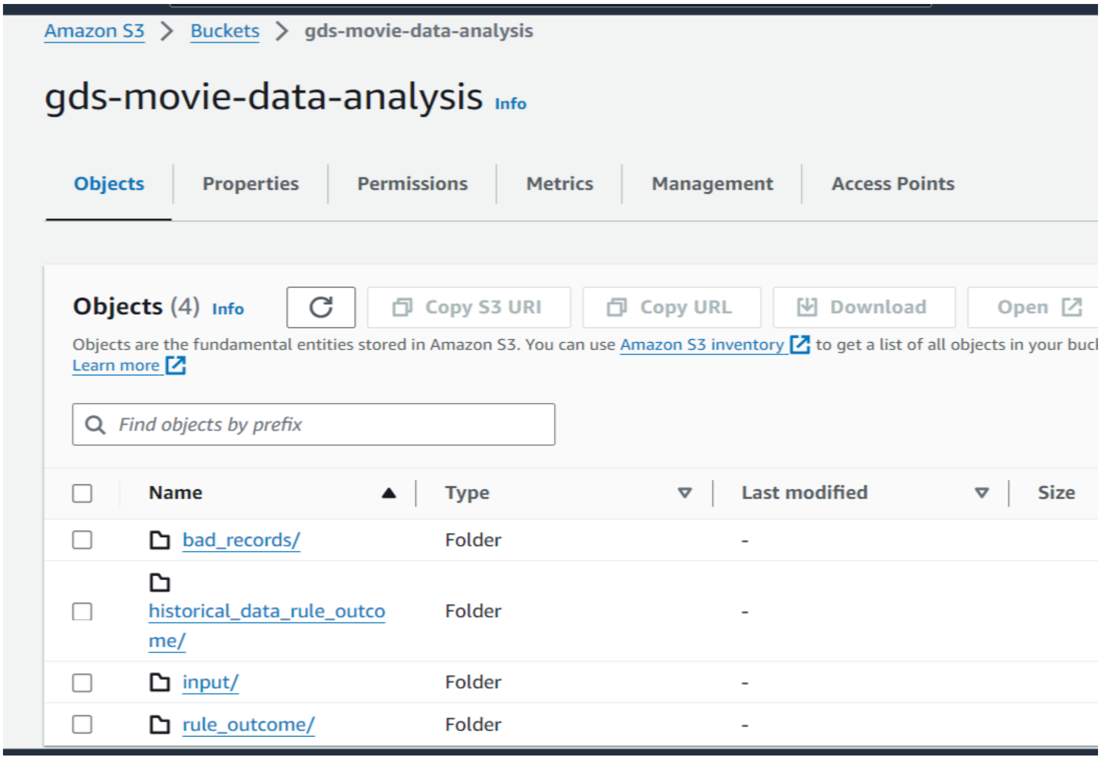
Quality Movie Data Analysis

1. create s3 bucket with bad records, historical_data_rule_outcome, input, rule_outcome



Amazon S3 > Buckets > gds-movie-data-analysis
Objects | Properties | Permissions | Metrics | Management | Access Points
Copy S3 URI
Copy URL
Download
Open

gds-movie-data-analysis Info

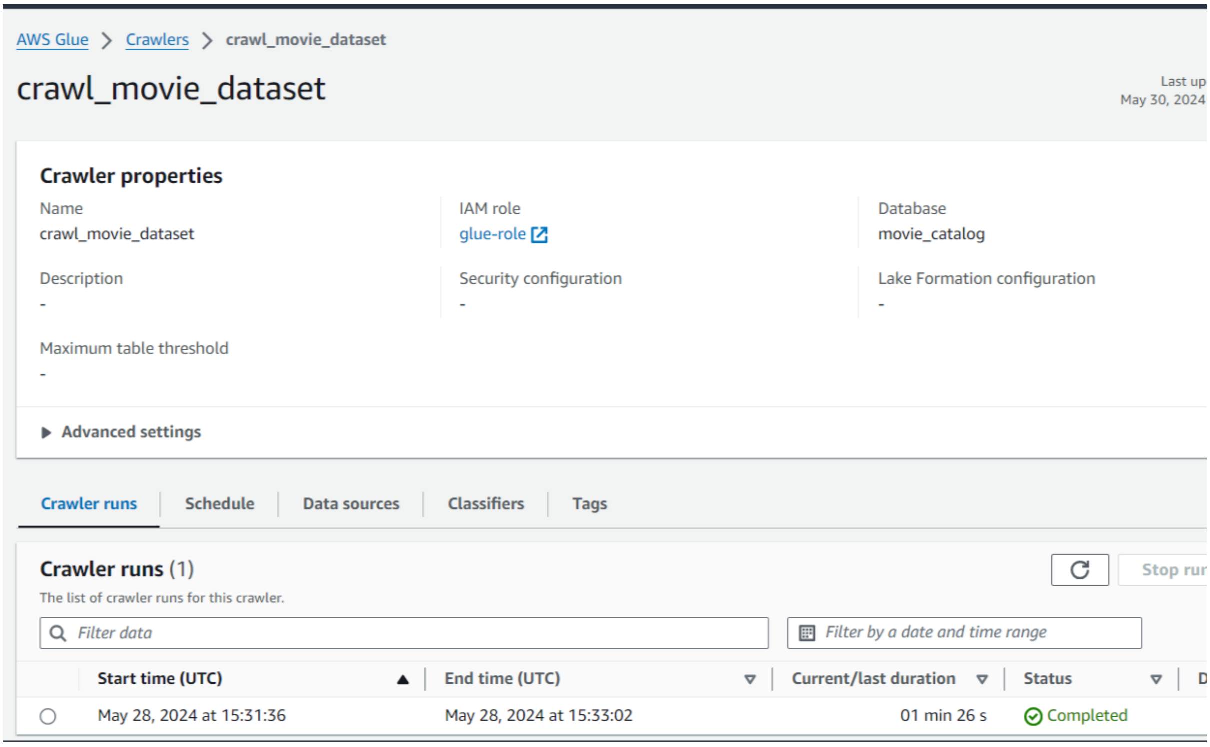
Objects (4) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucl
Learn more

