

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

gds-movie-data-analysis [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (4) [Info](#) [Refresh](#) [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 bucket. [Learn more](#)

| <input type="checkbox"/> | Name | Type | Last modified | Size |
|--------------------------|---|--------|---------------|------|
| <input type="checkbox"/> | bad_records/ | Folder | - | |
| <input type="checkbox"/> | historical_data_rule_outcome/ | Folder | - | |
| <input type="checkbox"/> | input/ | Folder | - | |
| <input type="checkbox"/> | rule_outcome/ | Folder | - | |

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

AWS Glue > Crawlers > crawl_movie_dataset

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](#) | [Schedule](#) | [Data sources](#) | [Classifiers](#) | [Tags](#)

Crawler runs (1) [Refresh](#) [Stop runs](#)

The list of crawler runs for this crawler.

| <input type="checkbox"/> | Start time (UTC) | End time (UTC) | Current/last duration | Status |
|--------------------------|--------------------------|--------------------------|-----------------------|-----------|
| <input type="radio"/> | May 28, 2024 at 15:31:36 | May 28, 2024 at 15:33:02 | 01 min 26 s | Completed |

input source catalog table:

Services [Alt+S]

| | | | |
|--|---|---|---------|
| 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.HiveIgnoreKey TextOutputFormat | Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe | |

Schema | Partitions | Indexes | Column statistics - new

Schema (16)

View and manage the table schema.

| # | 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'

[AWS Glue](#) > [Databases](#) > [movie_catalog](#) > [Tables](#) > [input](#) > Ruleset details

movies_data_quality_check Info

May 30,

Ruleset details

| | | |
|------------------|--|--|
| Description - | Created on May 28, 2024 at 18:54:13 | Last modified May 28, 2024 at 18:54 |
|------------------|--|--|

