Assignment11:Movie Quality Data Ingestion

1)We have created relevant folders in s3 bucket with movie data inside input_data folder in s3 bucket

2)created redshift cluster

3)created vpc security groups inbound and outbound rules along with s3 bucket as vpc endpoint

4)created glue database as named airlines_datamart

5)created redshift schema and table for movie data in query editor v2

6)we have made a data quality rule set

7)created visual etl glue job and set the eventbridge rule according to the data quality rule

8)After successfully running glue job we check the table in redshift with the relevant columns

9)We have also received bad records in our folder located in s3 bucket

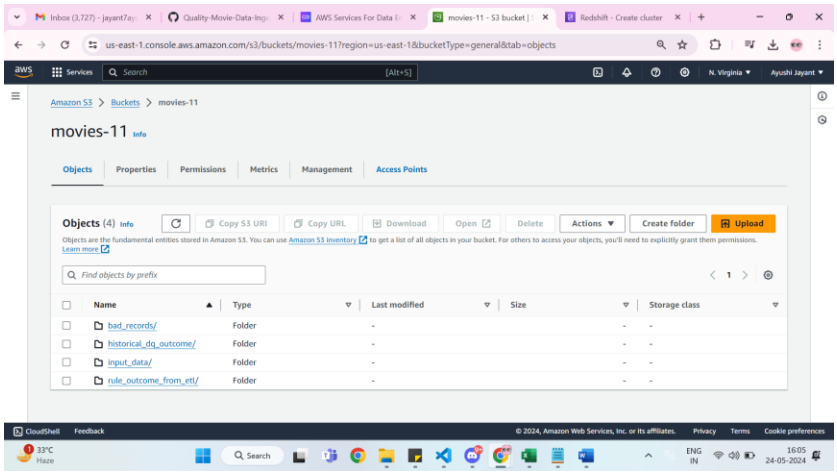
10)created a materialised view from our redshift table and query on basis of that

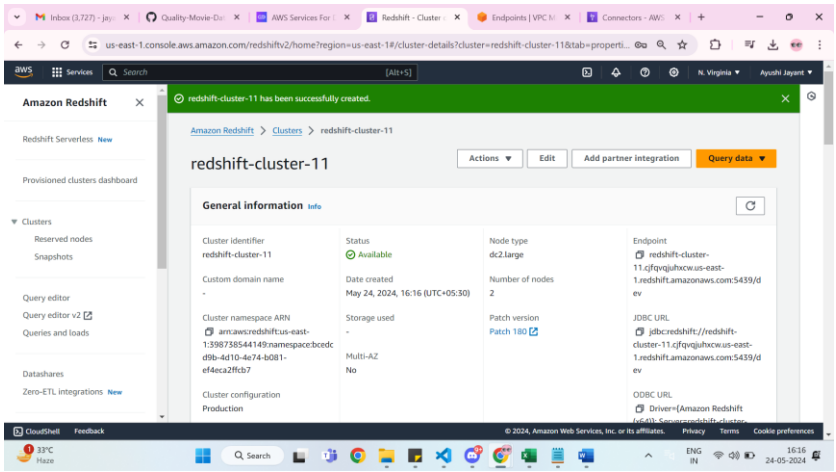
11)created step function to orchestrate our pipeline and attached sns topic with the same for success/failure notifications

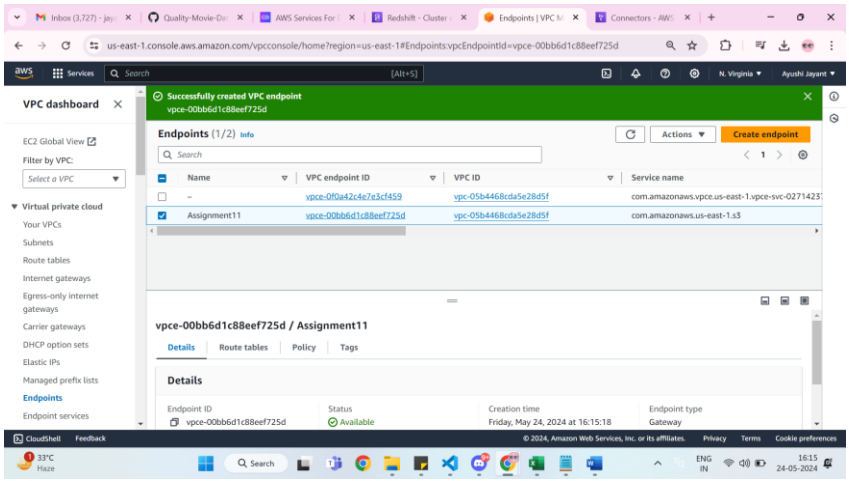
12)We have created an eventbridge rule toinvoke the step function as soon as an object is created in the movie-data bucket