Find objects by prefix

□	Name ▲	Type	Last modified	Size
□	bad_records/	Folder	-	
□	historical_data_rule_outcome/	Folder	-	
□	input/	Folder	-	
□	rule_outcome/	Folder	-	

2. created crawler on s3 input source where we also set a data quality rule to monitor the quality of our data assets
crawler:



AWS Glue > Crawlers > crawl_movie_dataset
Crawler runs | Schedule | Data sources | Classifiers | Tags
Stop rur

crawl_movie_dataset

Last up
May 30, 2024

Crawler properties

Name: crawl_movie_dataset
IAM role: glue-role
Database: movie_catalog
Description: -
Security configuration: -
Lake Formation configuration: -
Maximum table threshold: -

▶ Advanced settings

Crawler runs (1)

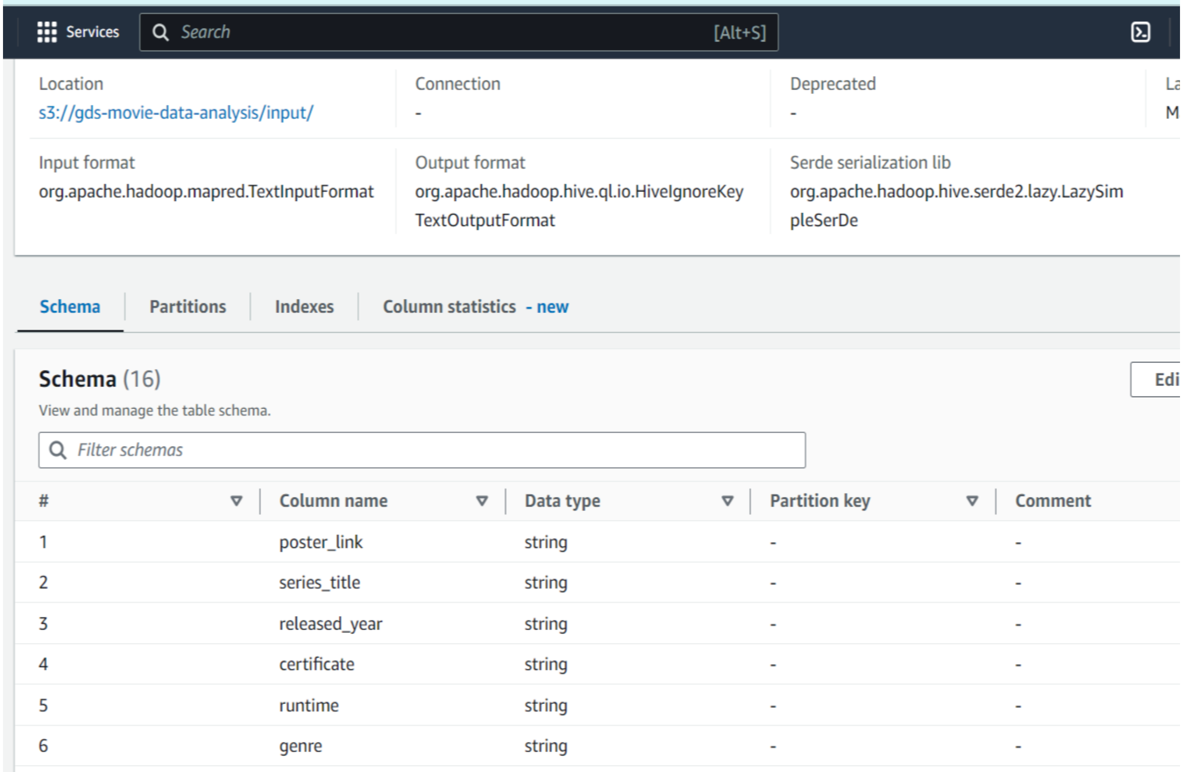
The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

	Start time (UTC) ▲	End time (UTC)	Current/last duration	Status	D
○	May 28, 2024 at 15:31:36	May 28, 2024 at 15:33:02	01 min 26 s	Completed	

input source catalog table:



Location
s3://gds-movie-data-analysis/input/

Connection
-

Deprecated
-

La
M

Input format
org.apache.hadoop.mapred.TextInputFormat

Output format
org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

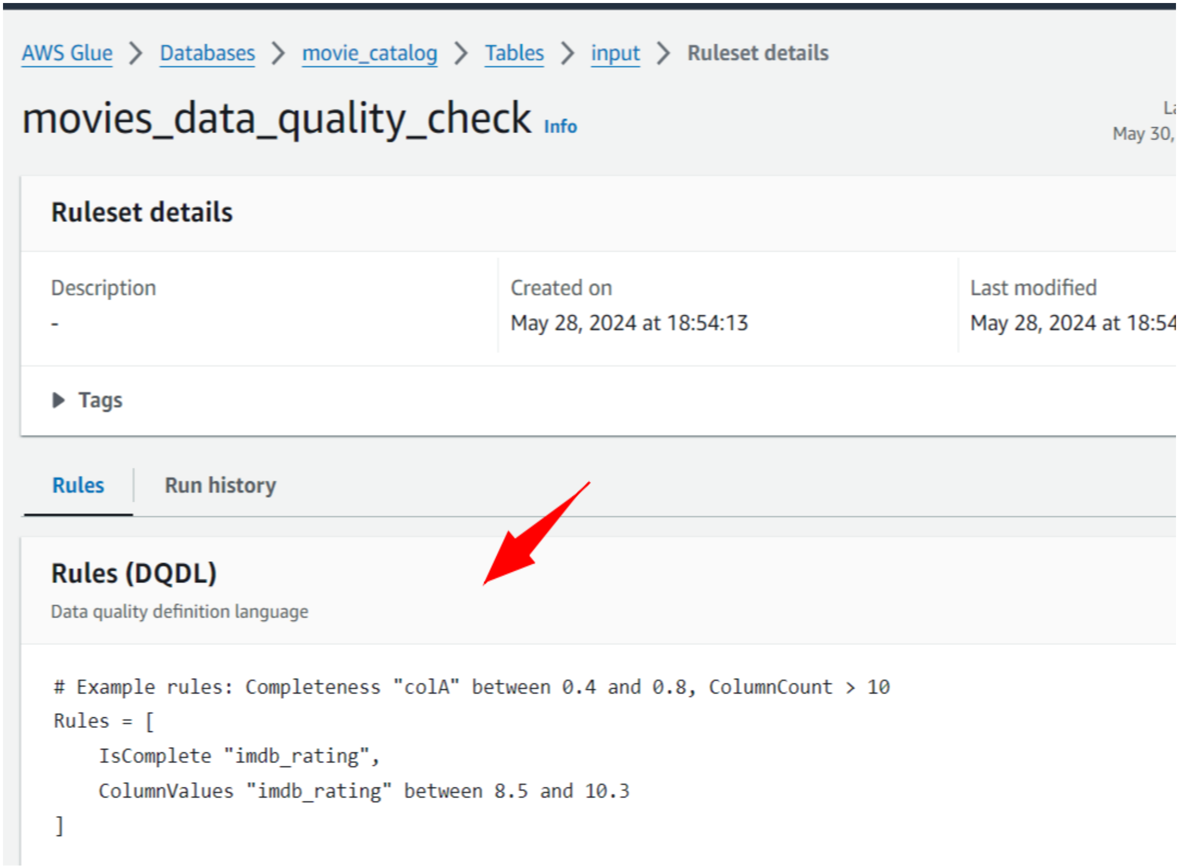
Serde serialization lib
org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Schema (16)
View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key	Comment
1	poster_link	string	-	-
2	series_title	string	-	-
3	released_year	string	-	-
4	certificate	string	-	-
5	runtime	string	-	-
6	genre	string	-	-