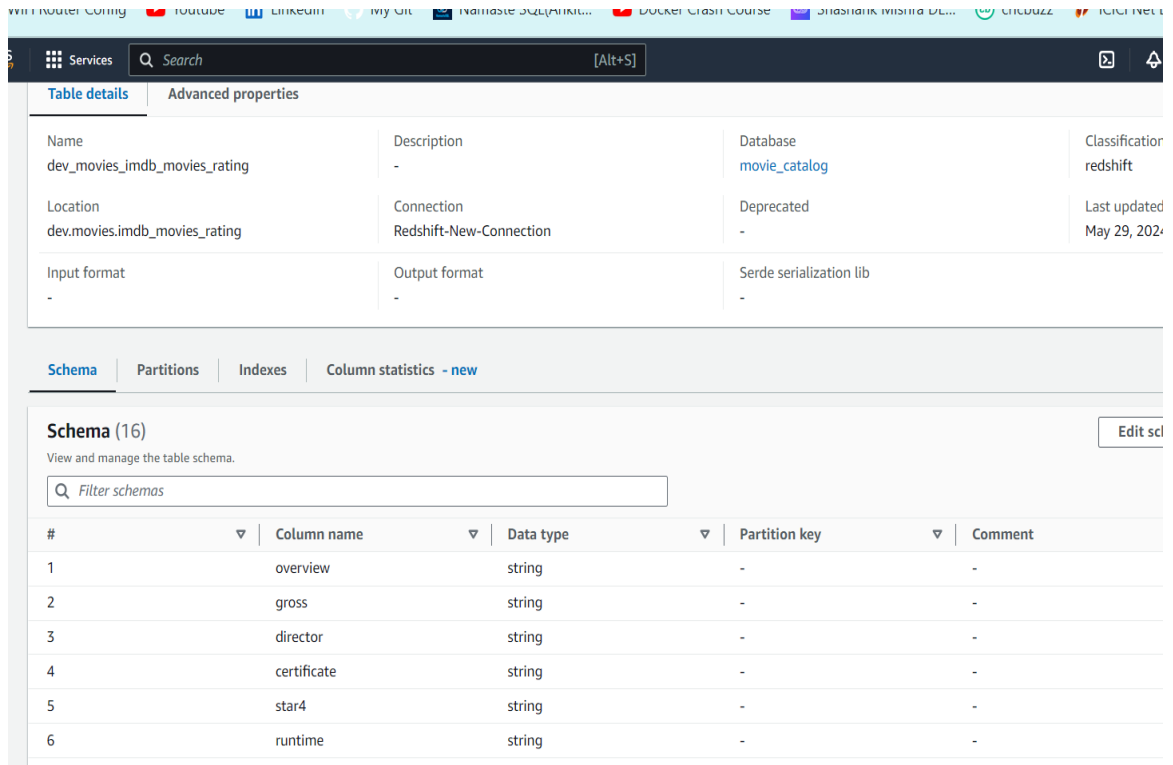
Tags

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
]
```

- 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



The screenshot shows the AWS Glue console interface. At the top, there's a search bar and tabs for 'Table details' and 'Advanced properties'. The 'Table details' tab is active, displaying a table with 4 columns: Name, Description, Database, and Classification. Below this, there's a section for 'Schema (16)' with a search bar and a table listing 6 columns: #, Column name, Data type, Partition key, and Comment. The table lists columns: overview (string), gross (string), director (string), certificate (string), star4 (string), and runtime (string).

| Name | Description | Database | Classification |
|-------------------------------|-------------------------|-------------------------|----------------|
| 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 | |
| - | - | - | |

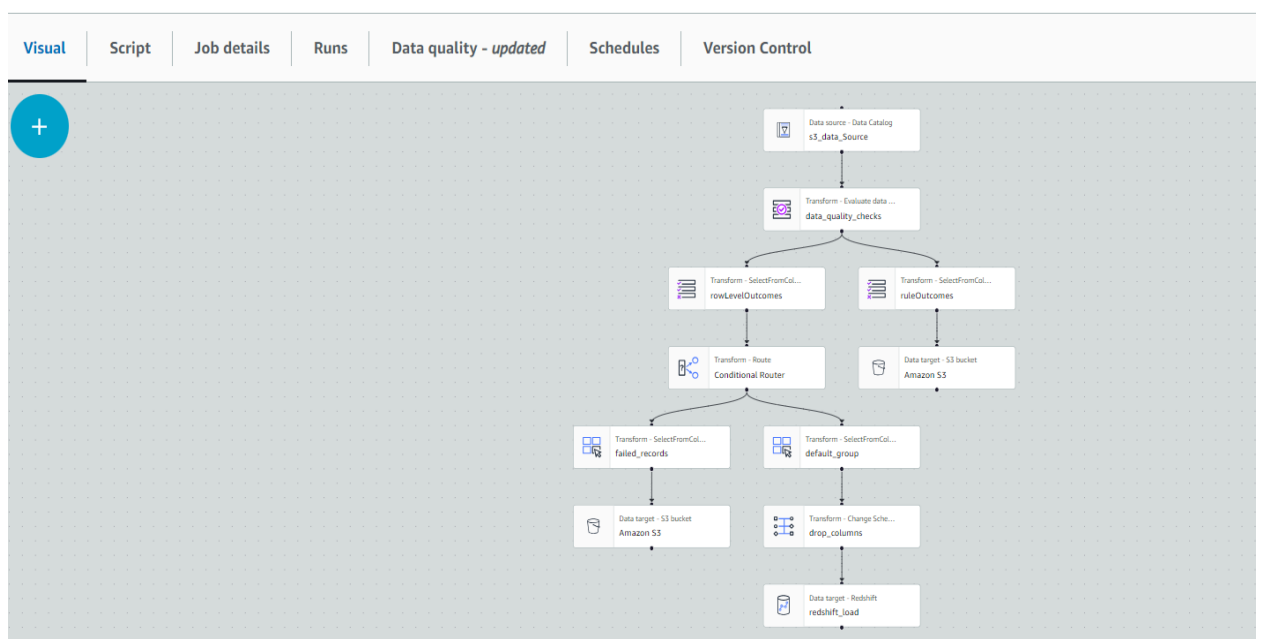
| # | Column name | Data type | Partition key | Comment |
|---|-------------|-----------|---------------|---------|
| 1 | overview | string | - | - |
| 2 | gross | string | - | - |
| 3 | director | string | - | - |
| 4 | certificate | string | - | - |
| 5 | star4 | string | - | - |
| 6 | runtime | string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

- 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:

etl_movie_data_analysis

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

Actions

Sa

Visual

Script

Job details

Runs

Data quality - updated

Schedules

Version Control

Job runs (1/1)

Info

Last updated (UTC)
May 30, 2024 at 12:36:46

View details

Stop job run

Table View

Ca

Filter job runs by property

< 1

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

Run details

Input arguments (11)

Continuous logs

Run insights

Metrics

Spark UI

Job name

Start time (Local)

Glue version

Last modified on (Local)

etl_movie_data_analysis

05/29/2024 18:12:09

4.0

05/29/2024 18:15:17

Id

End time (Local)

Worker type

Log group name

jr_e7e0bcd1711f3df6adaaf597a2ff6e581a399821769f2
e7bad6e486faf2308a3

05/29/2024 18:15:17

G.1X

/aws-glue/jobs

Run status

Start-up time

Max capacity

Number of workers

Succeeded

13 seconds

2 DPUs

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'.

Amazon S3

>

Buckets

>

gds-movie-data-analysis

>

rule_outcome/

rule_outcome/

Objects

Properties

Objects (2)

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 [Learn more](#)

Find objects by prefix

Name

Type

Last modified

run-1716986692854-part-r-00000

-

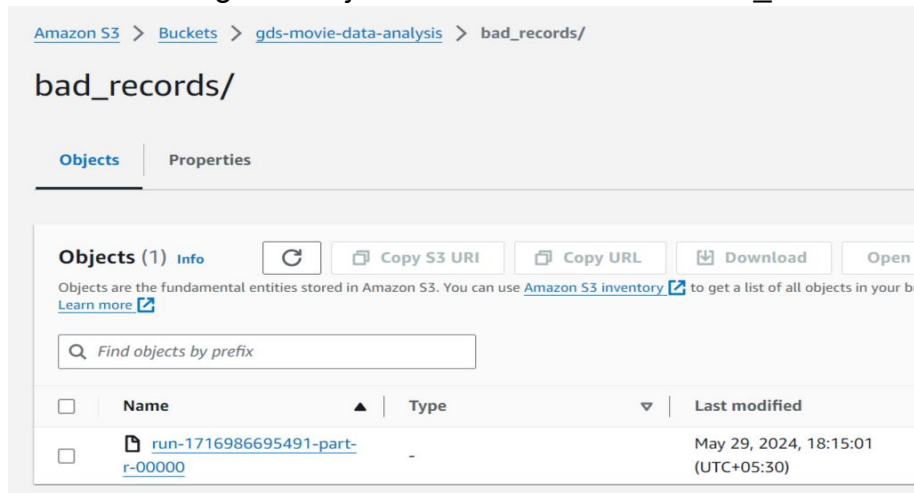
May 29, 2024, 18:14:56 (UTC+05:30)

run-1716986692854-part-r-00001

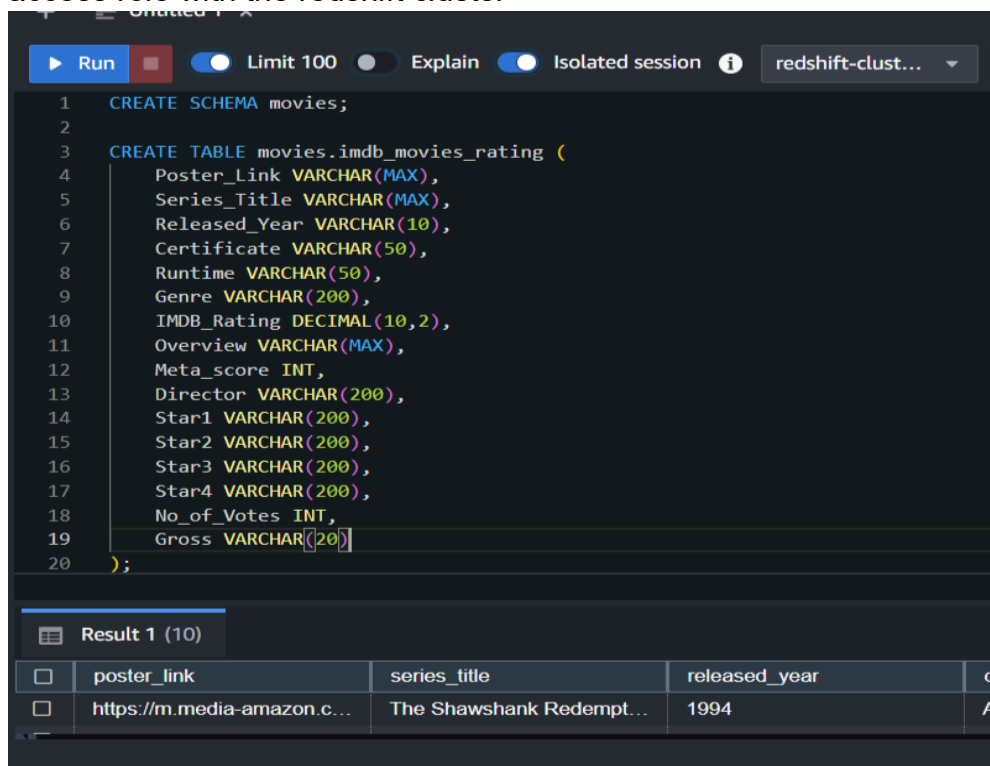
-

May 29, 2024, 18:14:56 (UTC+05:30)

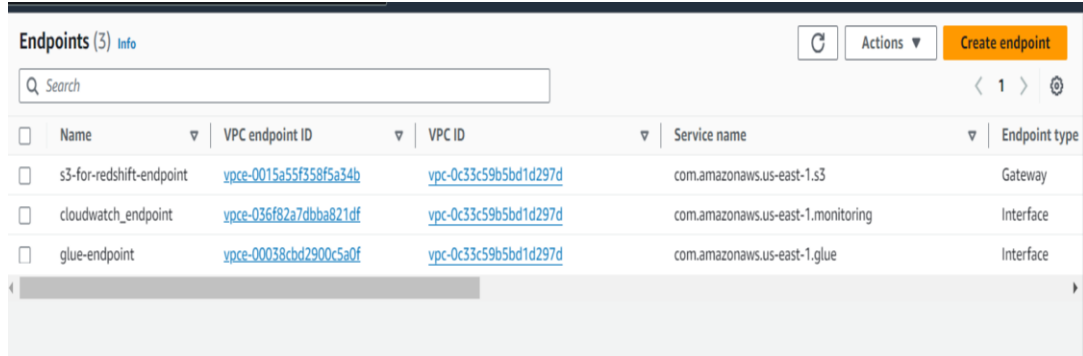
7. On the other hand, rowLevelOutcomes will be having row wise outcome with pass or fail reason. Following rowLevelOutcomes, we further next added a conditional router where we checked with filter condition whether 'DataQualityEvaluationResult' matches 'Passed' value or not.
8. once we have groups on failed and passed records group. For failed records, we were storing data in json format at s3 bucket 'bad_records' folder.



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



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



The screenshot shows the AWS VPC Endpoints console. At the top, there's a header with 'Endpoints (3)' and an 'Info' link. To the right are buttons for 'Create endpoint' and 'Actions'. Below the header is a search bar. The main content is a table with the following columns: Name, VPC endpoint ID, VPC ID, Service name, and Endpoint type. There are three rows of endpoints listed.

| 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 with event type 'Data Quality Evaluation Results Available'.

Services

Search

[Alt+S]

N. Virginia

SB F1 AWS

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

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

1

Enter the event JSON

Copy

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

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

```
1 {
2   "source": ["aws.glue-dataquality"],
3   "detail-type": ["Data Quality Evaluation Results Availab
4   "detail": {
5     "state": ["SUCCEEDED", "FAILED"]
6   }
7 }
```

Copy Test pattern Edit pattern

Cancel

Previous

Next

CloudShell

Feedback

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon EventBridge > Rules > movieDQCheckStatus_Rule

movieDQCheckStatus_Rule

Edit Enable Delete

Rule details [Info](#)

| | | | |
|--------------------------------------|--|--|------------------|
| Rule name movieDQCheckStatus_Rule | Status ⊖ Disabled | Event bus name default | Type Standard |
| Description | Rule ARN arn:aws:events:us-east-1:339713057891:rule/movieDQCheckStatus_Rule | Event bus ARN arn:aws:events:us-east-1:339713057891:event-bus/default | |

Event pattern [Info](#)

```
1 {
2   "source": ["aws.glue-dataquality"],
3   "detail-type": ["Data Quality Evaluation Results Available"],
4   "detail": {
5     "state": ["SUCCEEDED", "FAILED"]
6   }
7 }
```

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

first_sns

Details

| | |
|---|-----------------------------|
| Name first_sns | Display name - |
| ARN arn:aws:sns:us-east-1:339713057891:first_sns | Topic owner 339713057891 |
| Type Standard | |

[Subscriptions](#) | [Access policy](#) | [Data protection policy](#) | [Delivery policy \(HTTP/S\)](#) | [Delivery status logging](#) | [Encryption](#) | [Tags](#) | [Integrations](#)

Subscriptions (1)

Edit Delete Request confirmation Confirm

Q Search

| ID | Endpoint | Status | Protocol |
|-----------|----------------------|-------------|----------|
| ○ Deleted | sb26021995@gmail.com | ✔ Confirmed | EMAIL |