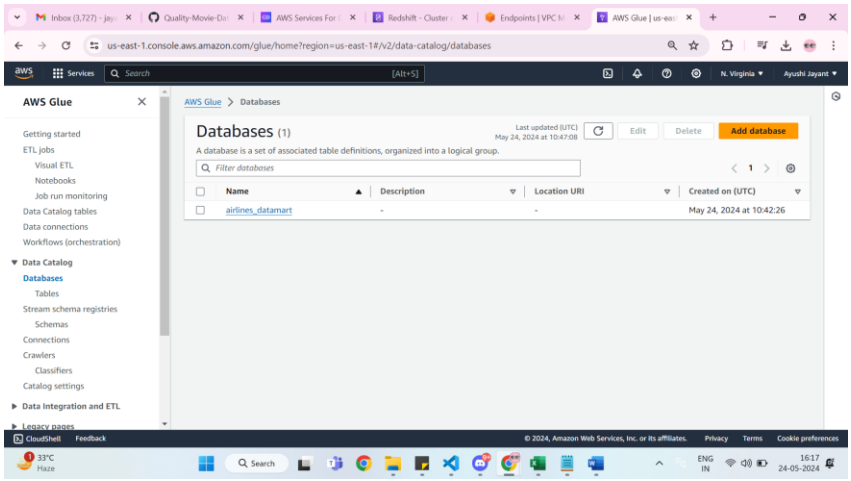
13)step functions ran successfully and we got the notification on mail

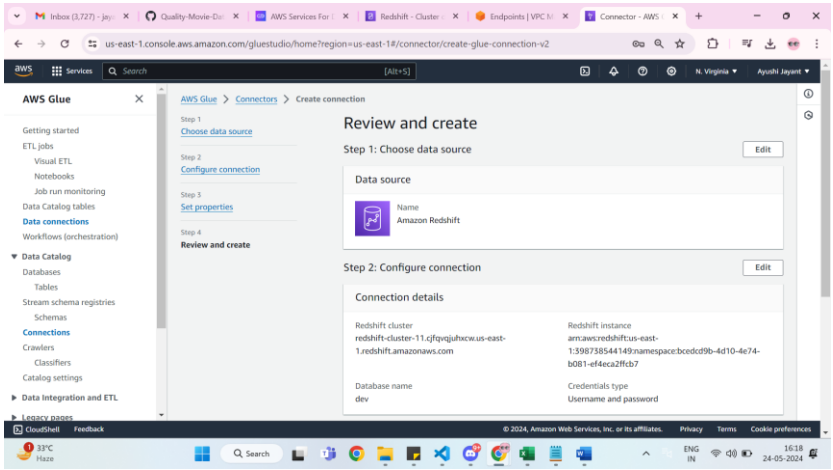
14)After querying in the redshift, we get the total count of records as 99 from 33 because we have the refreshed data I the materialised view


movies-11 - S3 bucket
us-east-1.console.aws.amazon.com/s3/buckets/movies-11?region=us-east-1&bucketType=general&tab=objects
Amazon S3 > Buckets > movies-11
movies-11
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (4)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
bad_records/
historical_dq_outcome/
input_data/
rule_outcome_from_etl/
Folder
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Redshift - Cluster
us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#/cluster-details?cluster=redshift-cluster-11&tab=properti...
Amazon Redshift
redshift-cluster-11 has been successfully created.
Amazon Redshift > Clusters > redshift-cluster-11
redshift-cluster-11
Actions
Edit
Add partner integration
Query data
General information
Cluster identifier
redshift-cluster-11
Status
Available
Node type
dc2.large
Endpoint
redshift-cluster-11.cjfqvqjuhxcw.us-east-1.redshift.amazonaws.com:5439/dev
Custom domain name
Date created
May 24, 2024, 16:16 (UTC+05:30)
Number of nodes
2
Cluster namespace ARN
Storage used
Patch version
Patch 180
JDBC URL
Multi-AZ
No
Cluster configuration
Production
ODBC URL
Clusters
Reserved nodes
Snapshots
Query editor
Query editor v2
Queries and loads
Datashares
Zero-ETL integrations
Redshift Serverless
Provisioned clusters dashboard