3. set a data quality rule to check their must be an imdb rating(not blank). Another rule is rating should consist between 8.5 and 10.3. On running we get passed and failed rule outcomes which was getting stored in s3 bucket another folder 'historical_data_rule_outcome'



movies_data_quality_check

Ruleset details

Description
-

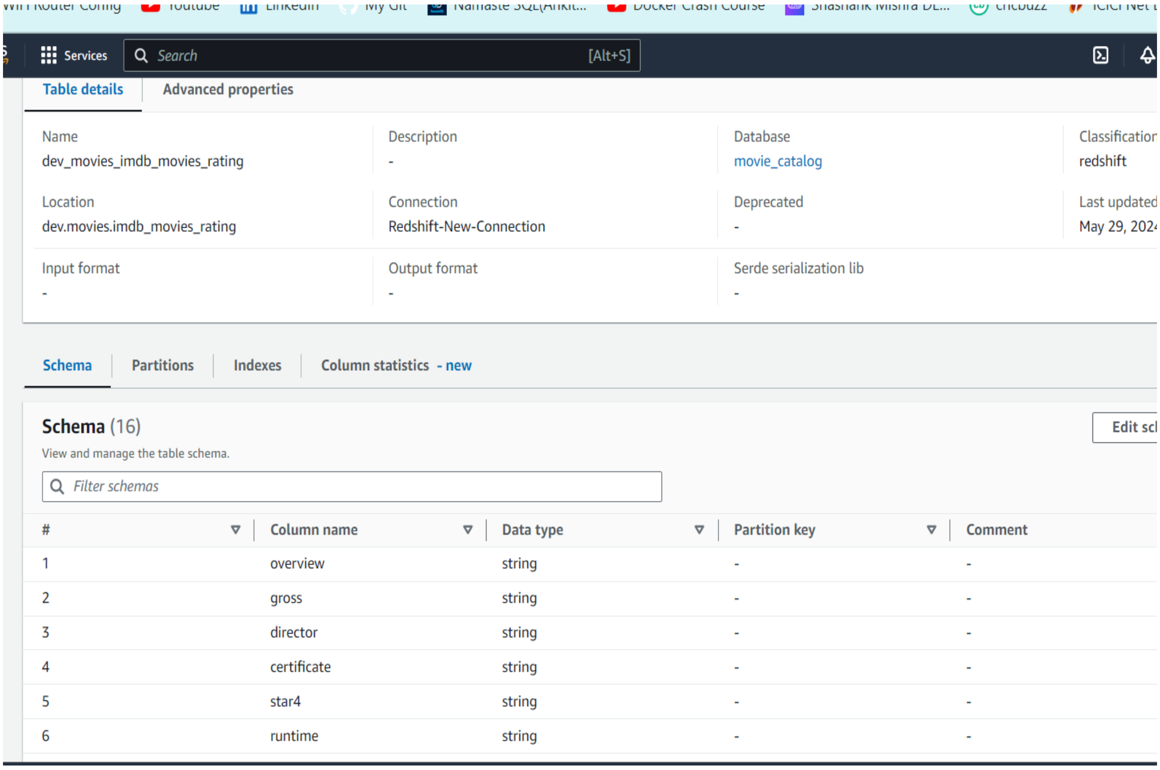
Created on
May 28, 2024 at 18:54:13

Last modified
May 28, 2024 at 18:54

Rules (DQDL)
Data quality definition language

```
# Example rules: Completeness "colA" between 0.4 and 0.8, ColumnCount > 10
Rules = [
    IsComplete "imdb_rating",
    ColumnValues "imdb_rating" between 8.5 and 10.3
]
```

4. Created a jdbc connection for connecting crawler to redshift target table. That crawler will be crawling the target redshift table where output results will be loaded



Name	Description	Database	Classificatior
dev_movies_imdb_movies_rating	-	movie_catalog	redshift

Location	Connection	Deprecated	Last updated
dev.movies.imdb_movies_rating	Redshift-New-Connection	-	May 29, 2024

