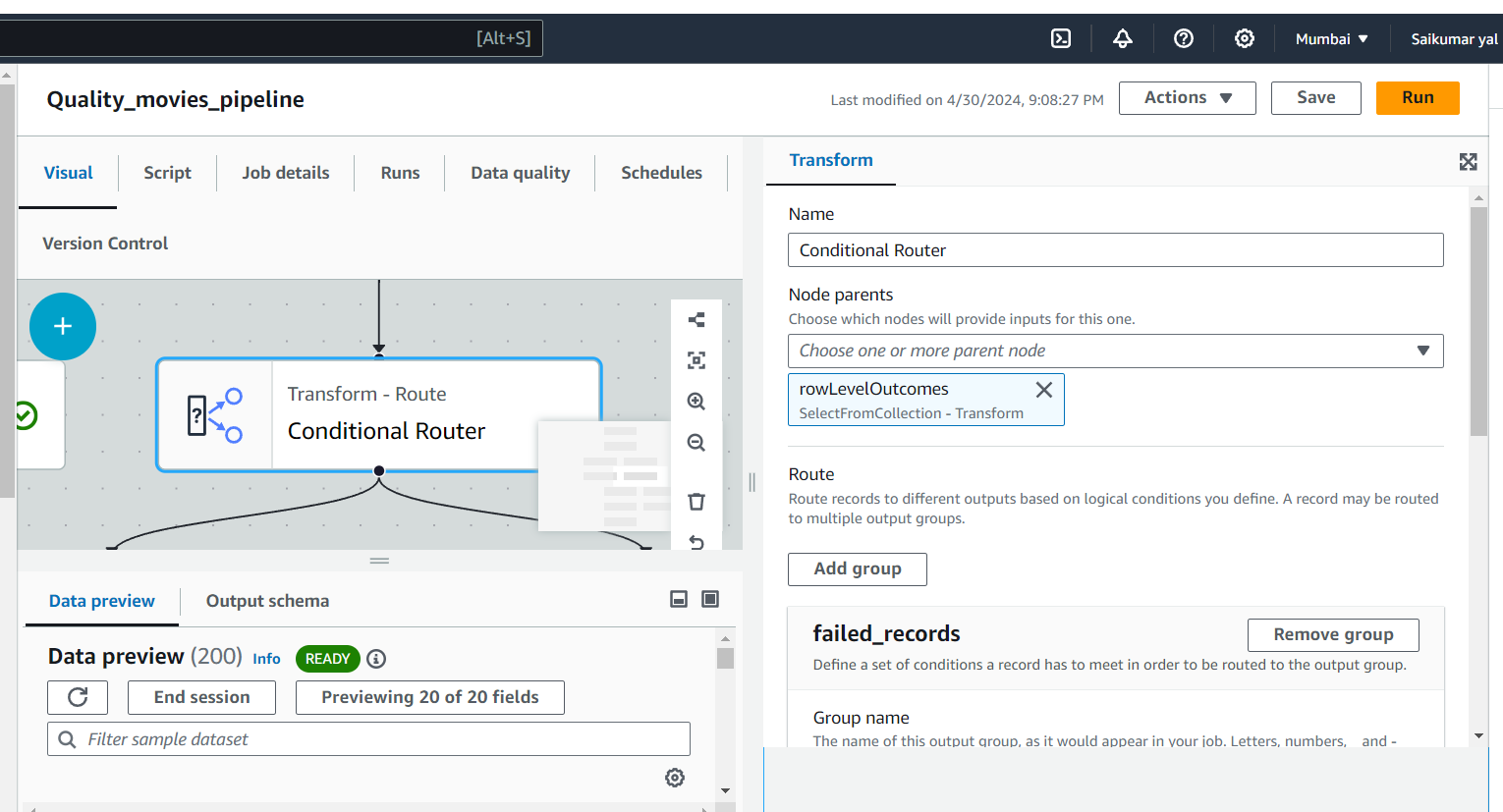
And from here if we want we connect to Athena or any other source to query this results

Next

For rowLevelOutcomes we add Transformer i.e Conditional Router to route records to one or more routes based on some conditions such as passed or failed

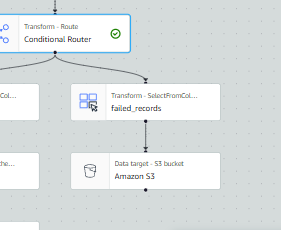


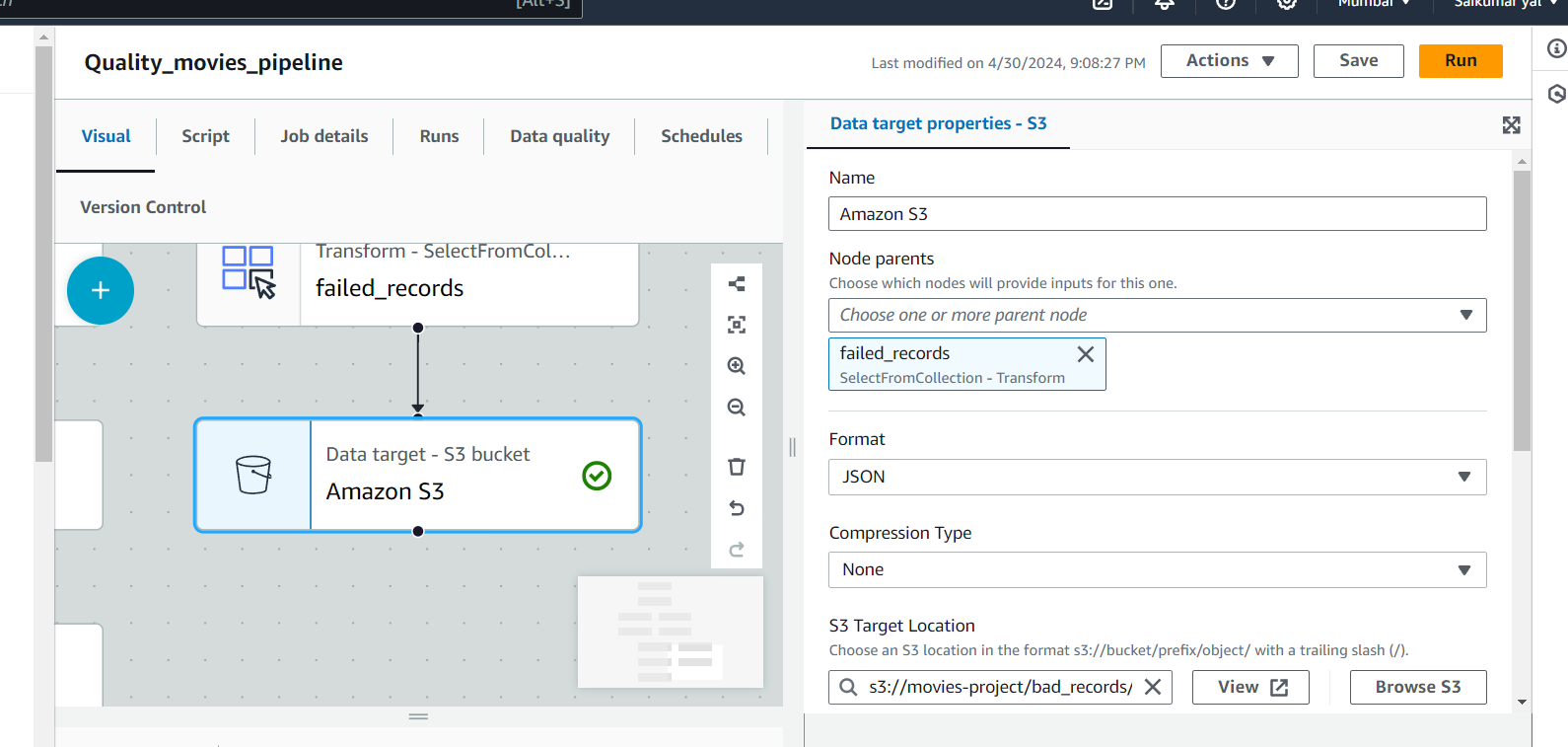
Select Add group in that name it failed\_records and in that in condition choose Data Quality Evaluation result and in key failed



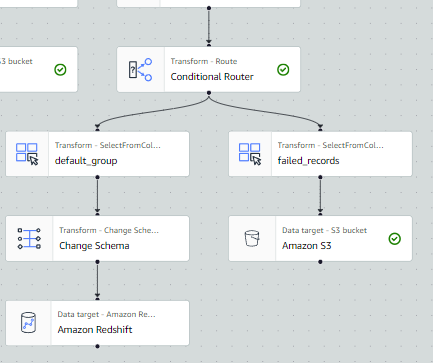


Next we will dump these records to another folder named bad records in S3 Bucket so add it