Endpoints | VPC M
us-east-1.console.aws.amazon.com/vpcconsole/home?region=us-east-1#Endpoints:vpcEndpointId=vpce-00bb6d1c88eef725d
VPC dashboard
Successfully created VPC endpoint
vpce-00bb6d1c88eef725d
Endpoints (1/2)
Actions
Create endpoint
Search
Name
VPC endpoint ID
VPC ID
Service name
vpce-0f0a42c4e7e3cf459
vpc-05b4468cda5e28d5f
com.amazonaws.vpce.us-east-1.vpce-svc-0271423
Assignment11
vpce-00bb6d1c88eef725d
vpc-05b4468cda5e28d5f
com.amazonaws.us-east-1.s3
vpce-00bb6d1c88eef725d / Assignment11
Details
Route tables
Policy
Tags
Endpoint ID
Status
Available
Creation time
Friday, May 24, 2024 at 16:15:18
Endpoint type
Gateway
EC2 Global View
Filter by VPC:
Select a VPC
Virtual private cloud
Your VPCs
Subnets
Route tables
Internet gateways
Egress-only internet gateways
Carrier gateways
DHCP option sets
Elastic IPs
Managed prefix lists
Endpoints
Endpoint services

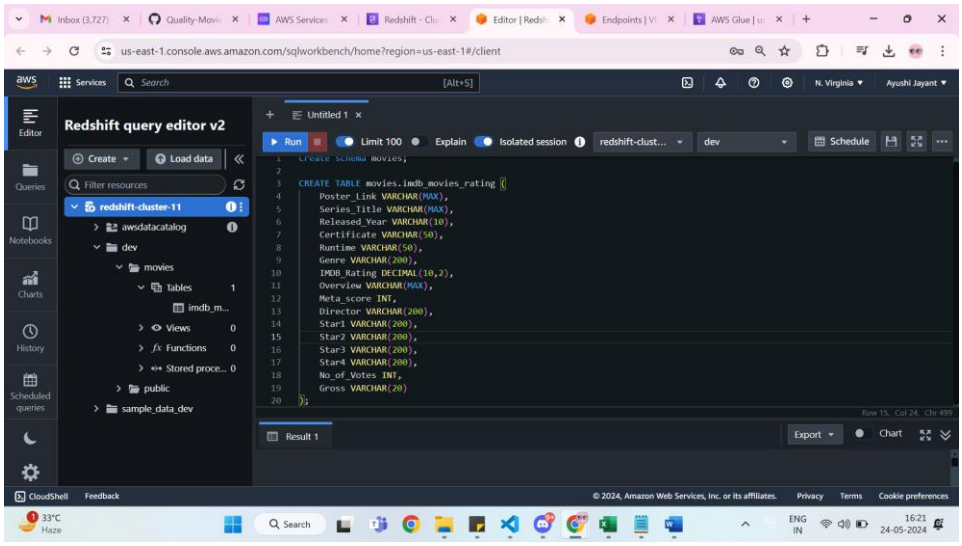


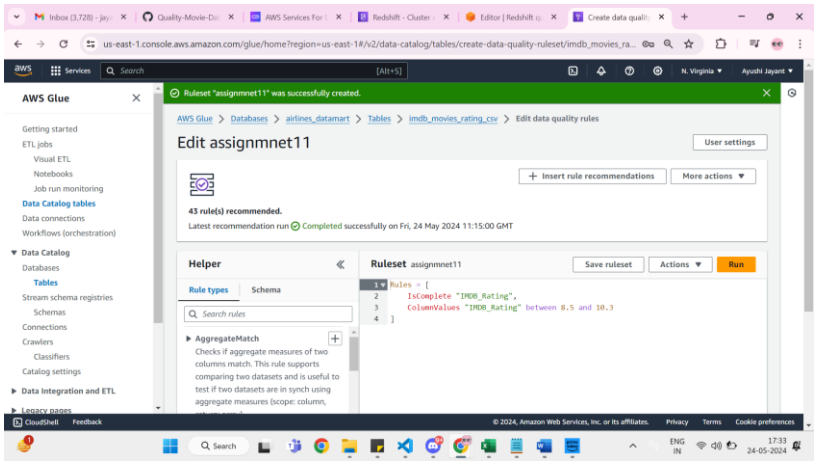


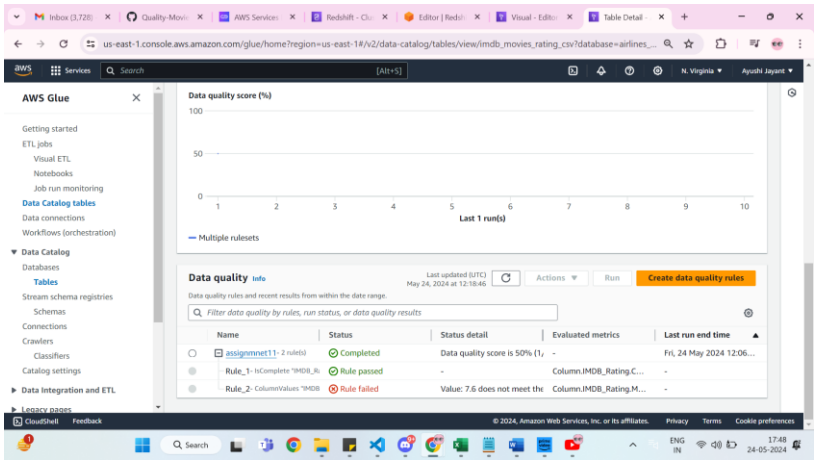
Create schema movies;

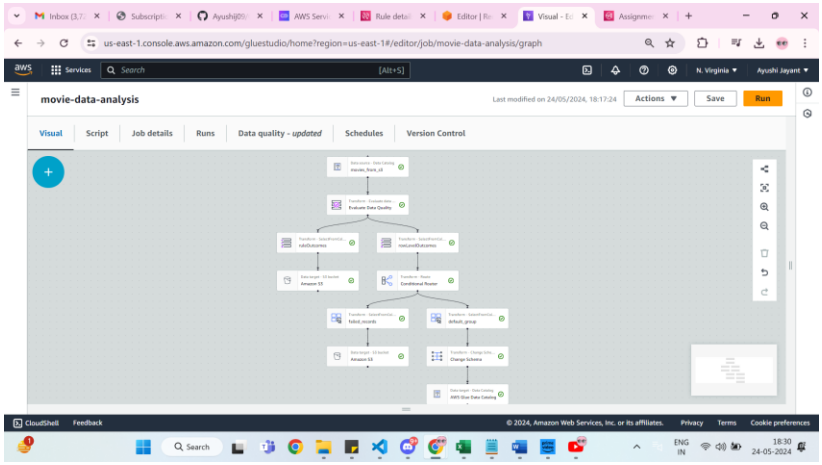
```
CREATE TABLE movies.imdb_movies_rating (
    Poster_Link VARCHAR(MAX),
    Series_Title VARCHAR(MAX),
    Released_Year VARCHAR(10),
    Certificate VARCHAR(50),
    Runtime VARCHAR(50),
    Genre VARCHAR(200),
    IMDB_Rating DECIMAL(10,2),
    Overview VARCHAR(MAX),
    Meta_score INT,
```

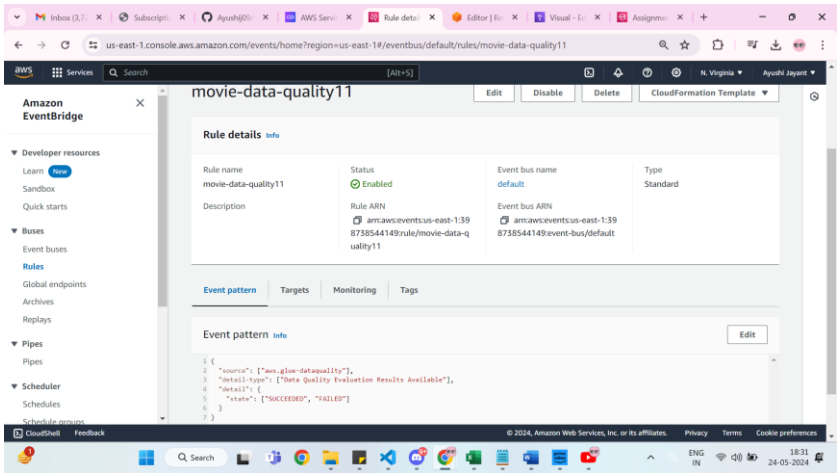
```
    Director VARCHAR(200),
    Star1 VARCHAR(200),
    Star2 VARCHAR(200),
    Star3 VARCHAR(200),
    Star4 VARCHAR(200),
    No_of_Votes INT,
    Gross VARCHAR(20)
);
```