Input format	Output format	Serde serialization lib
-	-	-

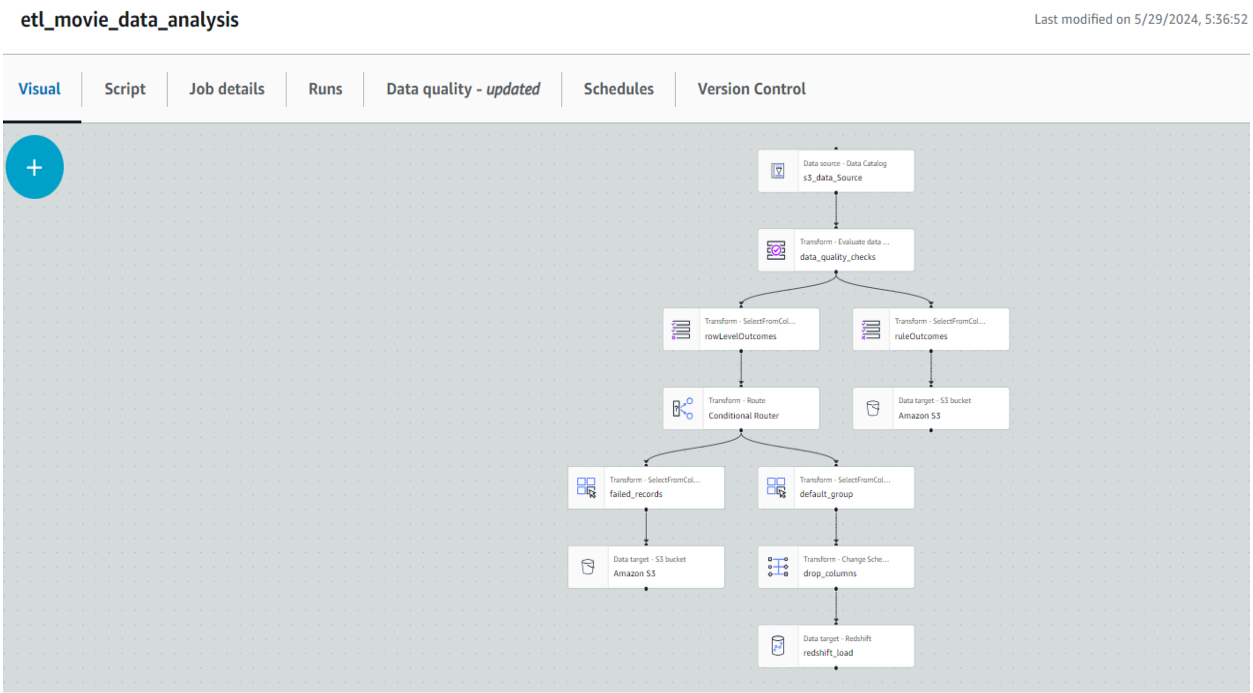
Schema (16)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	overview	string	-	-
2	gross	string	-	-
3	director	string	-	-
4	certificate	string	-	-
5	star4	string	-	-
6	runtime	string	-	-

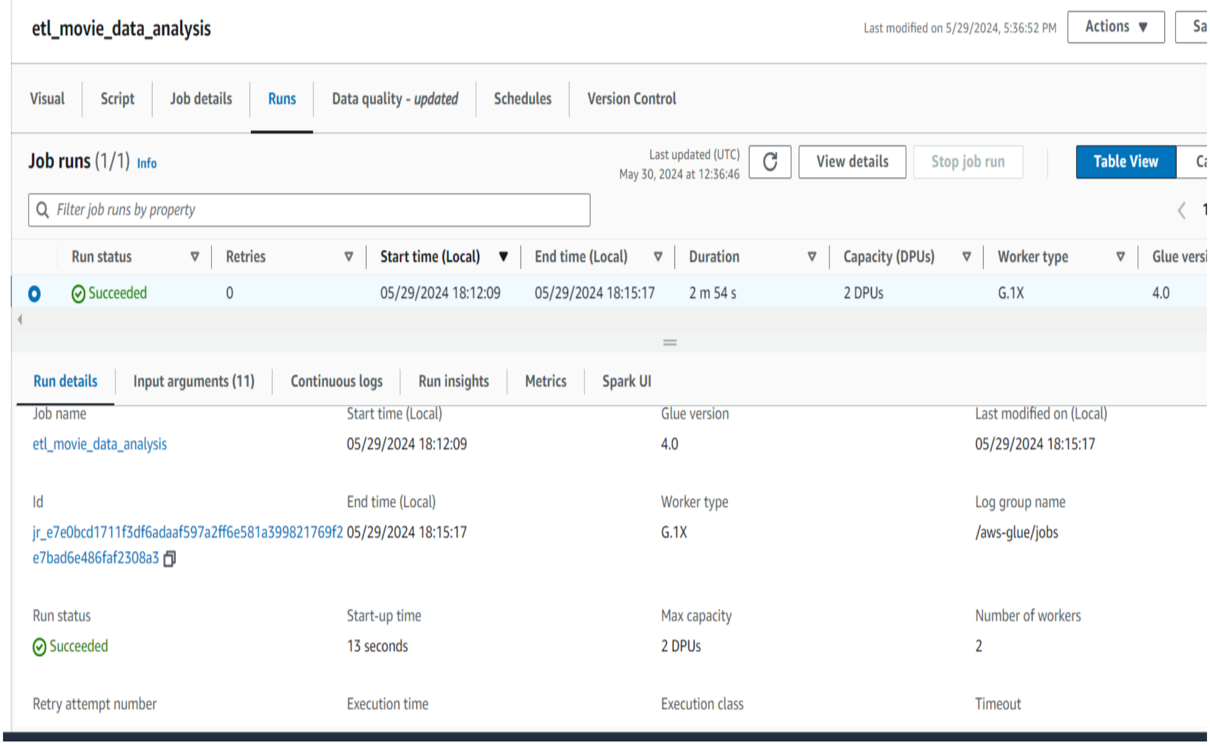
5. build an etl job where s3 input source being read from data catalog. Then we add our first transformation to include our data quality rule set where I also enabled 'add new columns on data quality errors'. We further applied two more transformations i.e., rowLevelOutcomes and ruleOutcomes

etl job visual representation:


etl_movie_data_analysis

Last modified on 5/29/2024, 5:36:52

success etl job runs:



	Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPUs)	Worker type	Glue versi
●	✓ Succeeded	0	05/29/2024 18:12:09	05/29/2024 18:15:17	2 m 54 s	2 DPUs	G.1X	4.0