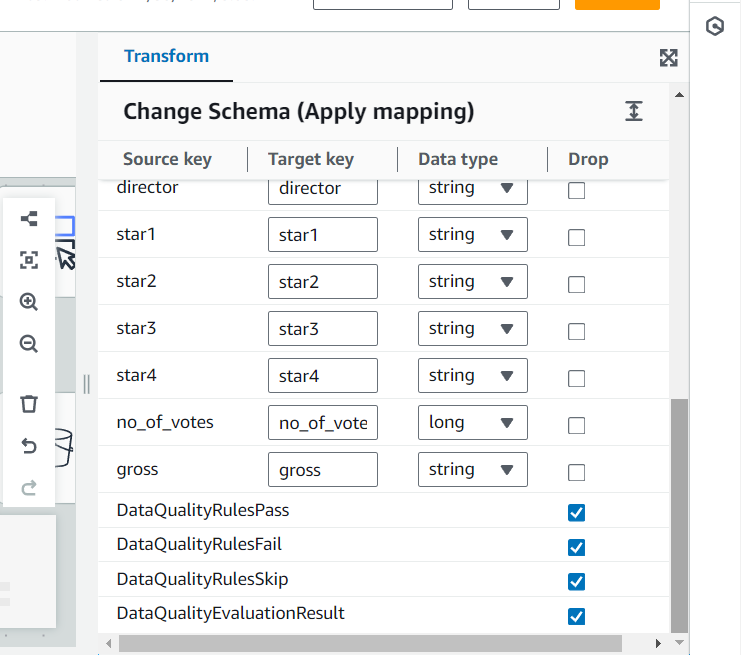




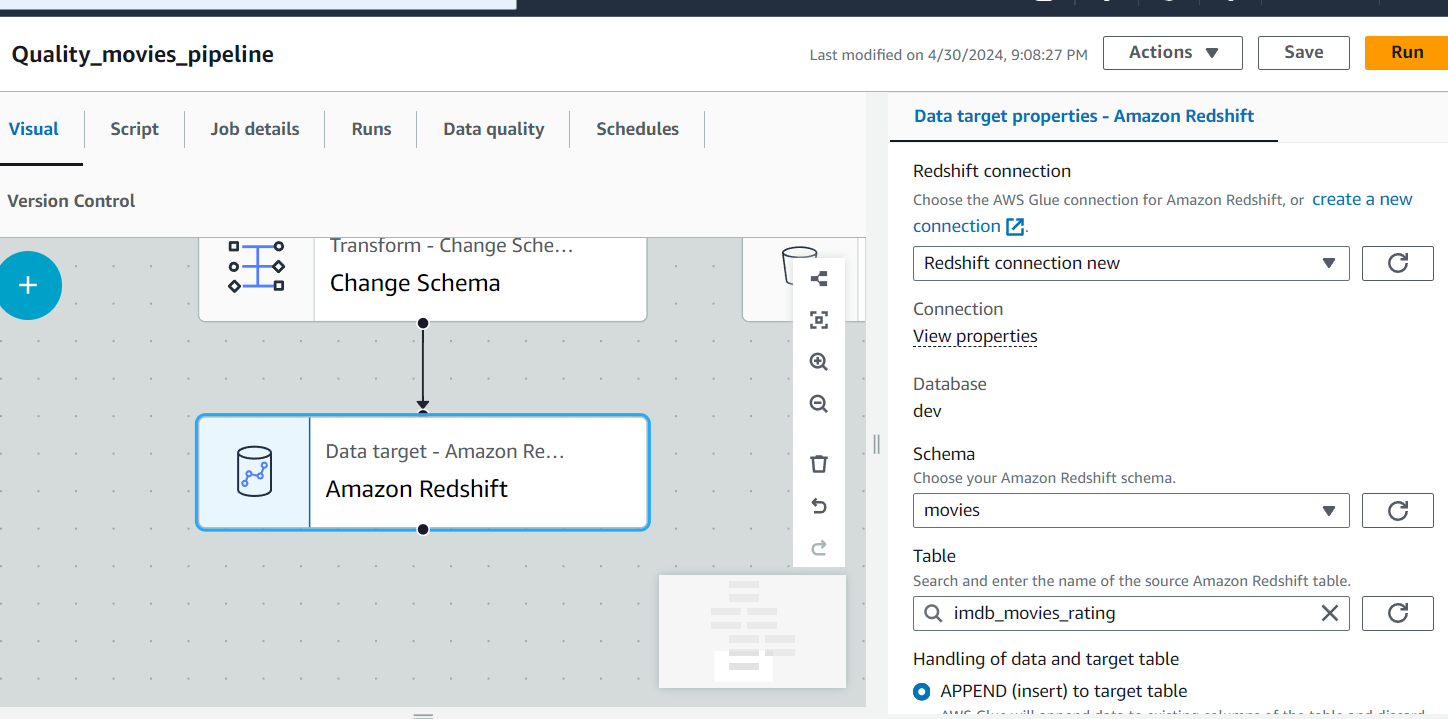
Next from default group we will change schema

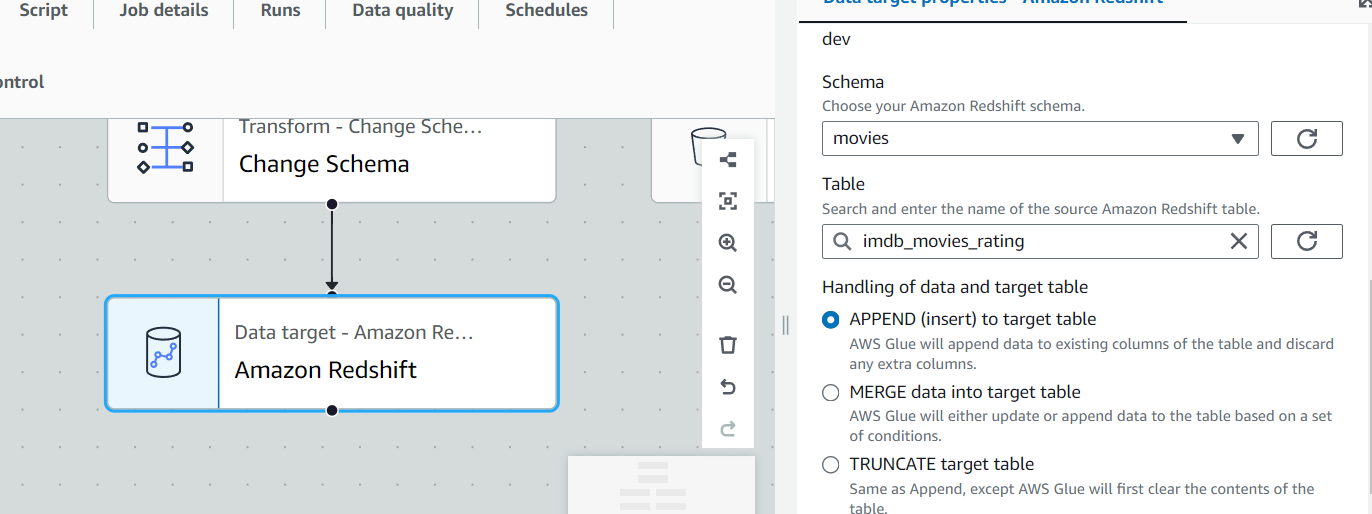


In change schema we will drop unnecessary records and we will change imdb\_rating record data type to decimal and meta\_score and no\_of\_votes –> int as per target table

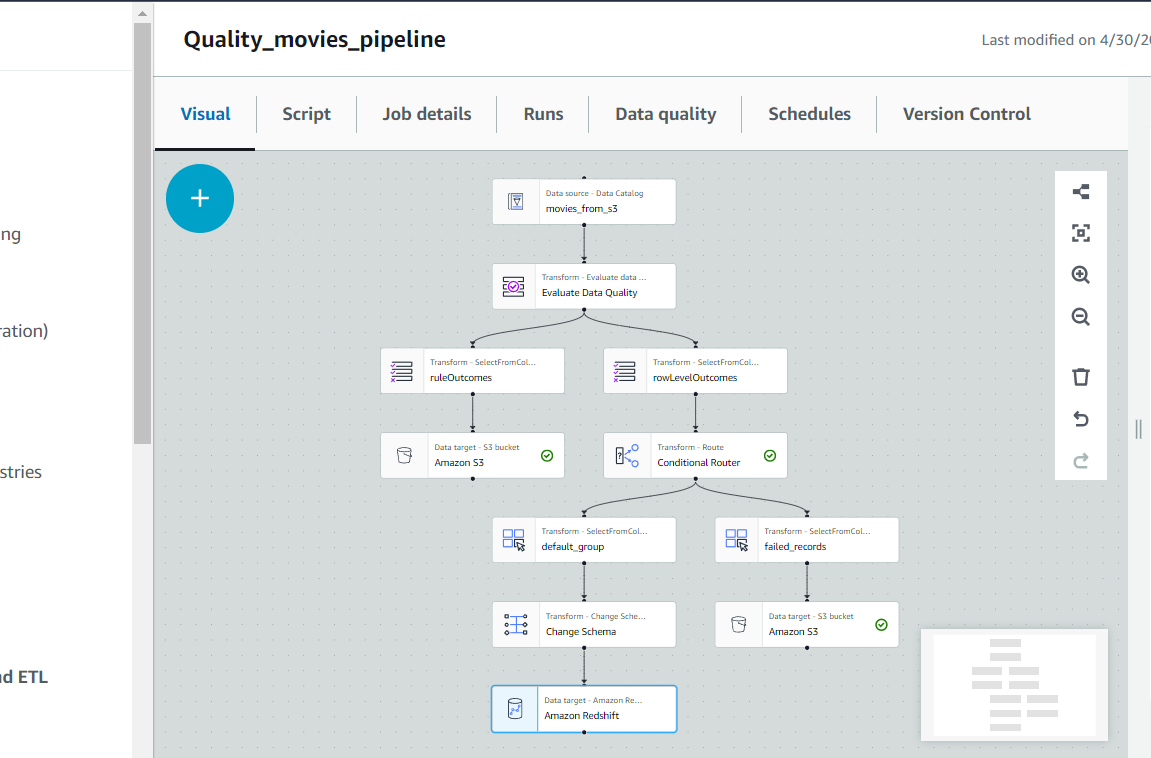


🡪Next we will Add Target i.e Redshift we will choose and add redshift connection , table and schema





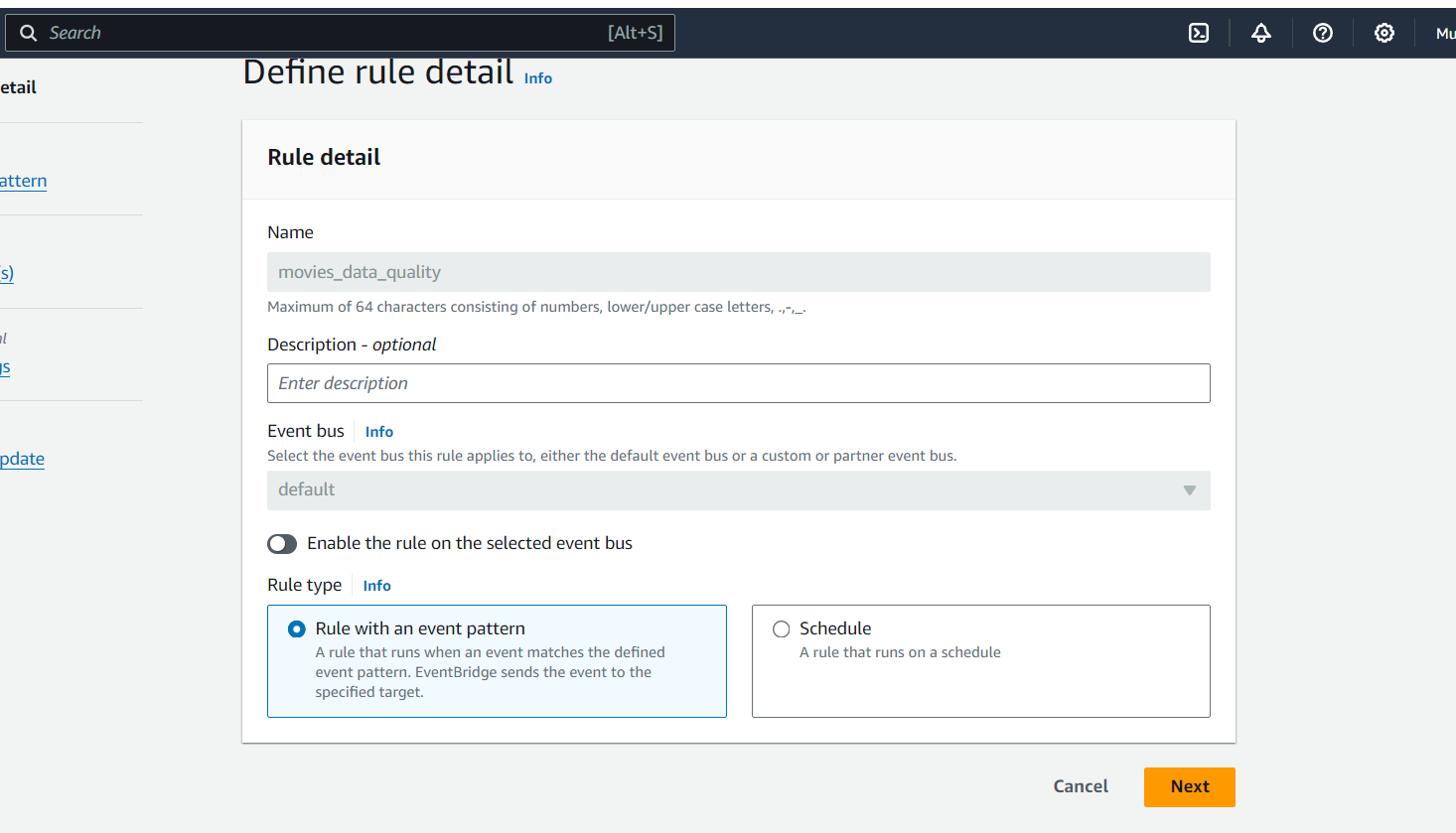
Final Pipeline is Here

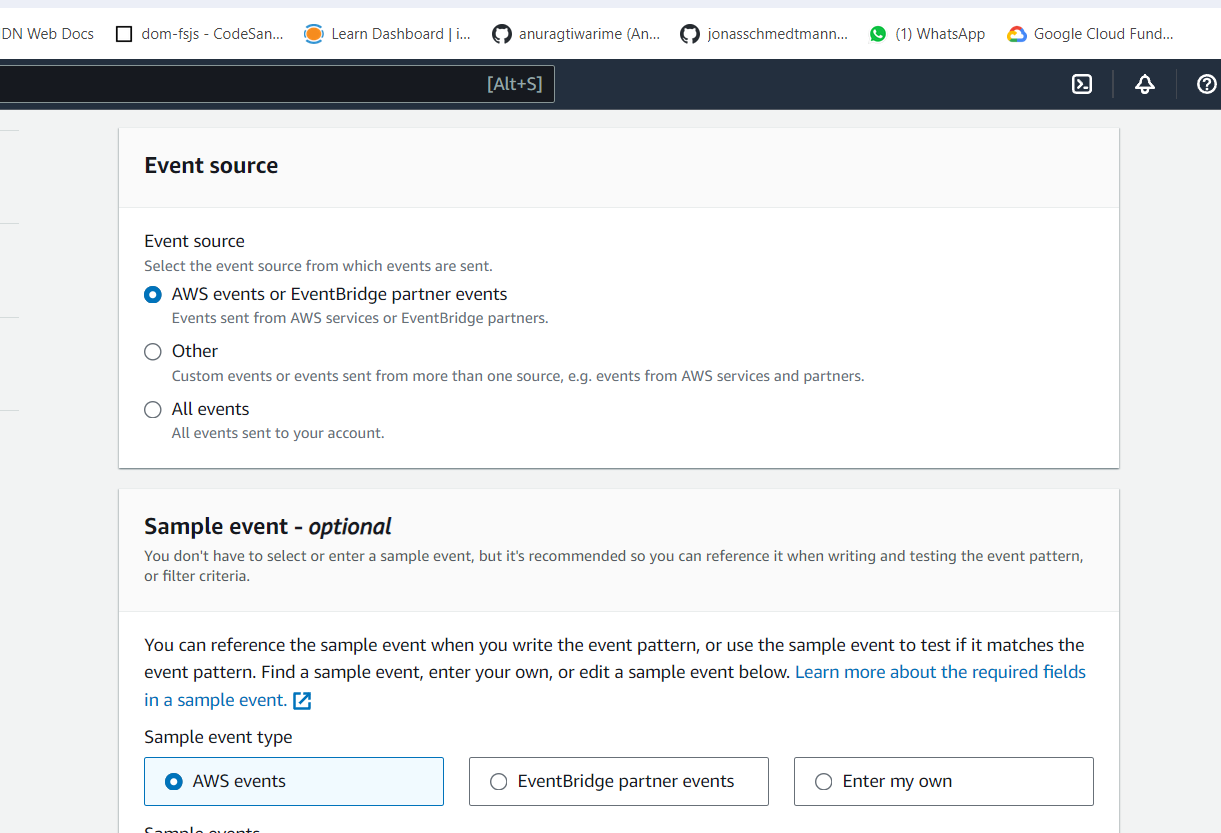


**Next EventBridge Rule**

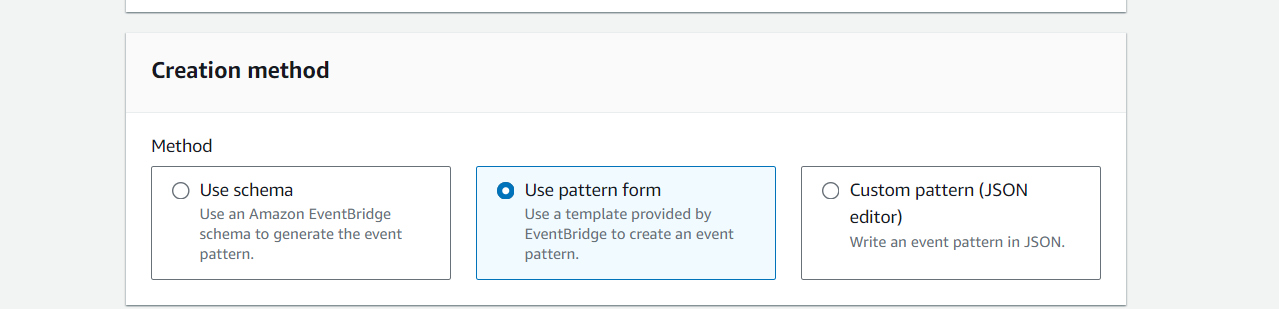
This Rule we are adding for Data Quality Ruleset

Click Create Rule and name Ex: movies\_data\_quality and in Event source select AWS events or Event Bridge partner events

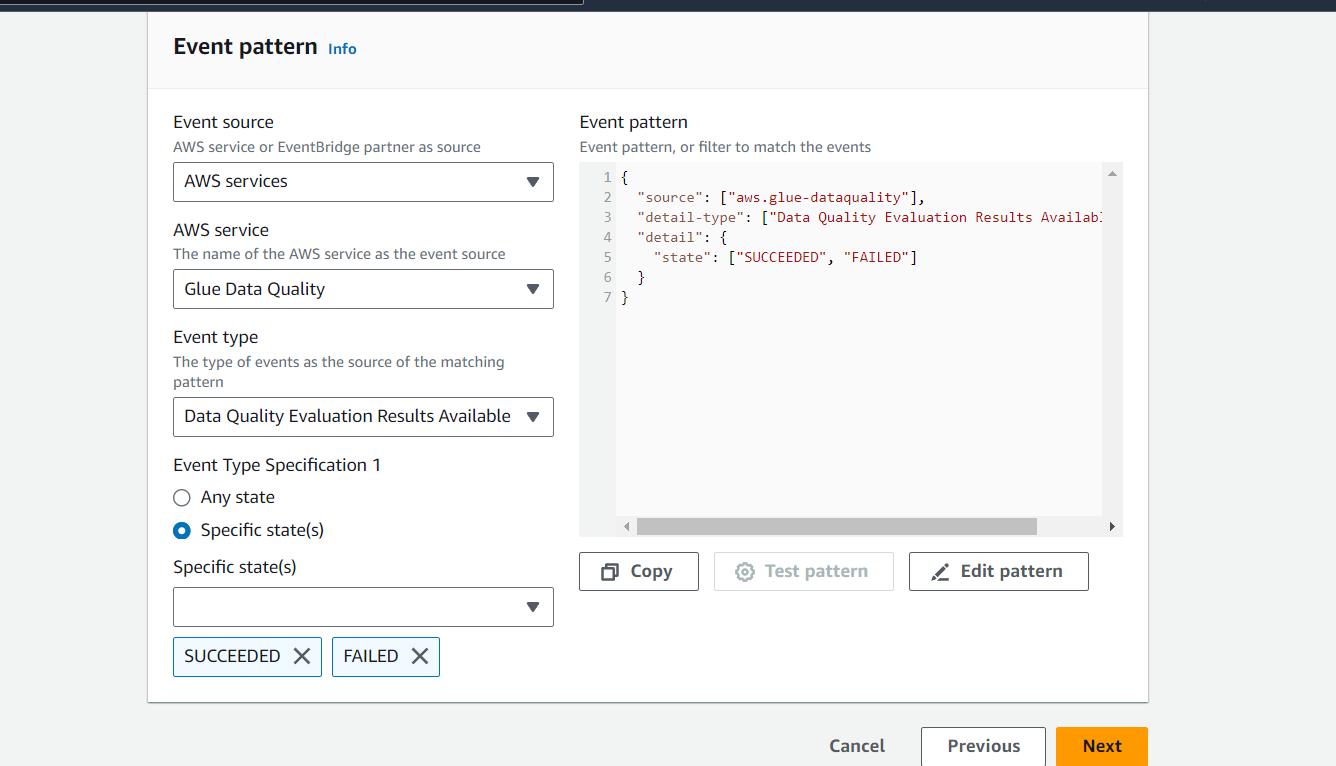
****

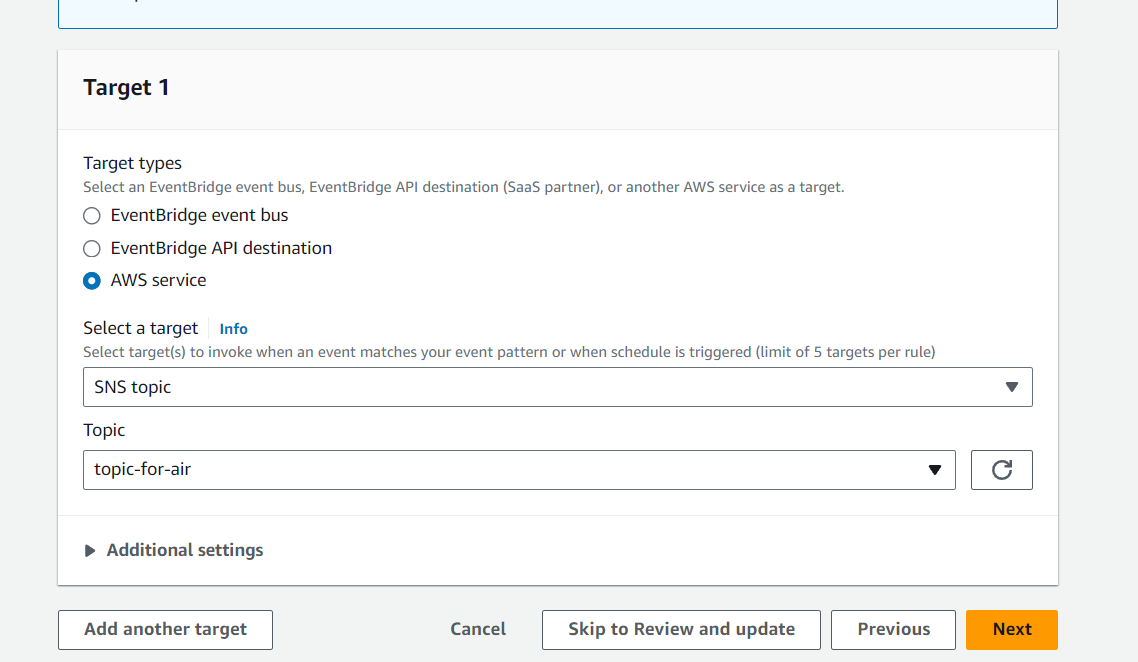


And select pattern form

****

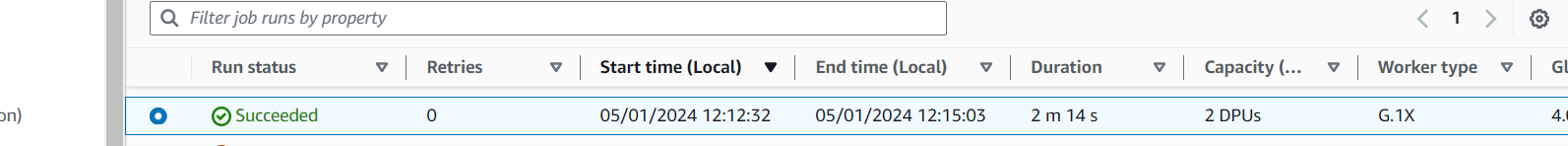
And Event source select AWS services and in that select Glue Data Quality and Add event type as Data quality evaluation results and specific states add Succeeded and failed and for Target add SNS Topic

****

****

NEXT Just manually run etl job so is it succeeding or any error in pipeline

It is succeded



🡪Now we will create materialized view in redshift

create MATERIALIZED VIEW movies.year\_aggregated\_genre AS select released\_year as year ,

genre , count(\*) as total\_movies from movies.imdb\_movies\_rating

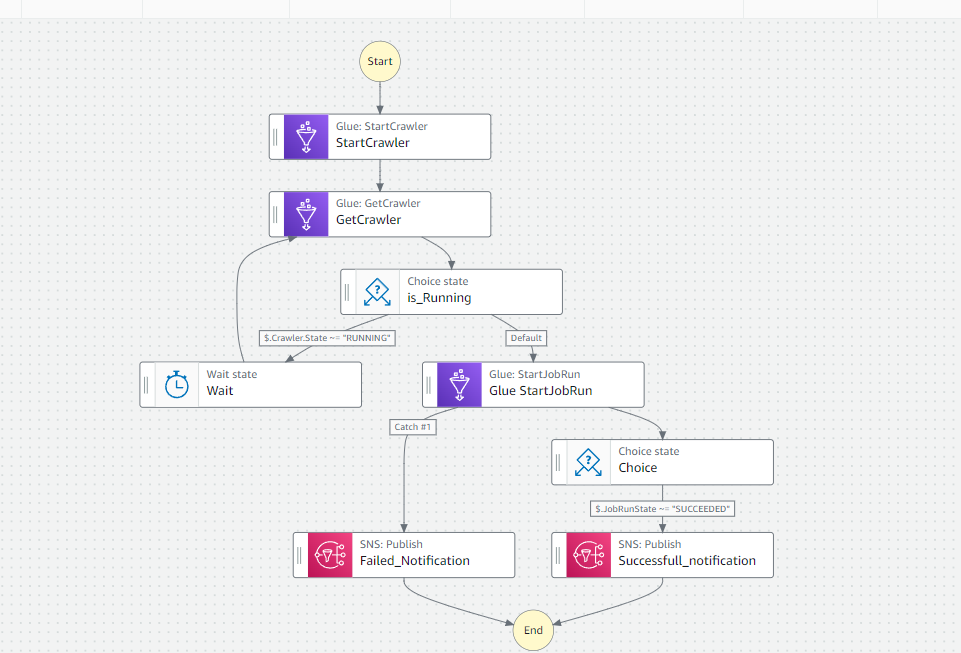
GROUP BY released\_year,genre;

And just execute and see

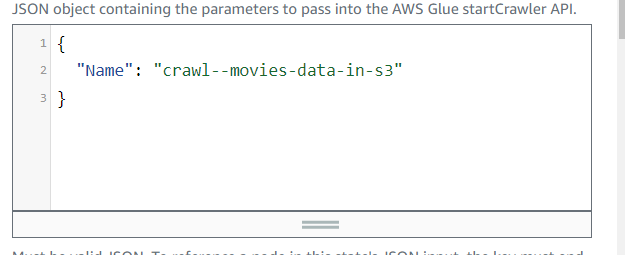
select \* from movies.year\_aggregated\_genre;

**Next Step Functions**

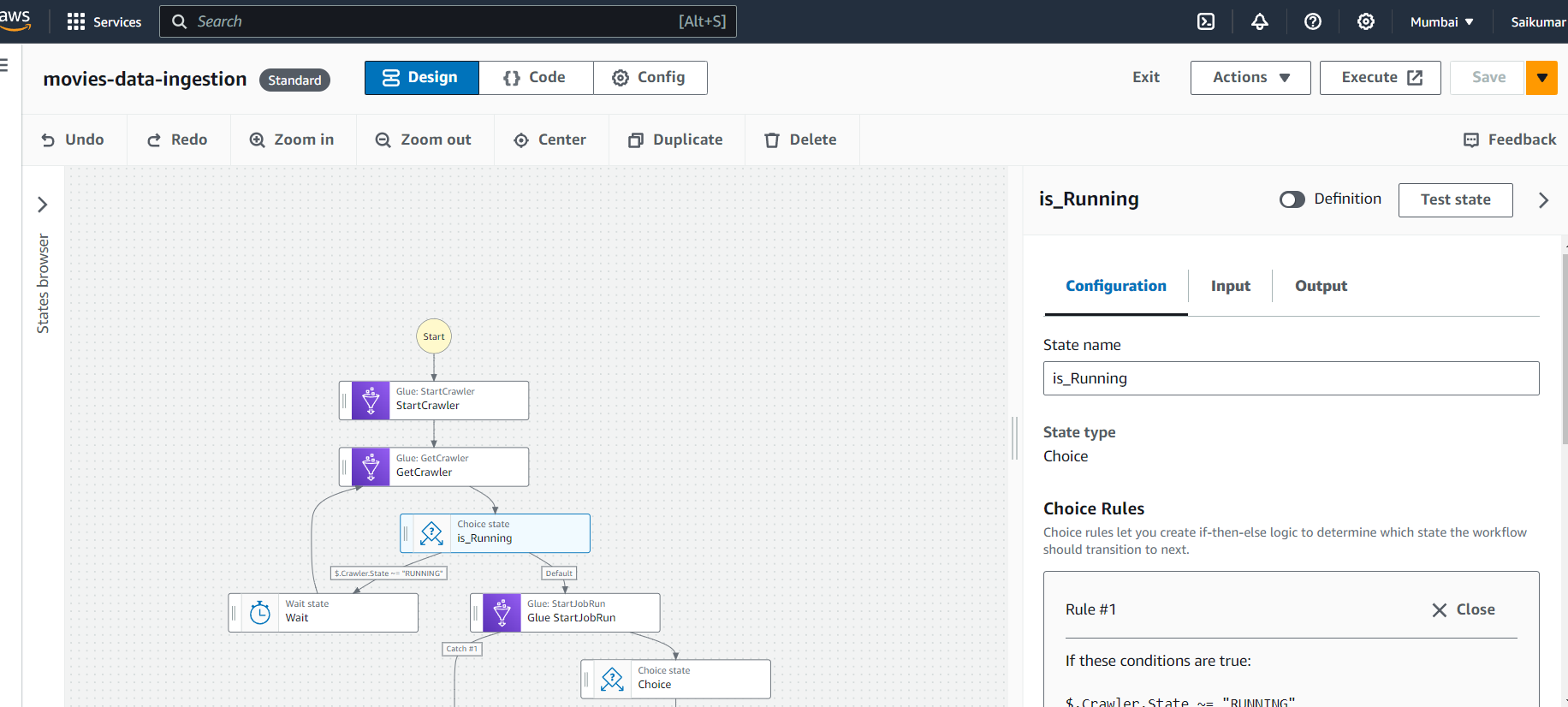
Created Below Step Function

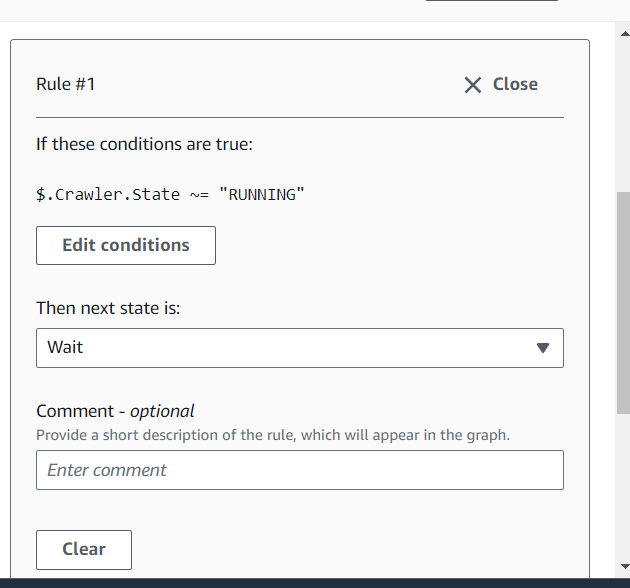
****

First In start crawler add the crawler which is to start I.e Crawler in S3



And same for Get Crawler and then next step is 🡪 choice state added and in that Rule is added

****

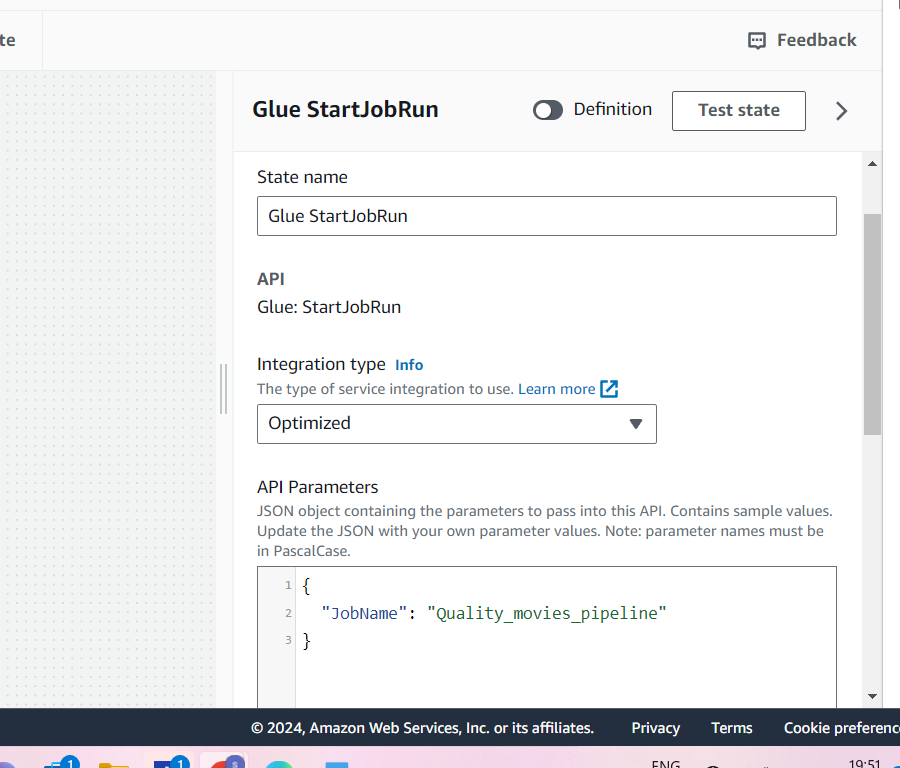
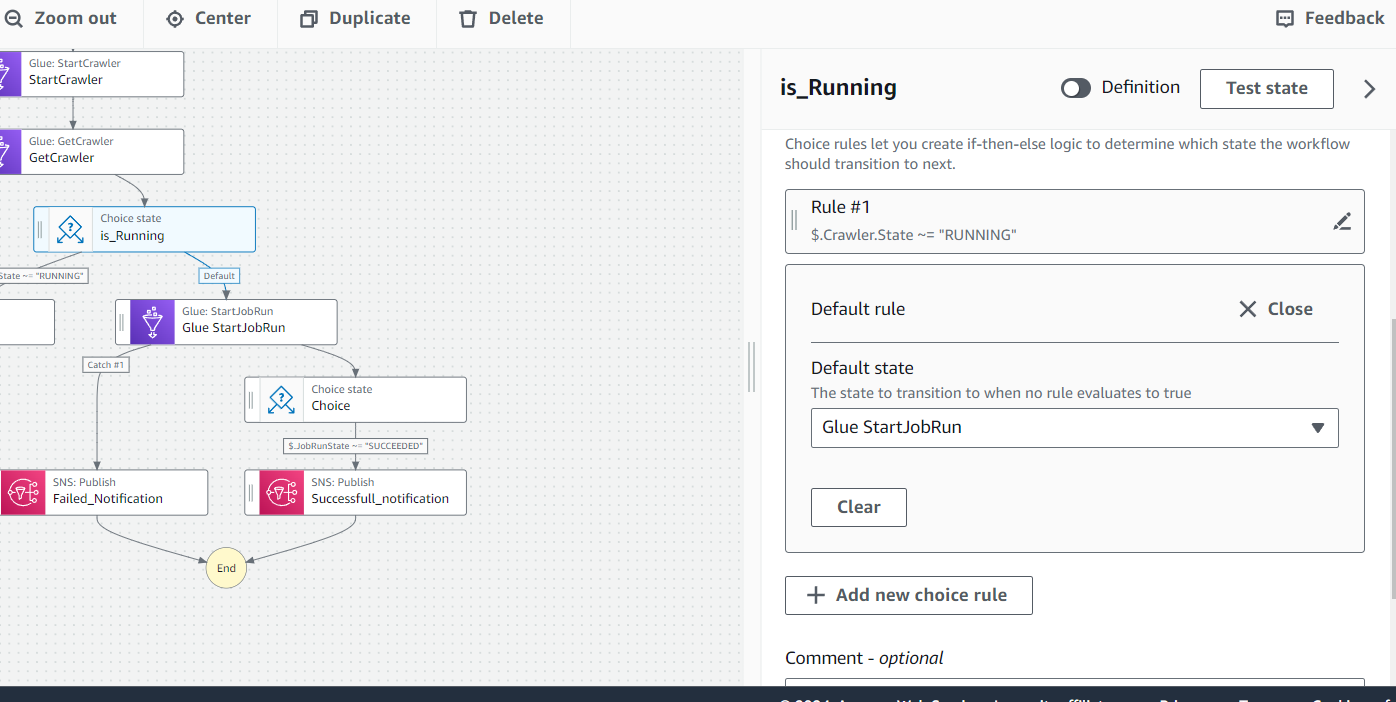
****

🡪Next state added as Wait



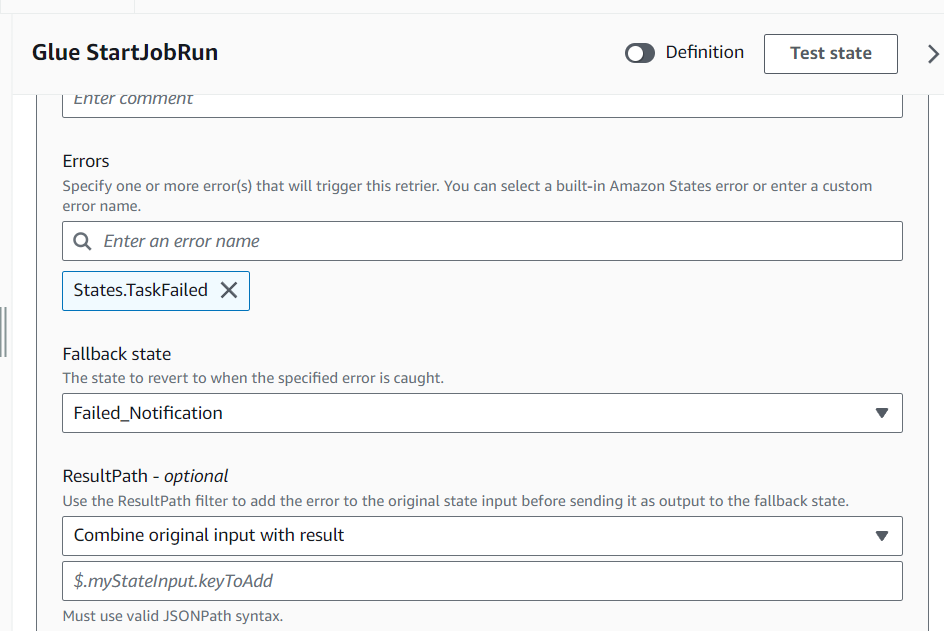
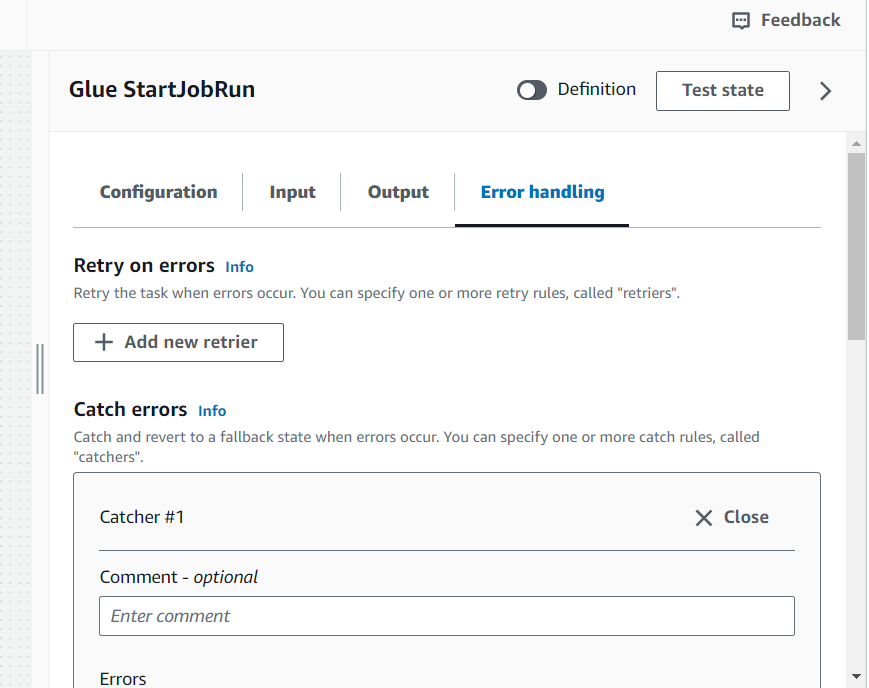
🡪

And for default of choice state the glue StartJobRun is added and here are the details i.e etl job is added

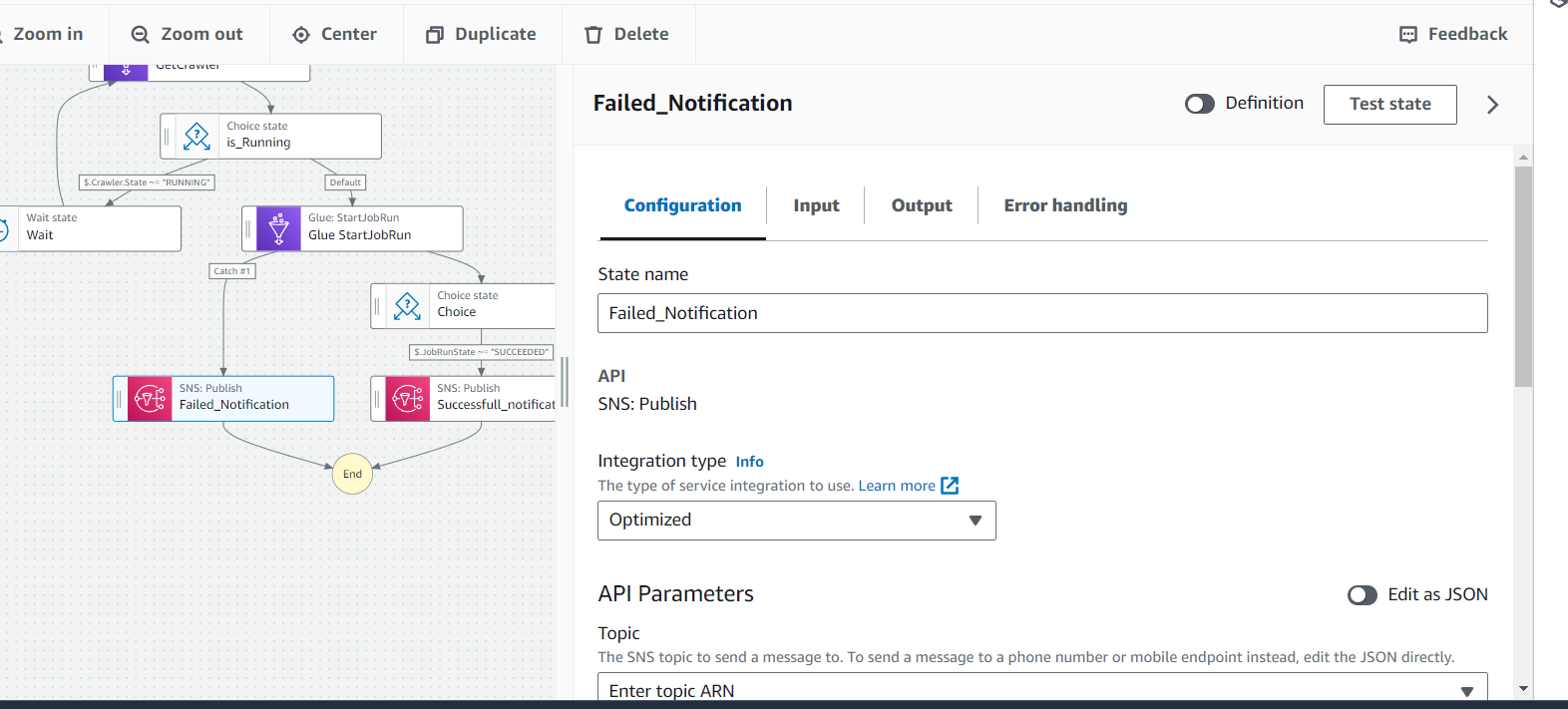


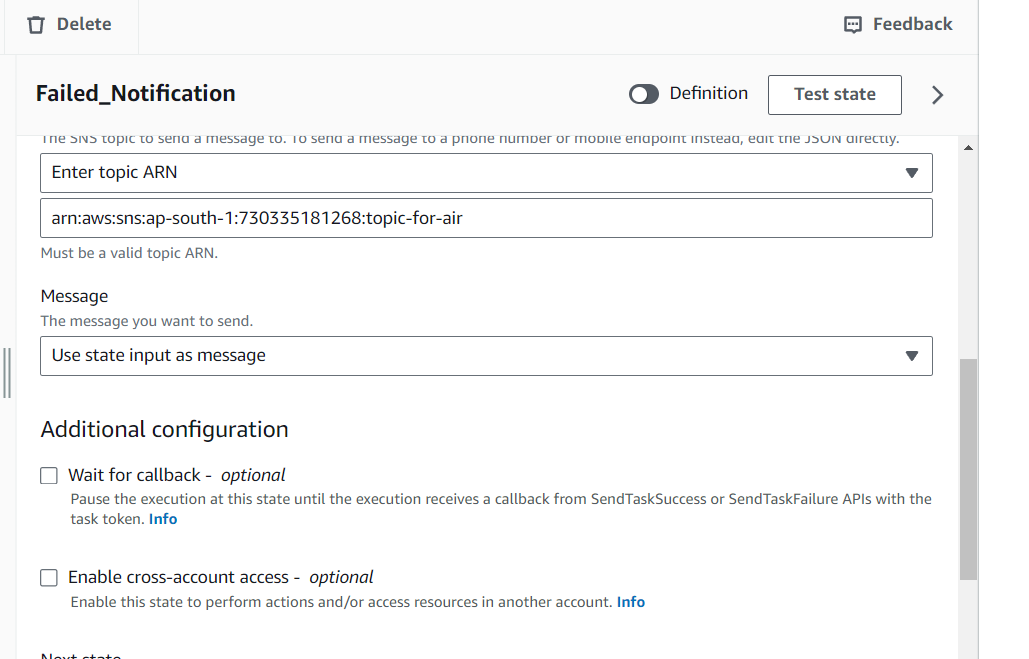
🡪Go to Error Handling and add Catcher

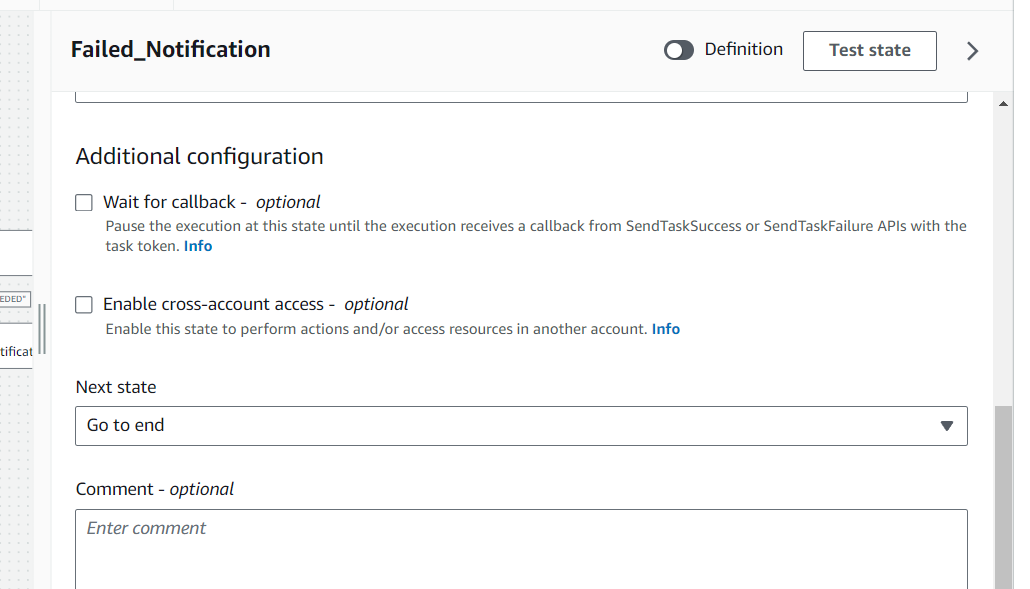
🡪And for this we will add 1 catcher that sends notifications to SNS , we have added as if this glueJobRun is failed it will send notifications to SNS , in that we added SNS topic arn so next step is SNS Publish



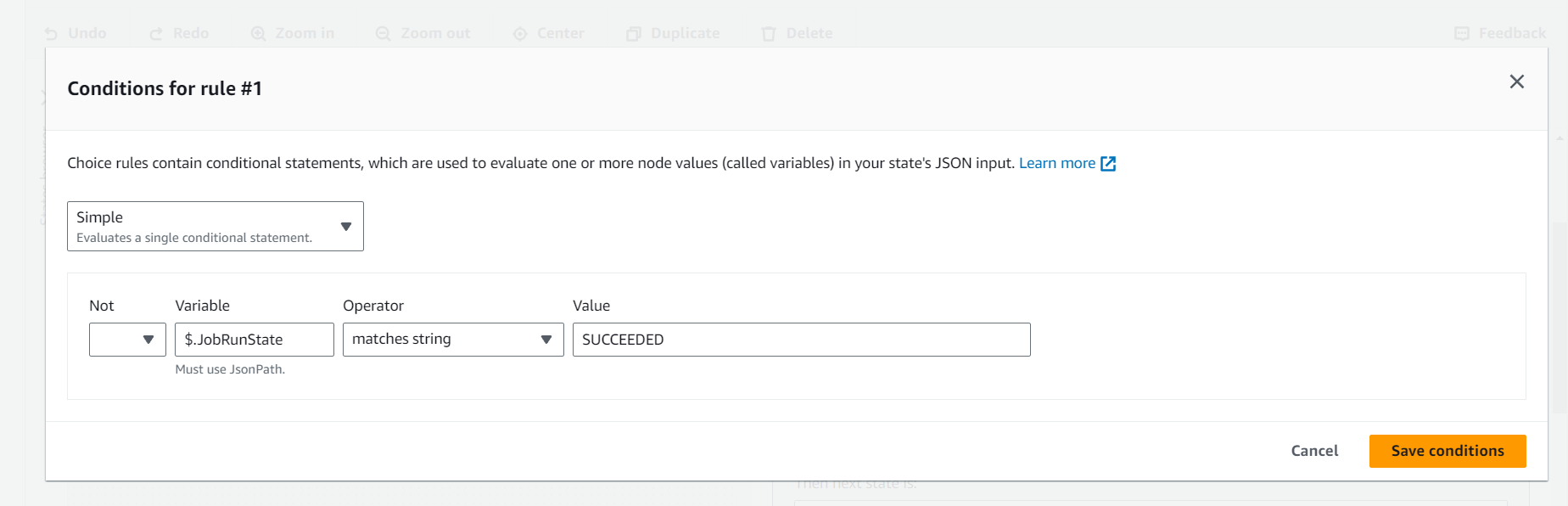
🡪SNS Publish I changed name to Failed\_Notification and Added SNS Topic ARN we can see below and the next state for this is End

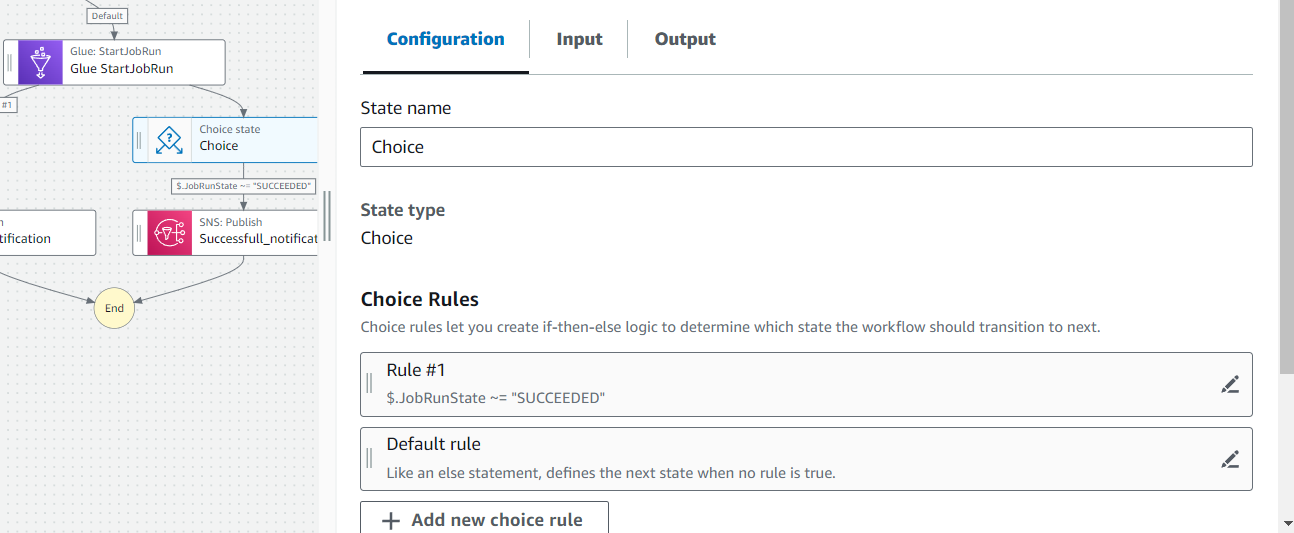






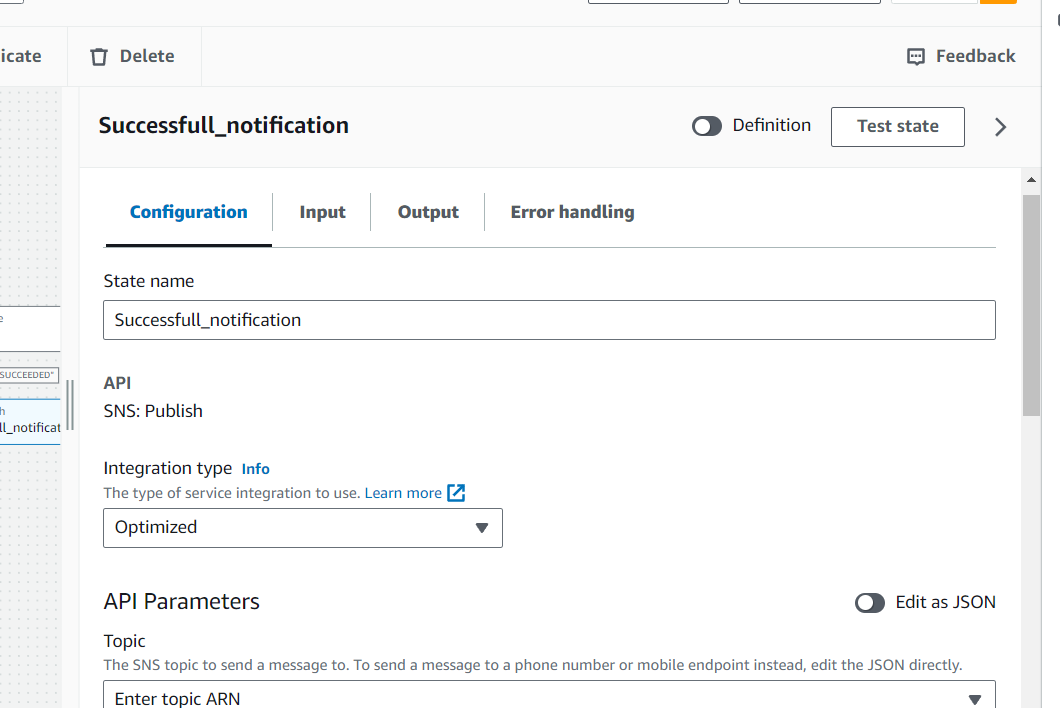
* Now for Glue StartJobRun we added Choice as next state and it that we added 1 Rule i.e and
* Can seen below

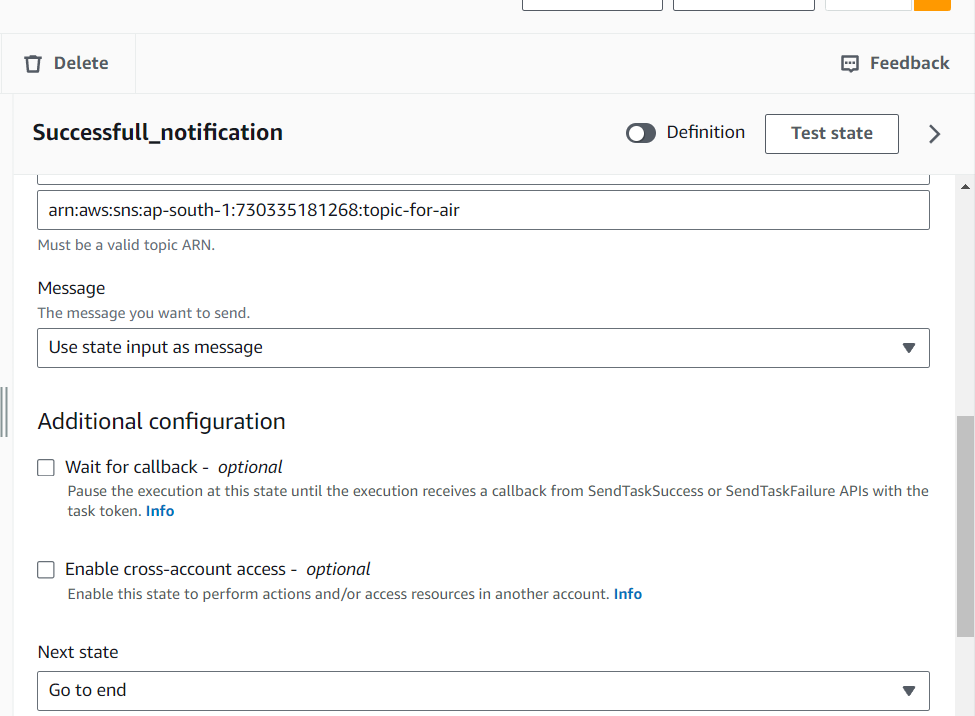




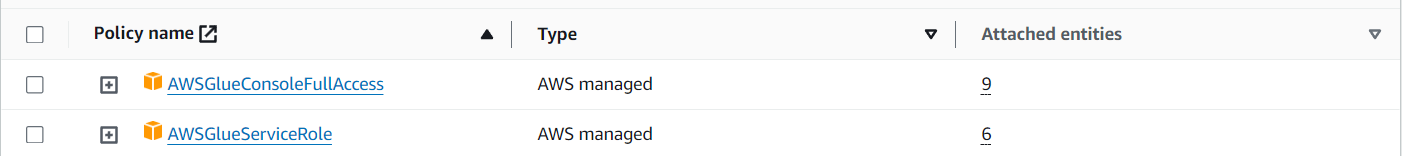
And next state we added to choice is SNS\_PUBLISH

* And sns\_publish name changed as Successful\_notification and added Topic Arn and the Next state as End





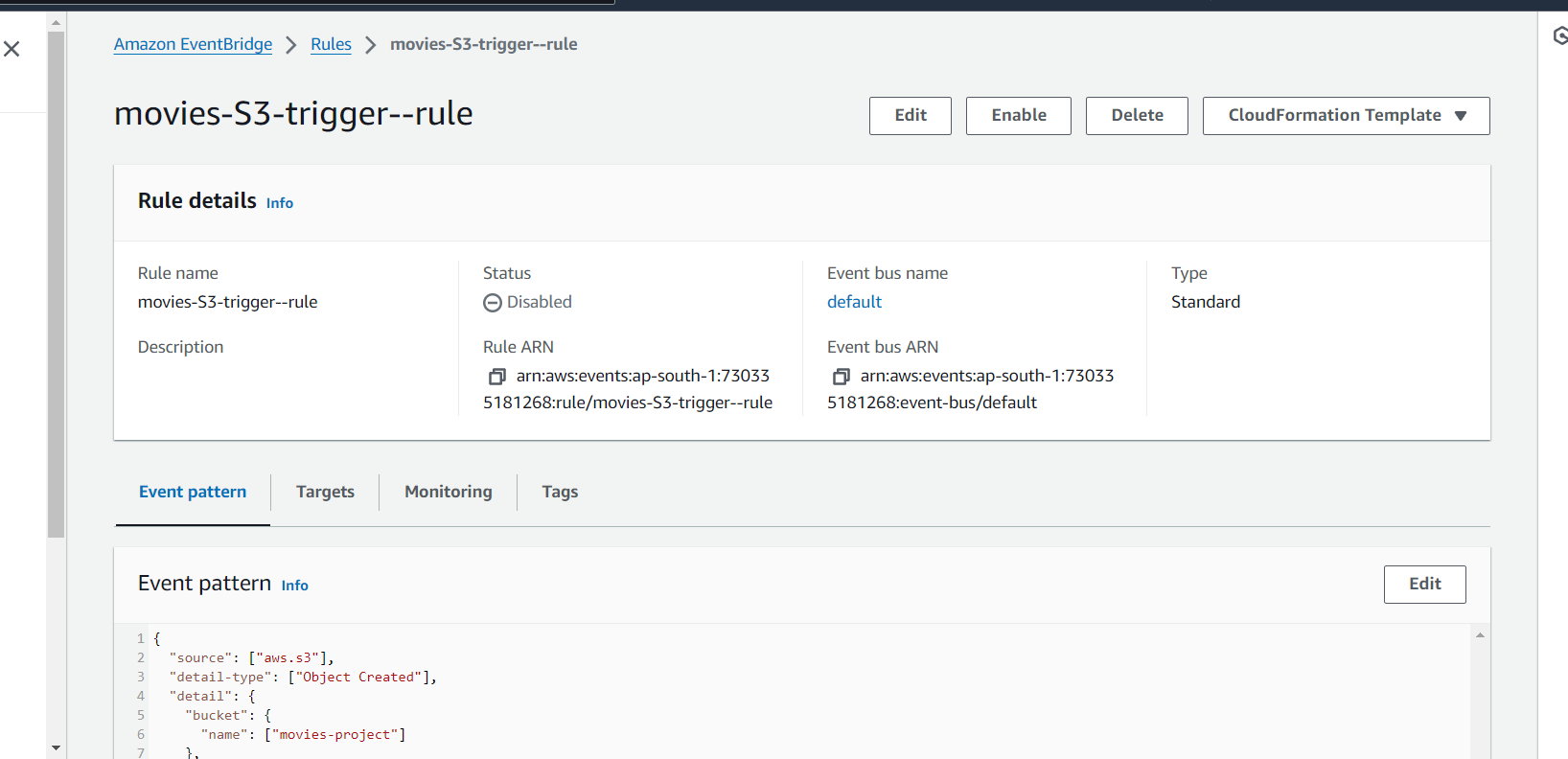
And add These Permissions for IAM Role of StepFunction

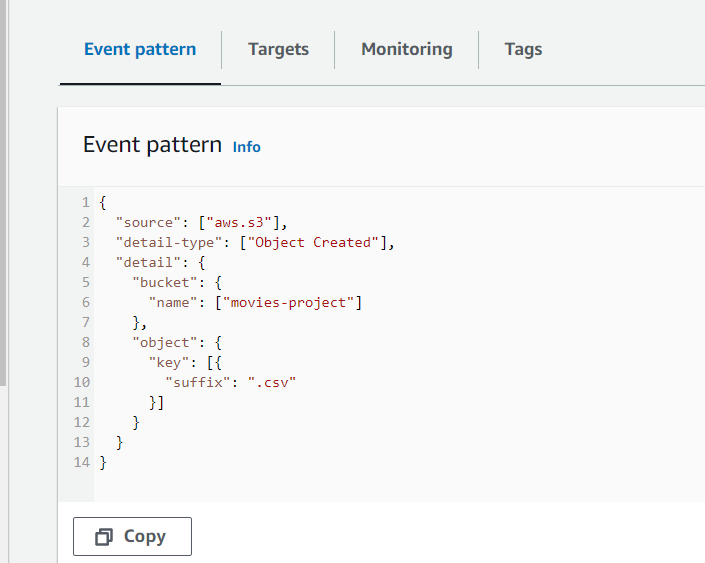


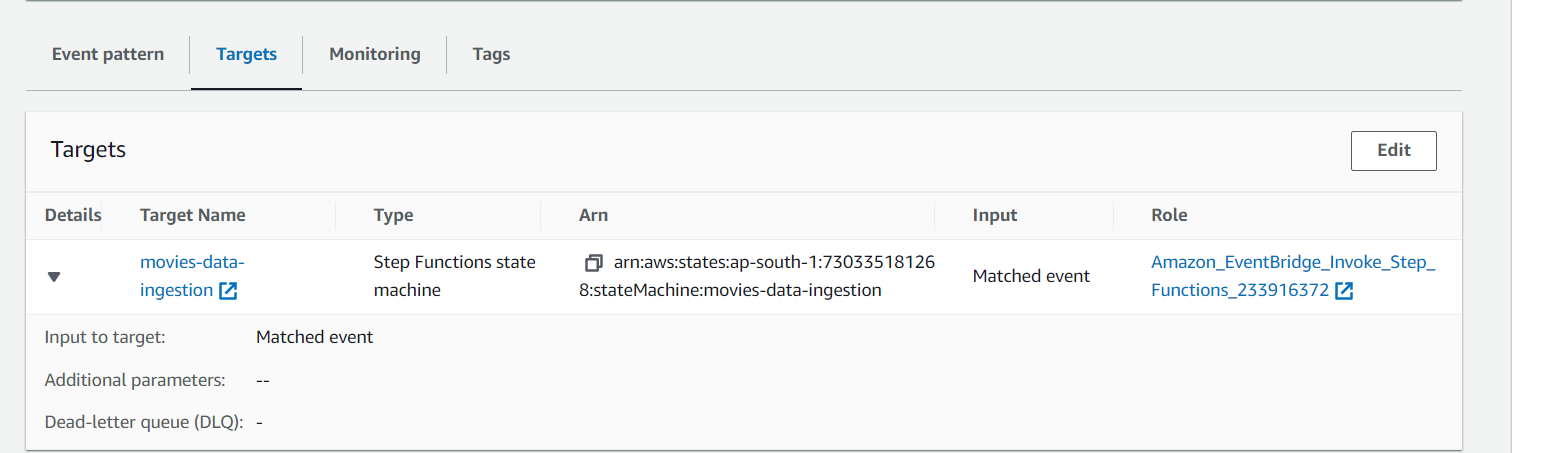
**Next Event Bridge Rule**

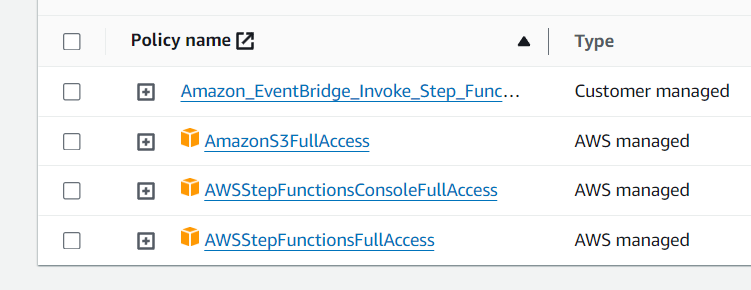
We will create this Rule in this S3 is added as source for Eventbridge rule and StepFuction is added as Target , So whenever a file is uploads in S3 this rule receives event and It Triggers the stepFunction so that pipeline gets started . here we added CSV as suffix so that eventbridgeRule will easily find that file

And add required permissions for the IAM Role





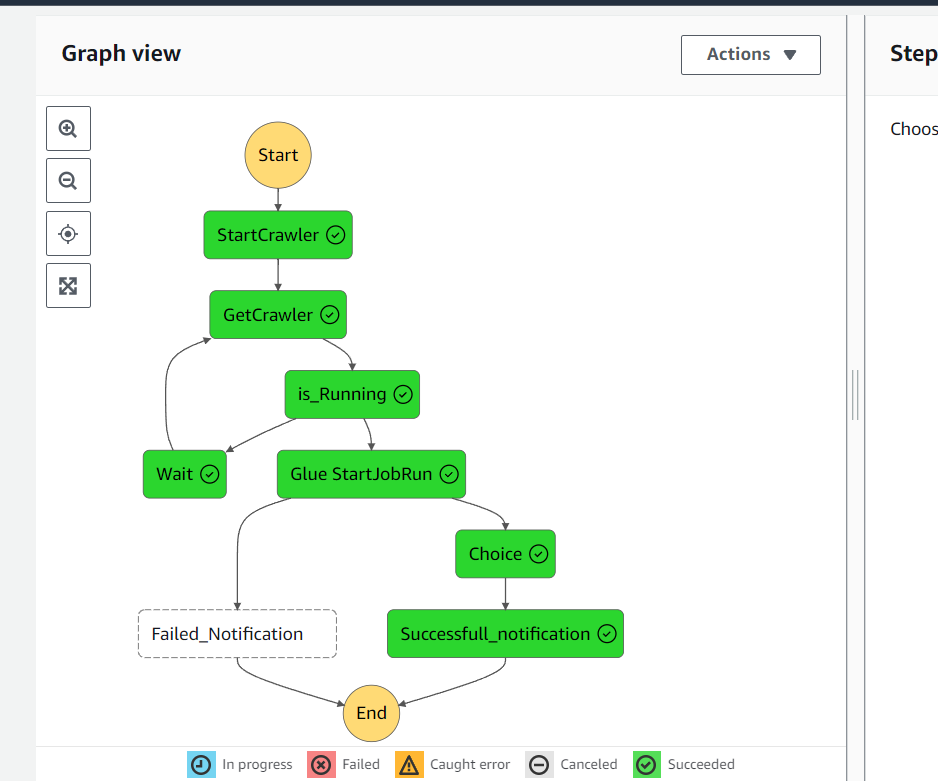


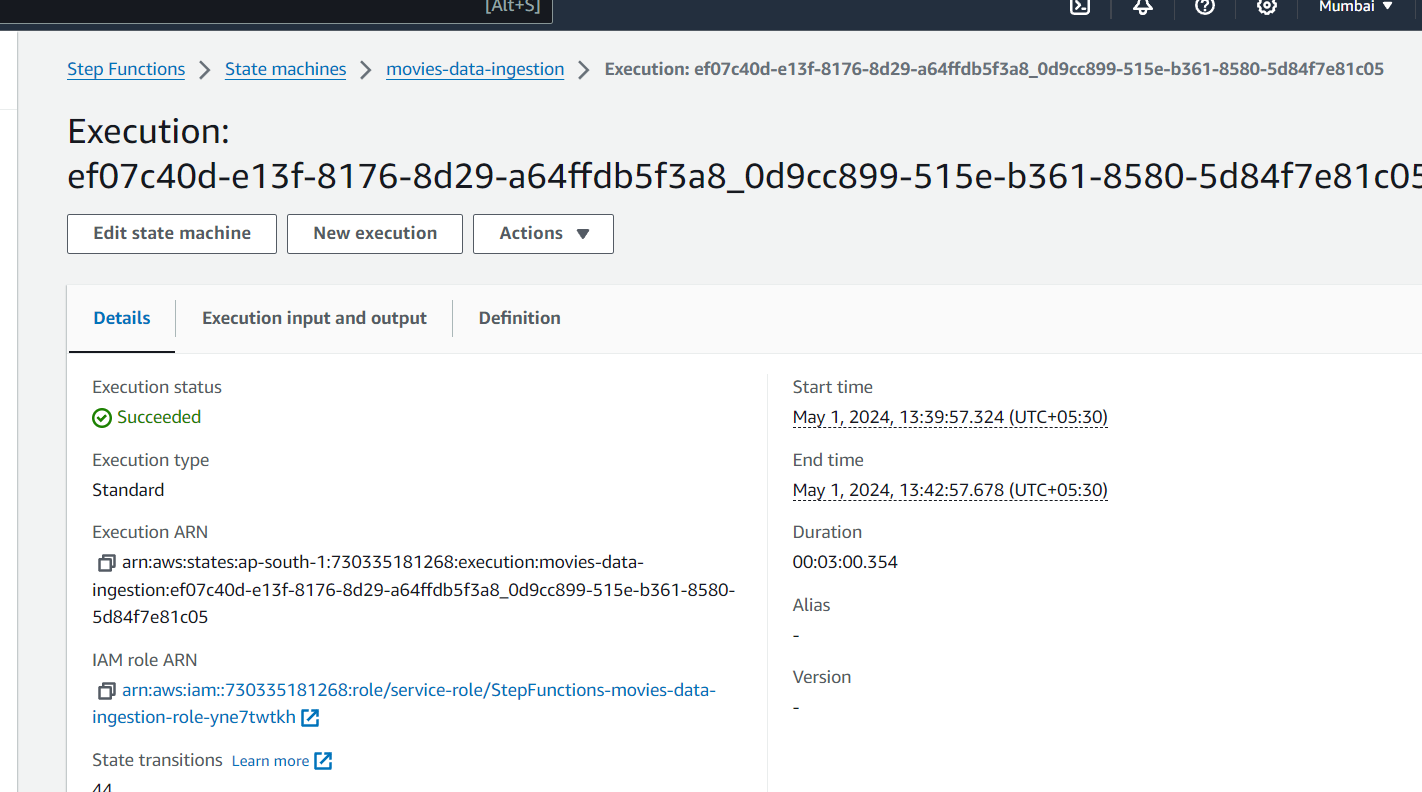


And we will turn ON the EventBridge Notification in S3 bucket so that it can send notifications(events) to EventBridge Rule



🡪Next we will upload Csv file in S3 bucket and as soon as it uploads the event bridge triggers StepFunction and it will start execution and we can see all states will Run and EtlJob also Run and finally Completes.

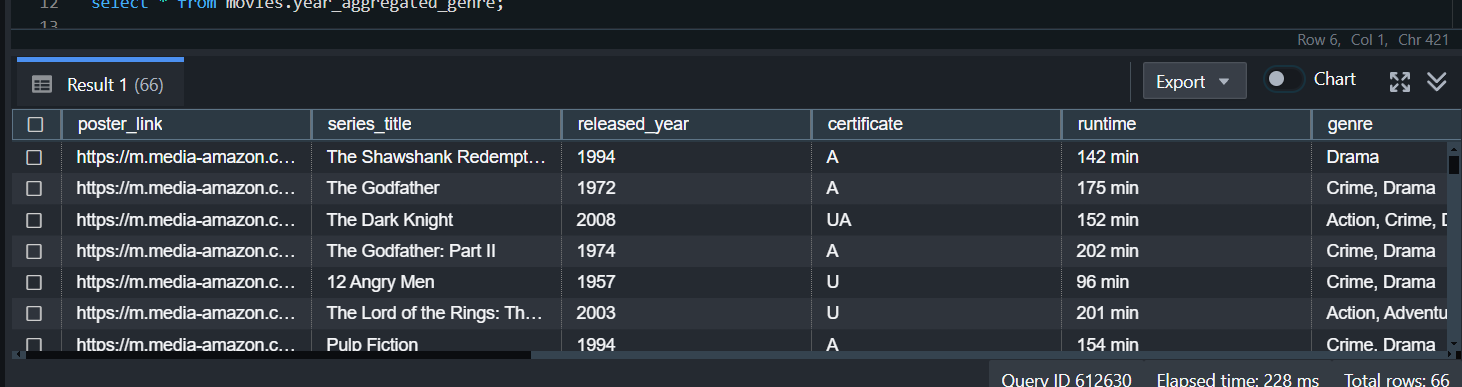




🡪Next in redshift if we execute this

We can see results below

select \* from movies.imdb\_movies\_rating;



🡪-🡪

And we will refresh materialized view so that it will update otherwise it will not update

refresh MATERIALIZED VIEW movies.year\_aggregated\_genre;

🡪and execute this and we can see Results

select \* from movies.year\_aggregated\_genre;