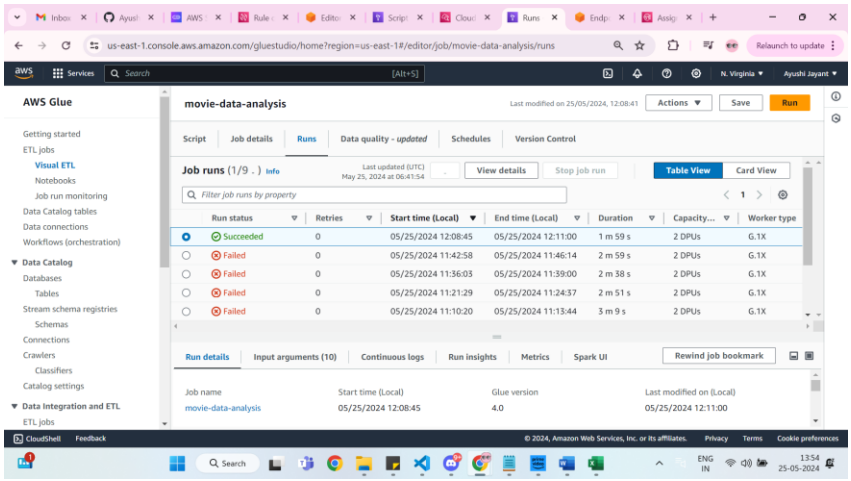


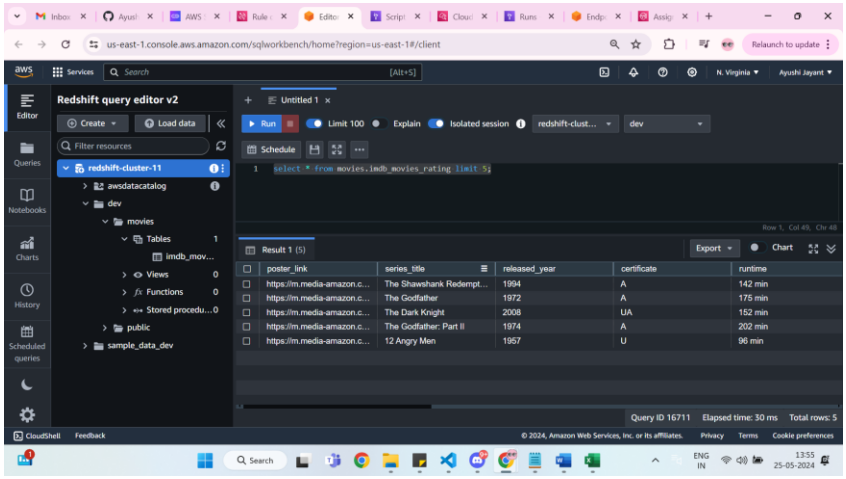


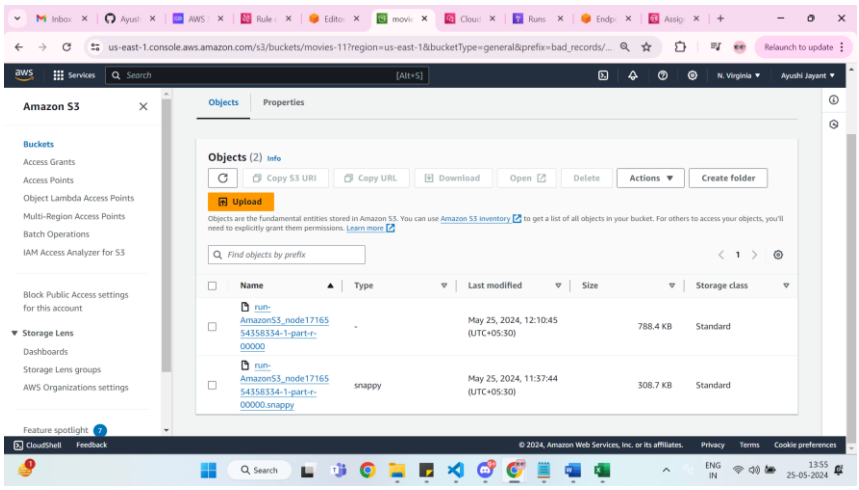

AWS Glue
Data quality score (%)
Last 1 run(s)
Multiple rulesets
Data quality
Create data quality rules
Name
Status
Status detail
Evaluated metrics
Last run end time
Completed
Rule passed
Rule failed

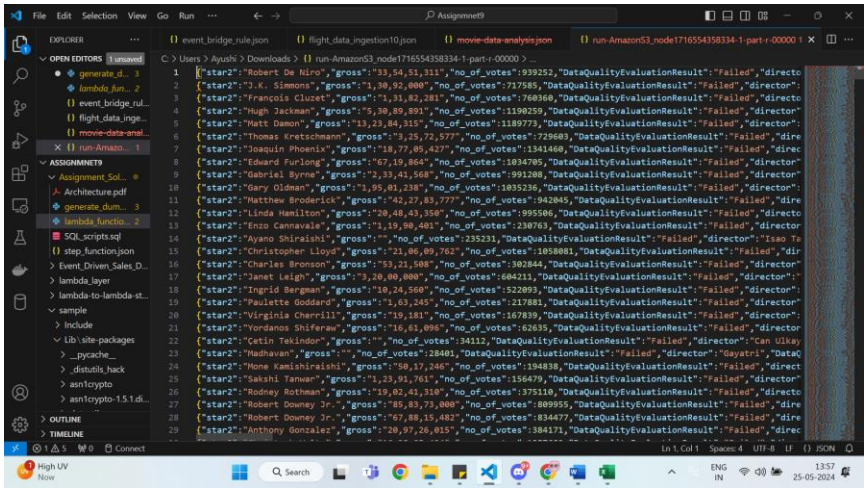

movie-data-analysis
Actions
Save
Run
Visual
Script
Job details
Runs
Data quality - updated
Schedules
Version Control


Amazon EventBridge
movie-data-quality11
Edit
Disable
Delete
CloudFormation Template
Rule details
Rule name
movie-data-quality11
Status
Enabled
Event bus name
default
Type
Standard
Description
Rule ARN
Event bus ARN
Event pattern
Targets
Monitoring
Tags
Edit