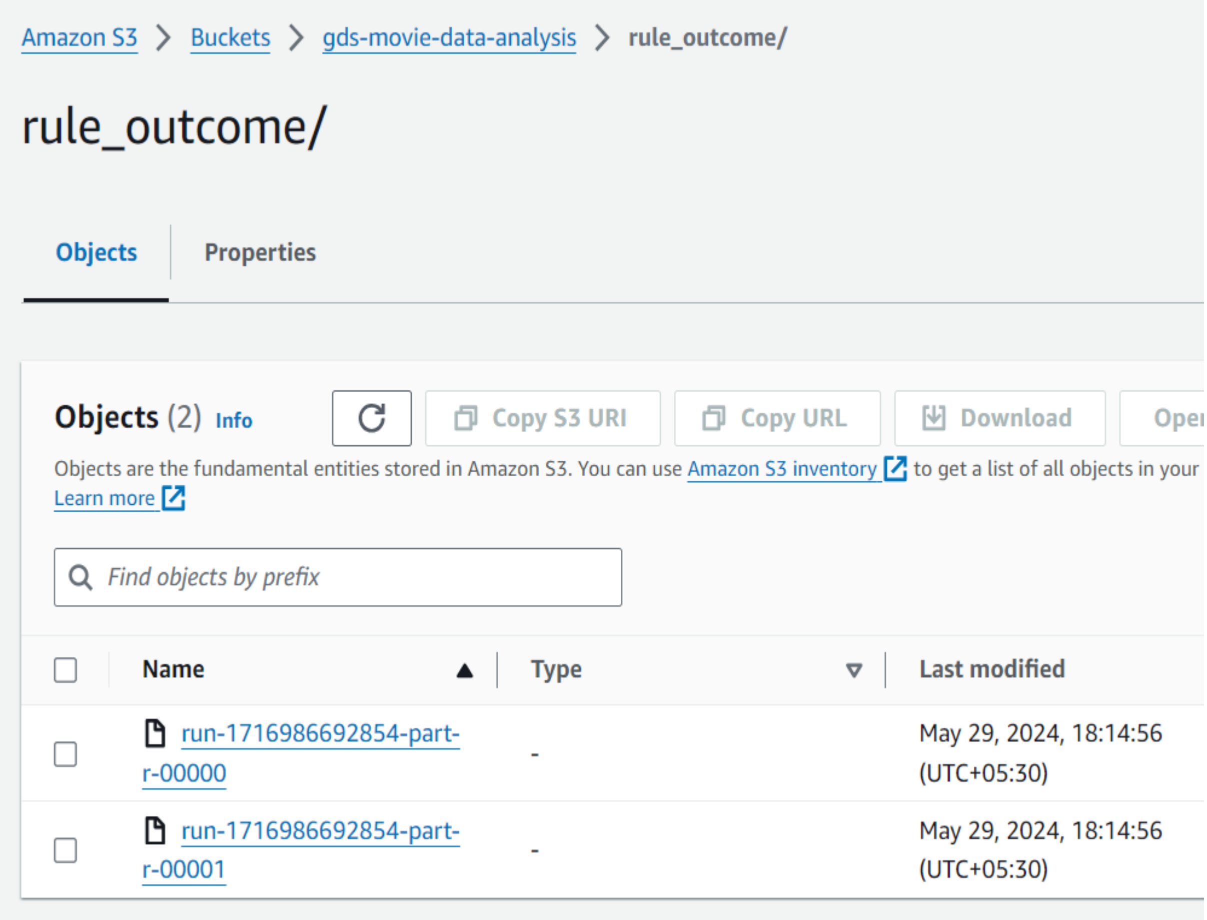
Job name: etl_movie_data_analysis
Start time (Local): 05/29/2024 18:12:09
Glue version: 4.0
Last modified on (Local): 05/29/2024 18:15:17

Id: jr_e7e0bcd1711f3df6adaaf597a2ff6e581a399821769f2e7bad6e486faf2308a3
End time (Local): 05/29/2024 18:15:17
Worker type: G.1X
Log group name: /aws-glue/jobs

Run status: ✓ Succeeded
Start-up time: 13 seconds
Max capacity: 2 DPUs
Number of workers: 2

Retry attempt number
Execution time
Execution class
Timeout

6. ruleOutcomes will be having rule wise outcome with pass or fail reason. And we were storing those in json format at s3 bucket another folder 'rule_outcome'.

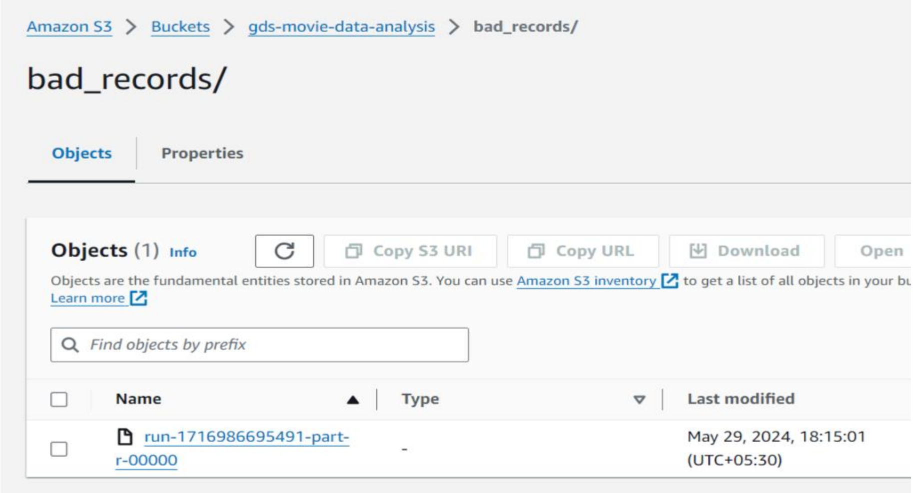


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your
Learn more

□	Name ▲	Type ▽	Last modified
□	run-1716986692854-part-r-00000	-	May 29, 2024, 18:14:56 (UTC+05:30)
□	run-1716986692854-part-r-00001	-	May 29, 2024, 18:14:56 (UTC+05:30)

7. On the other hand, rowLevelOutcomes will be having row wise outcome with pass or fail reason. Following rowLevelOutcomes, we further next added a conditional router where we checked with filter condition whether 'DataQualityEvaluationResult' matches 'Passed' value or not.

8. once we have groups on failed and passed records group. For failed records, we were storing data in json format at s3 bucket 'bad_records' folder.



Amazon S3 > Buckets > gds-movie-data-analysis > bad_records/

bad_records/

Objects | Properties

Objects (1) Info

Copy S3 URI | Copy URL | Download | Open

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bu

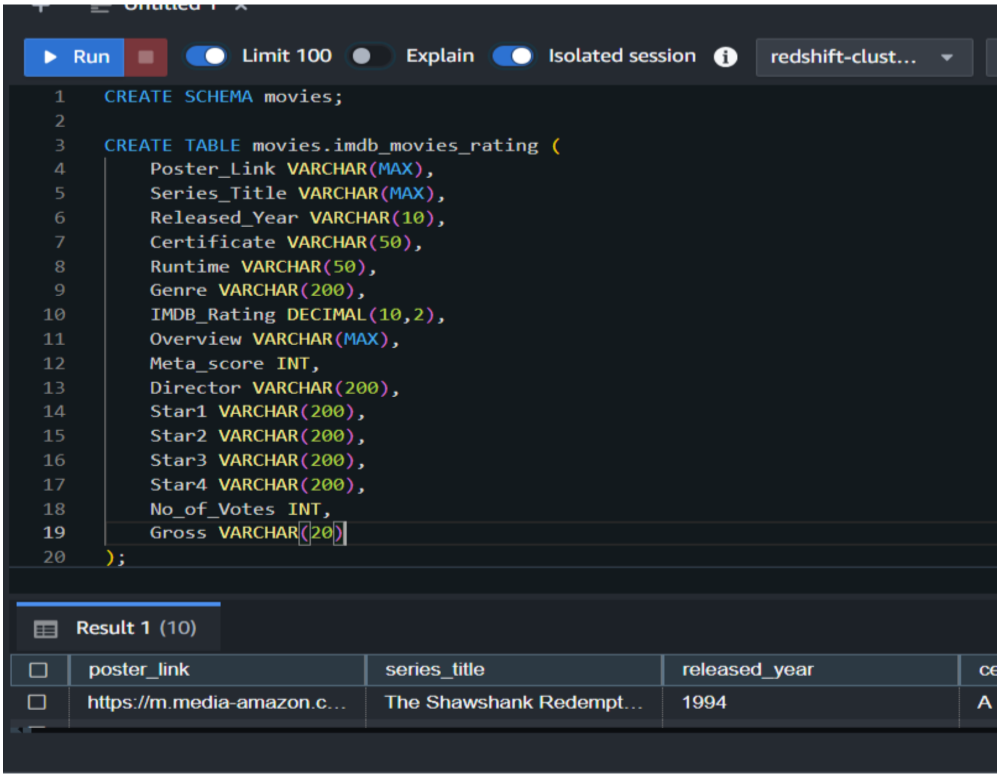
Learn more

Find objects by prefix

□	Name	Type	Last modified
□	run-1716986695491-part-r-00000	-	May 29, 2024, 18:15:01 (UTC+05:30)

9. For passed records, first we were applying 'change schema' transform where we matched both the source key and target key(redshift target table) column's data types and also dropping the 'data quality' additional columns.

10. As we already created a target redshift table 'imdb_movies_rating', so next we uploaded passed records to the target redshift table where assigned 'IAM' access role with the redshift cluster



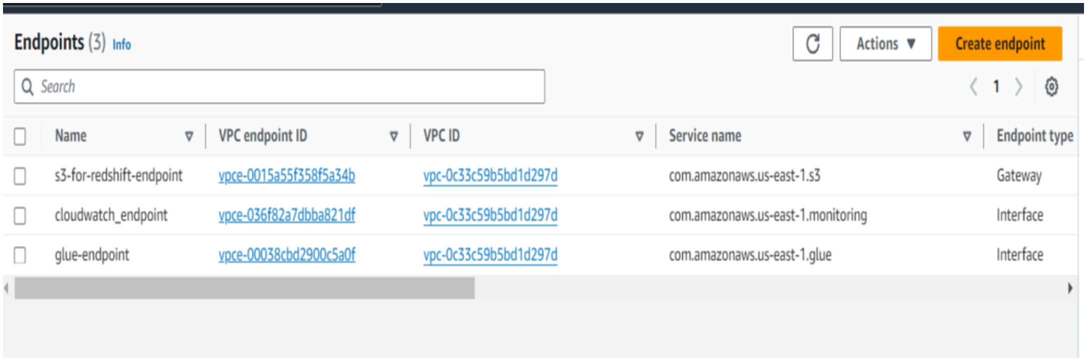
```
CREATE SCHEMA movies;

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)
);
```

Result 1 (10)

□	poster_link	series_title	released_year	ce
□	https://m.media-amazon.c...	The Shawshank Redempt...	1994	A