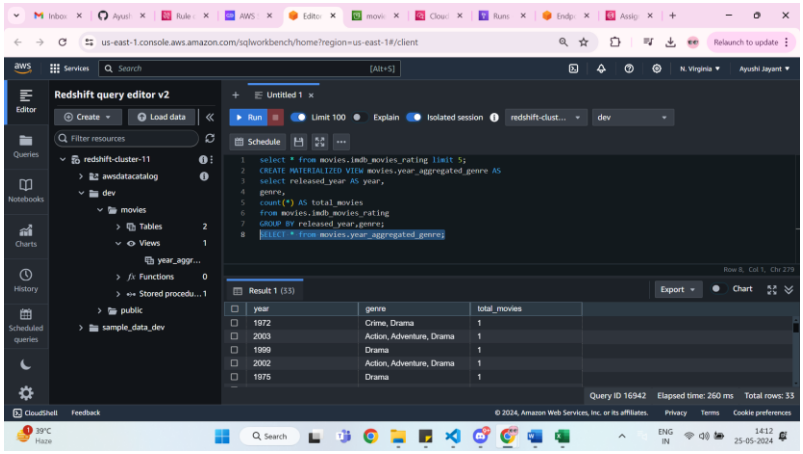

us-east-1.console.aws.amazon.com/gluestudio/home?region=us-east-1#/editor/job/movie-data-analysis/runs
AWS Glue
movie-data-analysis
Job runs (1/9 .) Info
Run status
Retries
Start time (Local)
End time (Local)
Duration
Capacity...
Worker type
Succeeded 0 05/25/2024 12:08:45 05/25/2024 12:11:00 1 m 59 s 2 DPUs G.1X
Failed 0 05/25/2024 11:42:58 05/25/2024 11:46:14 2 m 59 s 2 DPUs G.1X
Failed 0 05/25/2024 11:36:03 05/25/2024 11:39:00 2 m 38 s 2 DPUs G.1X
Failed 0 05/25/2024 11:21:29 05/25/2024 11:24:37 2 m 51 s 2 DPUs G.1X
Failed 0 05/25/2024 11:10:20 05/25/2024 11:13:44 3 m 9 s 2 DPUs G.1X
Job name
movie-data-analysis
Start time (Local)
05/25/2024 12:08:45
Glue version
4.0
Last modified on (Local)
05/25/2024 12:11:00

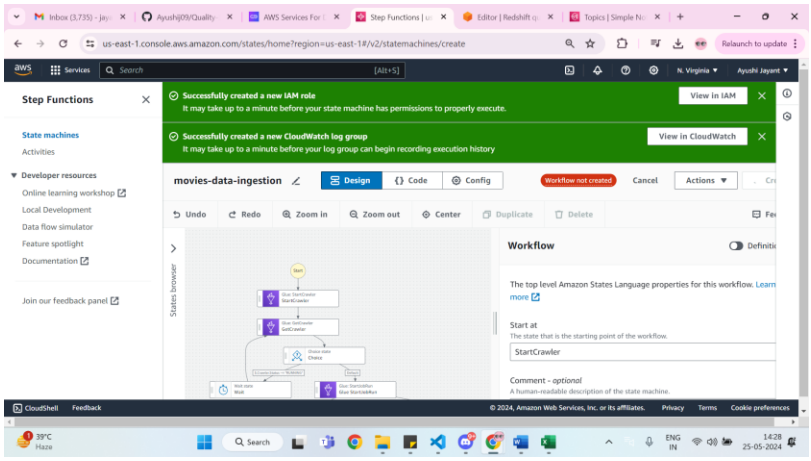

us-east-1.console.aws.amazon.com/sqlworkbench/home?region=us-east-1#/client
Redshift query editor v2
redshift-cluster-11
awsdatacatalog
dev
movies
imdb_mov...
public
sample_data_dev
select * from movies.imdb_movies_rating limit 5;
poster_link
series_title
released_year
certificate
runtime
https://m.media-amazon.c... The Shawshank Redempt... 1994 A 142 min
https://m.media-amazon.c... The Godfather 1972 A 175 min
https://m.media-amazon.c... The Dark Knight 2008 UA 152 min
https://m.media-amazon.c... The Godfather: Part II 1974 A 202 min
https://m.media-amazon.c... 12 Angry Men 1957 U 96 min
Query ID 16711
Elapsed time: 30 ms
Total rows: 5

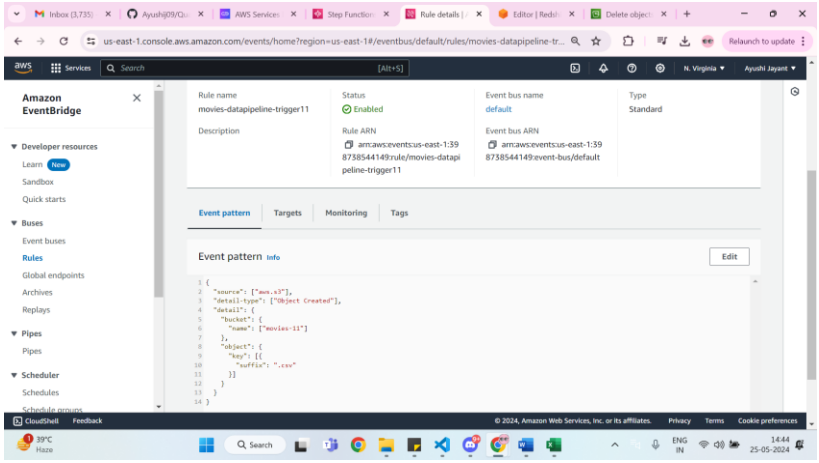

us-east-1.console.aws.amazon.com/s3/buckets/movies-11?region=us-east-1&bucketType=general&prefix=bad_records/...
Amazon S3
Objects
Properties
Objects (2) Info
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
run-AmazonS3_node1716554358334-1-part-r-00000
May 25, 2024, 12:10:45 (UTC+05:30)
788.4 KB
Standard
run-AmazonS3_node1716554358334-1-part-r-00000.snappy
snappy
May 25, 2024, 11:37:44 (UTC+05:30)
308.7 KB
Standard