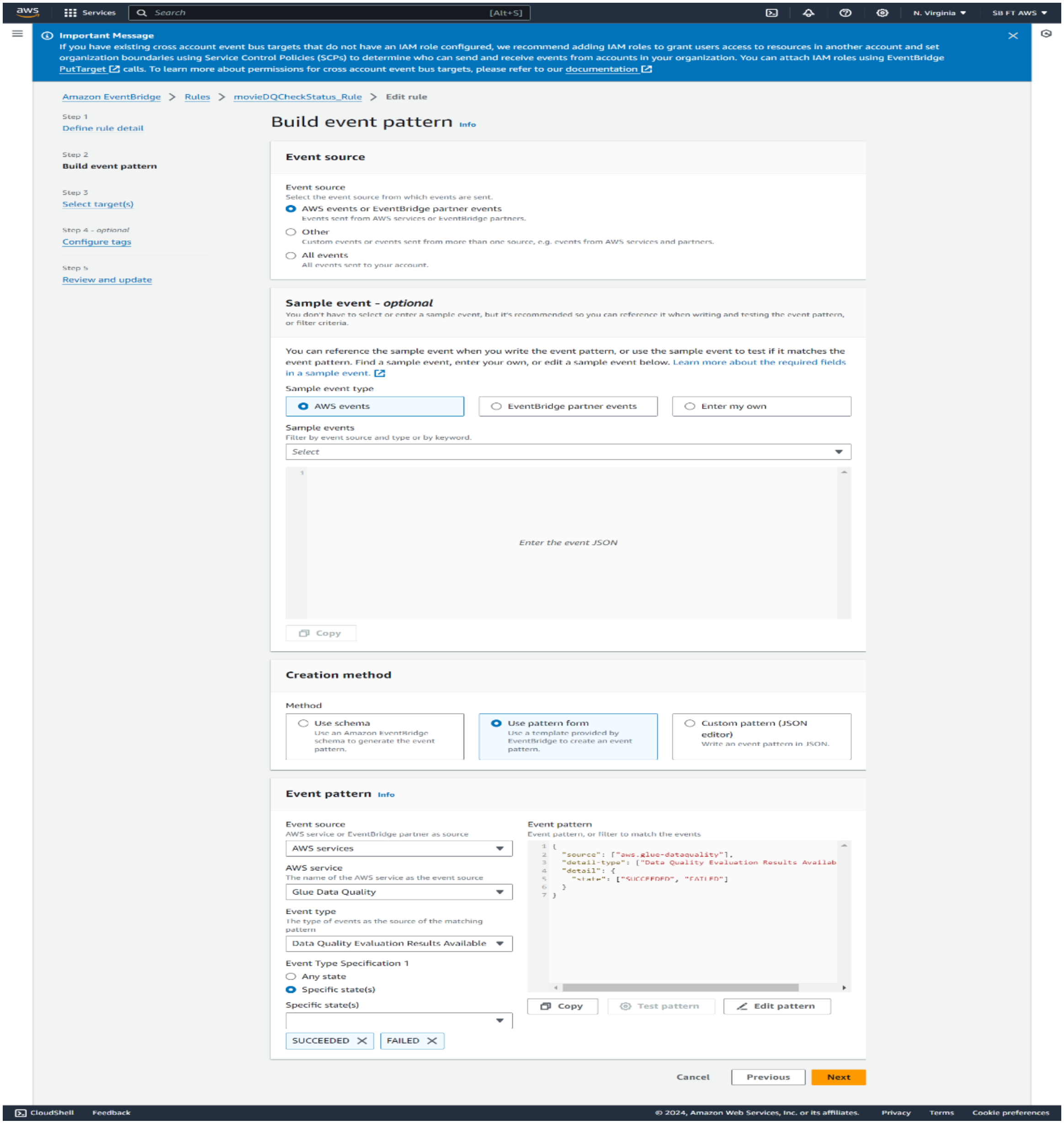
11. Also, added vpc endpoints for glue and cloudwatch under same vpc id(allows for private, secure communication between our Glue jobs & CloudWatch services with no expose to public internet) included in redshift cluster.



Endpoints (3) Info

Name	VPC endpoint ID	VPC ID	Service name	Endpoint type
s3-for-redshift-endpoint	vpce-0015a55f358f5a34b	vpc-0c33c59b5bd1d297d	com.amazonaws.us-east-1.s3	Gateway
cloudwatch_endpoint	vpce-036f82a7dbba821df	vpc-0c33c59b5bd1d297d	com.amazonaws.us-east-1.monitoring	Interface
glue-endpoint	vpce-00038cbd2900c5a0f	vpc-0c33c59b5bd1d297d	com.amazonaws.us-east-1.glue	Interface

12. We next made an eventbridge rule where a pattern is defined to check receiving event's data quality status using 'AWS Glue Data Quality' service’s event type 'Data Quality Evaluation Results Available'.



Important Message

If you have existing cross account event bus targets that do not have an IAM role configured, we recommend adding IAM roles to grant users access to resources in another account and set organization boundaries using Service Control Policies (SCPs) to determine who can send and receive events from accounts in your organization. You can attach IAM roles using EventBridge PutTarget calls. To learn more about permissions for cross account event bus targets, please refer to our documentation

Amazon EventBridge > Rules > movieDQCheckStatus_Rule > Edit rule

Step 1 Define rule detail
Step 2 Build event pattern
Step 3 Select target(s)
Step 4 - *optional* Configure tags
Step 5 Review and update

Build event pattern Info

Event source

Event source
Select the event source from which events are sent.
● AWS events or EventBridge partner events
Events sent from AWS services or EventBridge partners.
○ Other
Custom events or events sent from more than one source, e.g. events from AWS services and partners.
○ All events
All events sent to your account.

Sample event - *optional*

You don't have to select or enter a sample event, but it's recommended so you can reference it when writing and testing the event pattern, or filter criteria.

You can reference the sample event when you write the event pattern, or use the sample event to test if it matches the event pattern. Find a sample event, enter your own, or edit a sample event below. Learn more about the required fields in a sample event.

Sample event type
● AWS events ○ EventBridge partner events ○ Enter my own

Sample events
Filter by event source and type or by keyword.
Select

Enter the event JSON

Creation method

Method
○ Use schema
Use an Amazon EventBridge schema to generate the event pattern.
● Use pattern form
Use a template provided by EventBridge to create an event pattern.
○ Custom pattern (JSON editor)
Write an event pattern in JSON.

Event pattern Info

Event source
AWS service or EventBridge partner as source
AWS services

AWS service
The name of the AWS service as the event source
Glue Data Quality

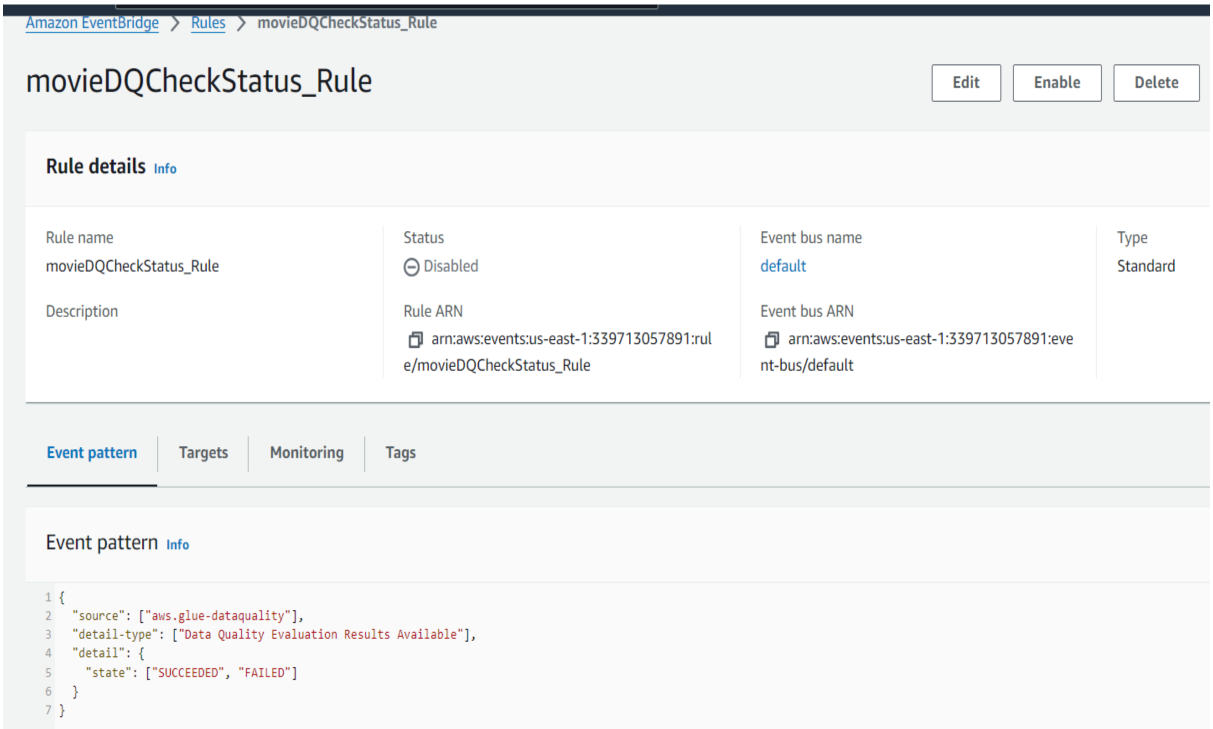
Event type
The type of events as the source of the matching pattern
Data Quality Evaluation Results Available

Event Type Specification 1
○ Any state
● Specific state(s)

Specific state(s)
SUCCEEDED FAILED

Event pattern
Event pattern, or filter to match the events

```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Availab
  "detail": {
    "state": ["SUCCEEDED", "FAILED"]
  }
}
```



movieDQCheckStatus_Rule

Rule details

Rule name
movieDQCheckStatus_Rule

Description

Status
⊖ Disabled

Rule ARN
arn:aws:events:us-east-1:339713057891:rule/movieDQCheckStatus_Rule

Event bus name
default

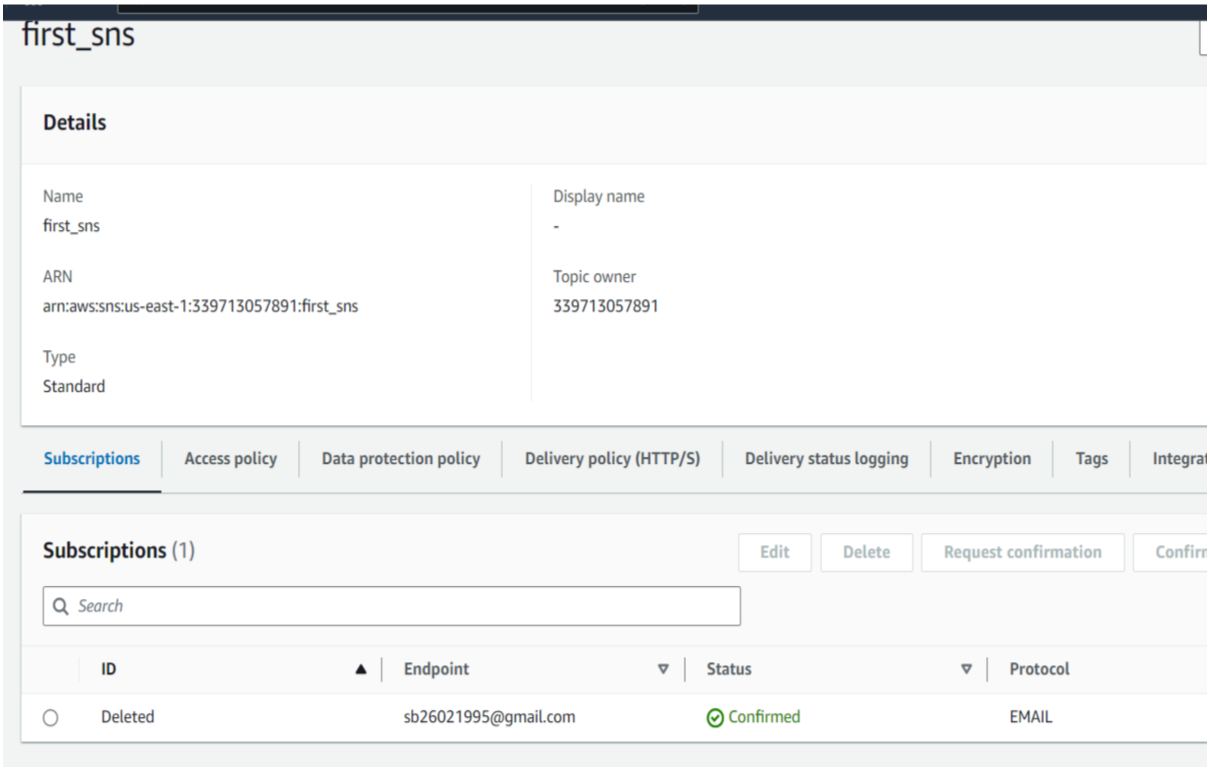
Event bus ARN
arn:aws:events:us-east-1:339713057891:event-bus/default

Type
Standard

Event pattern

```
{
  "source": ["aws.glue-dataquality"],
  "detail-type": ["Data Quality Evaluation Results Available"],
  "detail": {
    "state": ["SUCCEEDED", "FAILED"]
  }
}
```

13. We then invoke a target SNS topic service. We were sending success/fail notification using this service.



first_sns

Details

Name
first_sns

ARN
arn:aws:sns:us-east-1:339713057891:first_sns

Type
Standard

Display name
-

Topic owner
339713057891

Subscriptions (1)

	ID ▲	Endpoint ▽	Status ▽	Protocol
○	Deleted	sb26021995@gmail.com	✓ Confirmed	EMAIL