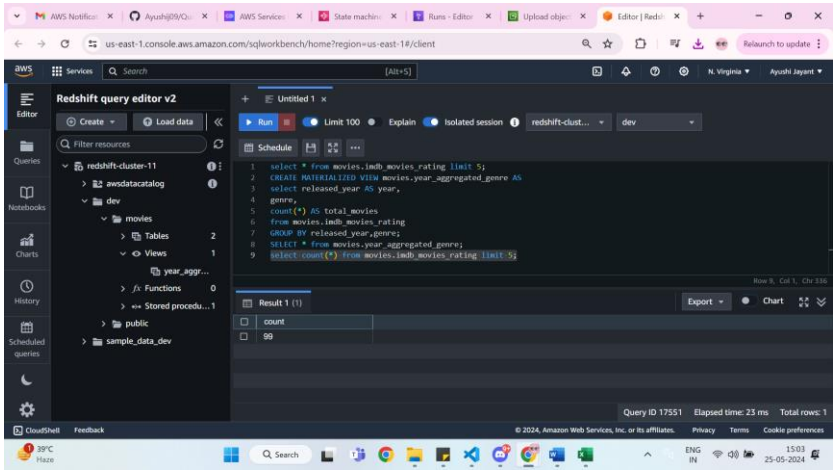


```sql
CREATE MATERIALIZED VIEW movies.year_aggregated_genre AS
select released_year AS year,
genre,
count(*) AS total_movies
from movies.imdb_movies_rating
GROUP BY released_year,genre;
```

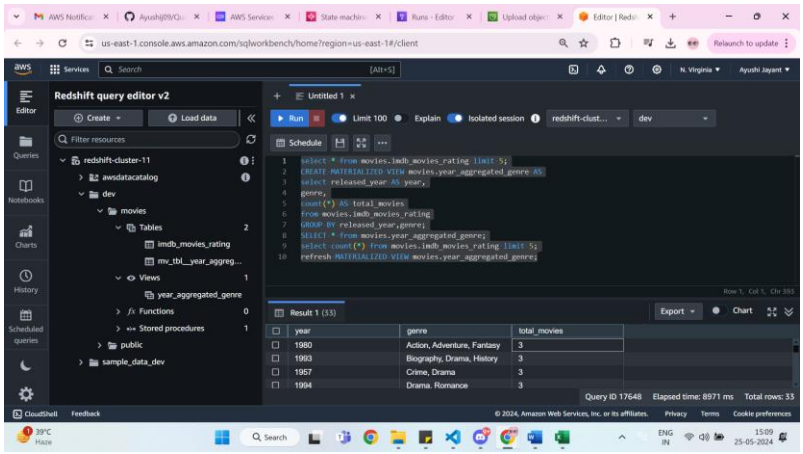








Now, if we check after refreshing the materialised vioew, the count has been tripled i.e. from 33→99


Redshift query editor v2
Create
Load data
Filter resources
redshift-cluster-11
awsdatacatalog
dev
movies
Tables 2
imdb_movies_rating
mv_tbl__year_aggreg...
Views 1
year_aggregated_genre
Functions 0
Stored procedures 1
public
sample_data_dev
Untitled 1
Run
Limit 100
Explain
Isolated session
redshift-clust...
dev
Schedule
select * from movies.imdb_movies_rating limit 5;
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;
SELECT * from movies.year_aggregated_genre;
select count(*) from movies.imdb_movies_rating limit 5;
refresh MATERIALIZED VIEW movies.year_aggregated_genre;
Result 1 (33)
Export
Chart
year
genre
total_movies
1980
Action, Adventure, Fantasy
3
1993
Biography, Drama, History
3
1957
Crime, Drama
3
1994
Drama, Romance
3
Query ID 17648
Elapsed time: 8971 ms
Total rows: 33
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences