

CESAR School

Engenharia em Análise de dados

Disciplina: Computação em Nuvem

Aluno: Erike Simon Costa Cativo do Nascimento

Desafio Final

## Escopo do Desafio

O novo aplicativo de streaming "5GFlix" está com o desafio de fazer estudos de mercado para formular a estratégia de negócio que irão adotar.

Para uma parte do estudo de mercado, a "5GFlix" precisa fazer análises em cima de filmes e séries que estão disponíveis na Netflix, a sua concorrente direta.

O CTO da "5GFlix", Alan Turing, entrou em contato com a turma para construir uma estrutura lógica que possibilite que o time de BI da "5GFlix" responda a várias perguntas de negócio relacionadas aos dados da Netflix, detalhadas nos entregáveis deste desafio.

Para poder realizar as análises foi fornecida a seguinte base de dados:

- Base1:  
<https://drive.google.com/file/d/1gLsCjaMrL91ECdThq58cZAzB9tPxG18g/view?usp=sharing>
- Base2:  
[https://drive.google.com/file/d/1C\\_T1w8fc7Oa8MeTo4LMTEcv90IfEOS-6/view?usp=sharing](https://drive.google.com/file/d/1C_T1w8fc7Oa8MeTo4LMTEcv90IfEOS-6/view?usp=sharing)

Para solucionar o desafio, você deve baixar os arquivos, processar os arquivos caso julgue necessários para facilitar responder as perguntas (o processamento pode ser feito em qualquer lugar, até localmente, pois não será avaliado neste trabalho), criar um bucket do Amazon S3, colocar os arquivos lá dentro, depois construir uma Tabela no Amazon Athena que consultará os dados desse Bucket e responder às perguntas utilizando SQL do Athena.

Descrição do Dataset:

Base 1:

1. ID do filme
2. título e ano de lançamento

Base 2:

1. Cust\_Id: ID do customer que fez a avaliação
2. Rating: avaliação (nota)

3. Date: data da avaliação
4. Movie\_Id: ID do filme

## Entregáveis

1. Queries utilizadas para responder às seguintes perguntas:
  - 1.1. Quantos filmes estão disponíveis no dataset?
  - 1.2. Qual é o nome dos 5 filmes com melhor média de avaliação?
  - 1.3. Quais os 9 anos com menos lançamentos de filmes?
  - 1.4. Quantos filmes que possuem avaliação maior ou igual a 4.7, considerando apenas os filmes avaliados na última data de avaliação do dataset?
  - 1.5. Quais os id's dos 5 customers que mais avaliaram filmes e quantas avaliações cada um fez?
2. Documentação
  - 2.1. Explicação do passo a passo necessário para rodar o código
  - 2.2. Print mostrando os dados no bucket do Amazon S3
  - 2.3. Print do Amazon Athena com a query executada e a resposta obtida
3. Você deverá compartilhar seu trabalho em um Documento (Google Docs, Word ou similar).

## Tecnologias

- Amazon S3
- Amazon Athena
- SQL
- Python (opcional para processar os dados caso julgue necessário)

## Desenvolvimento do Desafio

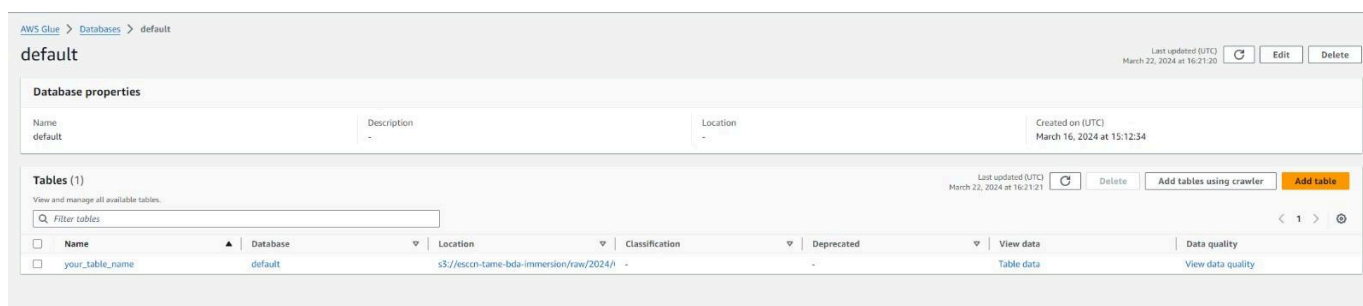
Para responder às perguntas de negócio do streaming 5GFlix, o primeiro passo foi realizar uma análise e pré-processamento nas bases 1 e 2, de maneira a corrigir algum possível problema nos dados e transformá-los em arquivo *.parquet*. Acesse o link do notebook [aqui](#). O formato *.parquet* possui várias vantagens em comparação a arquivos com outros formatos de dados (como JSON ou CSV) em relação a vários aspectos. É otimizado para consultas analíticas, implicando em um maior desempenho com um baixo tempo de processamento; possui compactação eficiente, reduzindo custos de armazenamento no Amazon S3, dentre outras vantagens.

O Amazon Athena permite consultar dados armazenados no Amazon S3 usando consultas SQL padrão. No contexto do Athena, um banco de dados é uma coleção lógica de tabelas que são usadas para acessar os dados, com cada tabela possuindo um único esquema que define os tipos de dados, colunas e outras propriedades. É possível usar o próprio Athena para criar um esquema e depois usá-lo ou utilizar o AWS Glue Catalog, que é um serviço de ETL que pode ser usado para catalogar automaticamente os metadados dos dados armazenados no S3, facilitando a criação e manutenção de tabelas no Athena, permitindo que os usuários consultem os dados com SQL padrão sem a necessidade de configurar manualmente os esquemas destas tabelas.

## Configurações AWS Glue Catalog

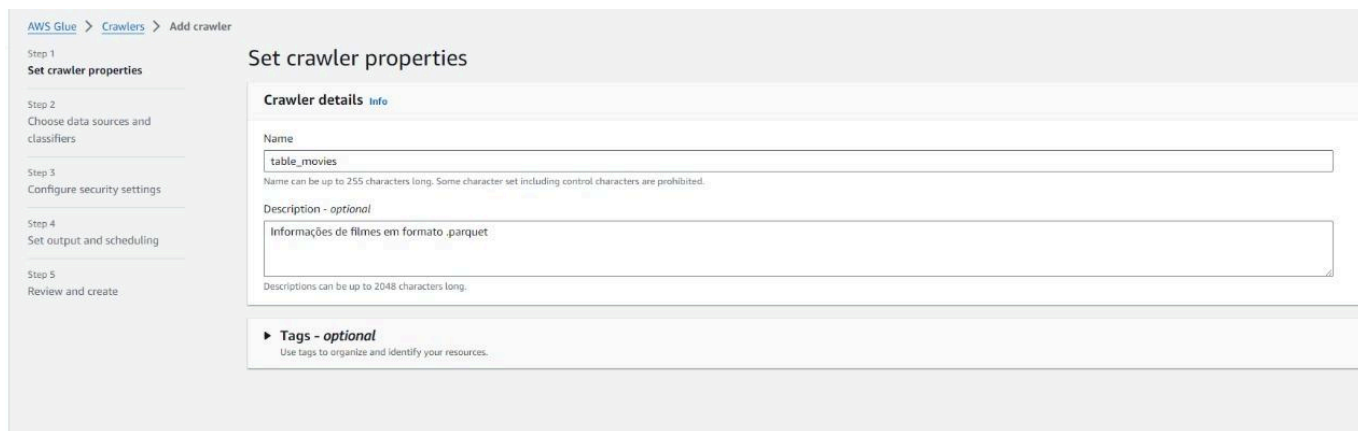
O primeiro passo é a criação e configuração do crawler no Glue para catalogar os dados e gerar a tabela.

1. Dentro do AWS Glue Catalog, acesse Data Catalog > Databases > Default e clique em Add tables using crawler:

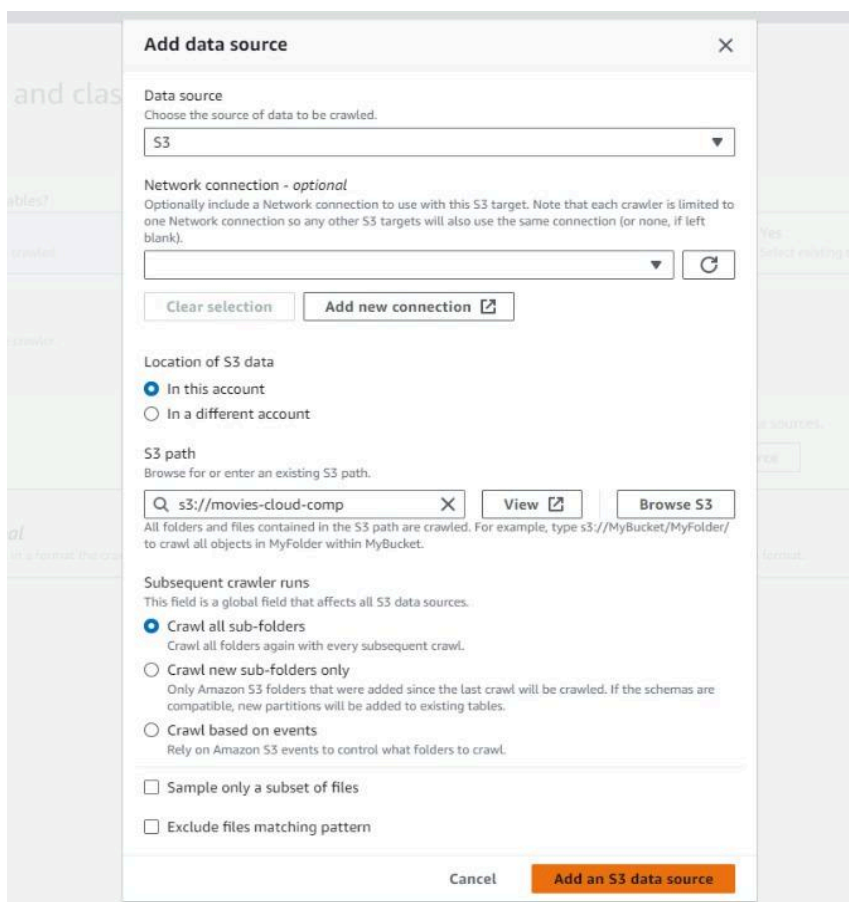


The screenshot shows the AWS Glue Catalog console. At the top, the breadcrumb navigation is 'AWS Glue > Databases > default'. Below this, the 'default' database is selected, showing its properties: Name (default), Description (-), Location (-), and Created on (UTC) (March 16, 2024 at 15:12:34). Under the 'Tables (1)' section, there is a table named 'your\_table\_name' with the following details: Database (default), Location (s3://escn-tame-bda-immersion/raw/2024/h), Classification (-), Deprecated (-), View data (Table data), and Data quality (View data quality). The console also includes buttons for 'Add tables using crawler' and 'Add table'.

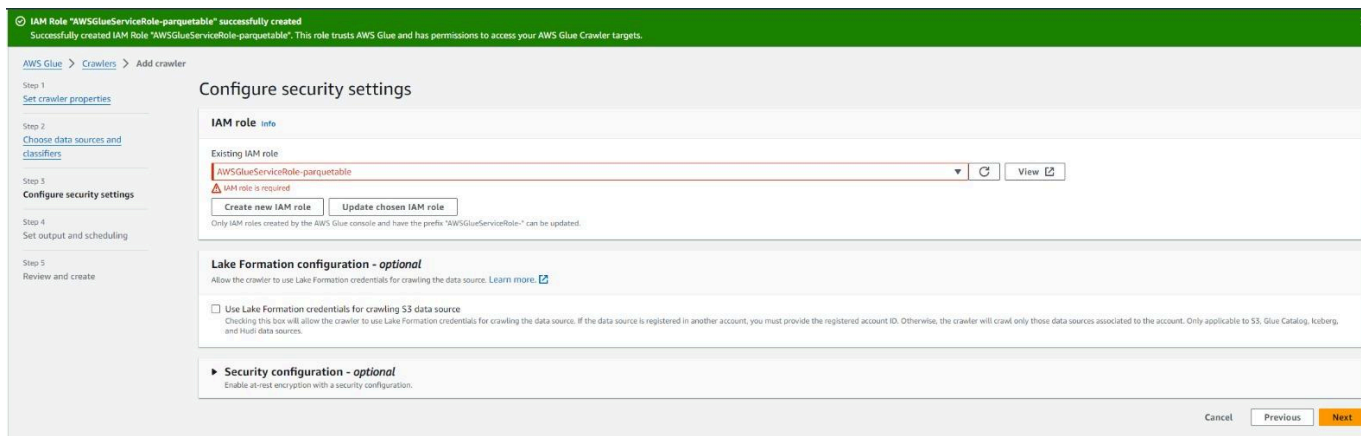
- No Step 1 das configurações do crawler, crie um nome, descrição (opcional) e clique em next:



- No Step 2, Selecione Not yet para Data source configuration, adicione o Data Source (Bucket S3) e clique em next:



- No Step 4, em IAM role, clique em Create new IAM role (AWSGlueServiceRole-parquetable foi o nome da role criada nesse tutorial) e clique em next:



**IAM Role "AWSGlueServiceRole-parquetable" successfully created**  
Successfully created IAM Role "AWSGlueServiceRole-parquetable". This role trusts AWS Glue and has permissions to access your AWS Glue Crawler targets.

**Configure security settings**

**IAM role info**

Existing IAM role: **AWSGlueServiceRole-parquetable** [View](#)

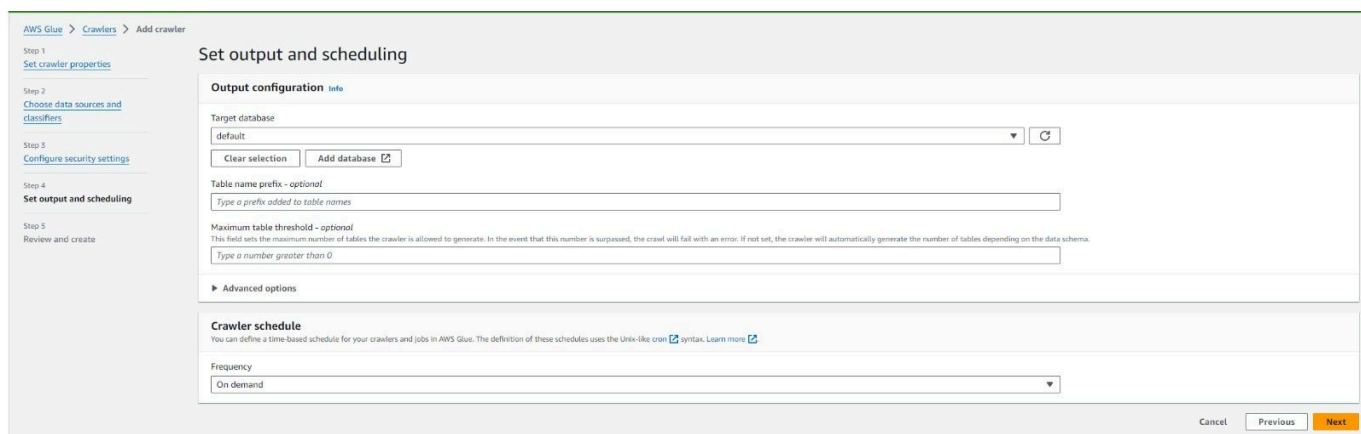
**Lake Formation configuration - optional**  
Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more](#)

☐ Use Lake Formation credentials for crawling S3 data source  
Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

**Security configuration - optional**  
Enable at-rest encryption with a security configuration.

[Cancel](#) [Previous](#) [Next](#)

- No Step 4, selecione o Target database como default, o Frequency do Crawler schedule como On demand e clique em next:



**Set output and scheduling**

**Output configuration info**

Target database: **default** [Add database](#)

Table name prefix - optional:  
Type a prefix added to table names

Maximum table threshold - optional  
This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.  
Type a number greater than 0

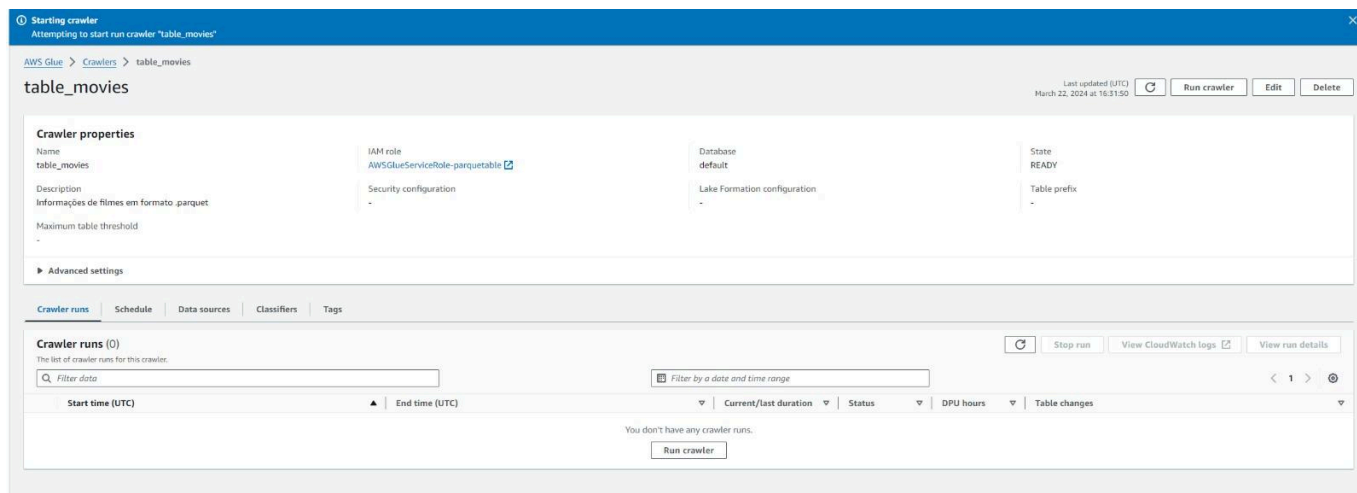
**Crawler schedule**  
You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency: **On demand**

[Cancel](#) [Previous](#) [Next](#)

- Revise as informações no Step 5 e clique em Create crawler.

- Em Data catalog > Crawlers é possível visualizar e editar o crawler criado (table\_movies). Clique em Run crawler para iniciar a catalogação dos dados armazenados no bucket S3:



**Starting crawler**  
Attempting to start run crawler "table\_movies"

[AWS Glue](#) > [Crawlers](#) > table\_movies

**table\_movies** Last updated (UTC) March 22, 2024 at 16:31:50 [Run crawler](#) [Edit](#) [Delete](#)

**Crawler properties**

Name table_movies	IAM role AWSGlueServiceRole-parquetable <a href="#">↗</a>	Database default	State READY
Description Informações de filmes em formato parquet	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

▶ **Advanced settings**

**Crawler runs** | **Schedule** | **Data sources** | **Classifiers** | **Tags**

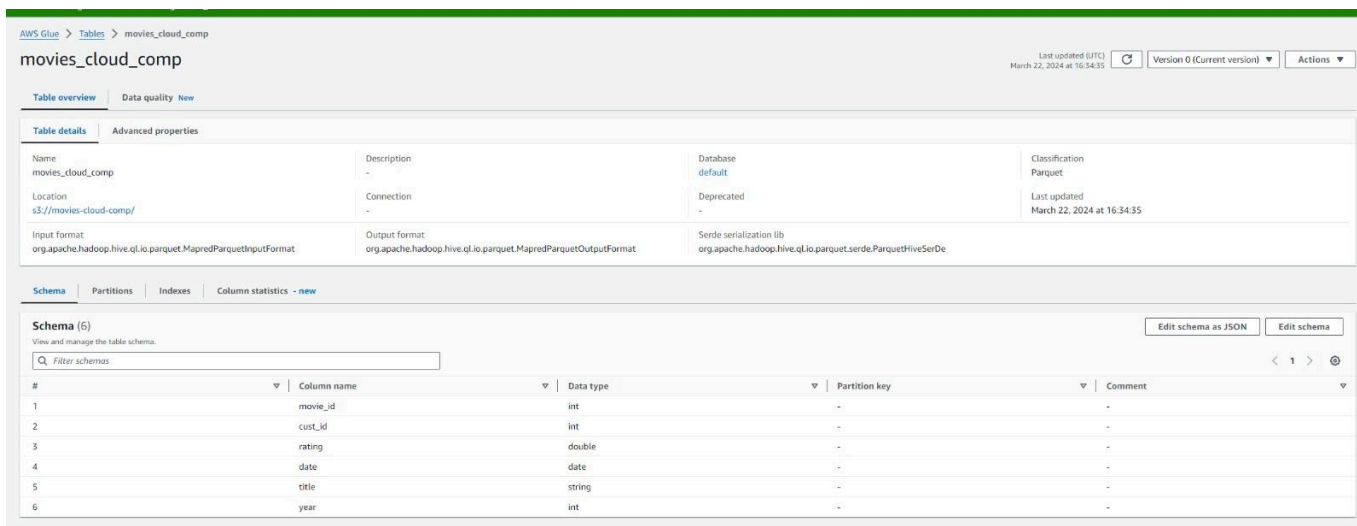
**Crawler runs (0)**  
The list of crawler runs for this crawler.

**Start time (UTC)** **End time (UTC)** **Current/last duration** **Status** **DPU hours** **Table changes**

You don't have any crawler runs.

[Run crawler](#)

- Em Data catalog > Databases > Tables é possível visualizar a tabela criada (movie\_cloud\_comp) a partir das informações dos dados armazenado no bucket S3:



[AWS Glue](#) > [Tables](#) > movies\_cloud\_comp

**movies\_cloud\_comp** Last updated (UTC) March 22, 2024 at 16:34:35 [Version 0 \(Current version\)](#) [Actions](#)

**Table overview** | **Data quality** [New](#)

**Table details** | **Advanced properties**

Name movies_cloud_comp	Description -	Database default	Classification Parquet
Location s3://movies-cloud-comp/	Connection -	Deprecated -	Last updated March 22, 2024 at 16:34:35
Input format org.apache.hadoop.hive.q1.io.parquet.MapredParquetInputFormat	Output format org.apache.hadoop.hive.q1.io.parquet.MapredParquetOutputFormat	Serde serialization lib org.apache.hadoop.hive.q1.io.parquet.serde.ParquetHiveSerDe	

**Schema** | **Partitions** | **Indexes** | **Column statistics** [- new](#)

**Schema (6)**  
View and manage the table schema.

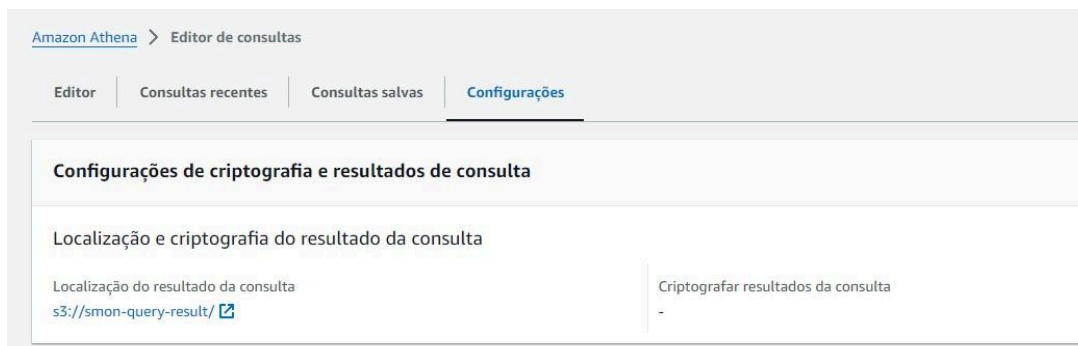
[Edit schema as JSON](#) [Edit schema](#)

#	Column name	Data type	Partition key	Comment
1	movie_id	int	-	-
2	cust_id	int	-	-
3	rating	double	-	-
4	date	date	-	-
5	title	string	-	-
6	year	int	-	-

## Configurações e consultas no Amazon Athena

Para começar a fazer consultas pelo Amazon Athena, primeiro é preciso configurar o bucket que irá armazenar os resultados dessas consultas.

1. Na página inicial do Athena, clique em Launch query editor > Settings e adicione o bucket de Query result location, onde serão armazenados os resultados das consultas realizadas (é preciso criar um bucket previamente no Amazon S3):



2. Em Query editor > Editor, configure o Data source para AwsDataCatalog e o Database para Default. Você logo verá a tabela criada pelo Glue na aba Tables. Clique nos três pontos da tabela movies\_cloud\_comp e clique em Generate table DDL. DDL (Data Definition Language) refere-se às instruções usadas para definir e manipular a estrutura de tabelas de dados. As instruções DDL são usadas para criar, alterar e excluir tabelas, bem como para definir suas propriedades e esquemas:


Concluído

```

CREATE EXTERNAL TABLE `movies_cloud_comp`(
  `movie_id` int,
  `cust_id` int,
  `rating` double,
  `date` date,
  `title` string,
  `year` int)
ROW FORMAT SERDE
  'org.apache.hadoop.hive ql.io.parquet.serde.ParquetHiveSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive ql.io.parquet.MapredParquetInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive ql.io.parquet.MapredParquetOutputFormat'
LOCATION
  's3://movies-cloud-comp/'
TBLPROPERTIES (
  'CrawlerSchemaDeserializerVersion'='1.0',
  'CrawlerSchemaSerializerVersion'='1.0',
  'UPDATED_BY_CRAWLER'='table_movies',
  'averageRecordSize'='5',
  'classification'='parquet',
  'compressionType'='none',
  'objectCount'='5',
  'recordCount'='24053764',
  'sizeKey'='141050881',
  'typeOfData'='file')

```

- O bucket dos dados com os arquivos *.parquet* que serão consumidos pelo Athena para realizar as consultas pode ser observado em no serviço da Amazon S3 > Buckets > movies\_cloud\_comp:

Amazon S3 > Buckets > movies-cloud-comp






**movies-cloud-comp** Informações

Objetos | Propriedades | Permissões | Métricas | Gerenciamento | Pontos de acesso

Objetos (5) Informações

Os objetos são as entidades fundamentais armazenadas no Amazon S3. Você pode usar o [inventário do Amazon S3](#) para obter uma lista de todos os objetos em seu bucket. Para outras pessoas acessarem seus objetos, você precisará conceder permissões explicitamente a eles. [Saiba mais](#)

Localizar objetos por prefixo


<input type="checkbox"/>	Nome	Tipo	Última modificação	Tamanho	Classe de armazenamento
<input type="checkbox"/>	 part-00000-89f5311a-ec3d-4078-9687-e5ffc1ec82a2-c000.snappy.parquet	parquet	21 Mar 2024 02:16:01 PM -03	28.3 MB	Padrão
<input type="checkbox"/>	 part-00001-89f5311a-ec3d-4078-9687-e5ffc1ec82a2-c000.snappy.parquet	parquet	21 Mar 2024 02:16:01 PM -03	27.3 MB	Padrão
<input type="checkbox"/>	 part-00002-89f5311a-ec3d-4078-9687-e5ffc1ec82a2-c000.snappy.parquet	parquet	21 Mar 2024 02:16:01 PM -03	27.3 MB	Padrão
<input type="checkbox"/>	 part-00003-89f5311a-ec3d-4078-9687-e5ffc1ec82a2-c000.snappy.parquet	parquet	21 Mar 2024 02:16:01 PM -03	27.1 MB	Padrão
<input type="checkbox"/>	 part-00004-89f5311a-ec3d-4078-9687-e5ffc1ec82a2-c000.snappy.parquet	parquet	21 Mar 2024 02:16:01 PM -03	24.5 MB	Padrão



## Respostas de negócio

Abaixo serão exibidas as consultas e SQL feitas no editor do Athena.

### 1.1 Quantos filmes estão disponíveis no dataset?


 **Consulta 5** :

1 SELECT COUNT(DISTINCT Movie\_Id) AS num\_filmes

2 FROM movies\_cloud\_comp;

SQL Ln 2, Col 24

Executar novamente

Explicar 


Cancelar

Limpar

Criar ▼

Resultados da consulta

Estatísticas da consulta

 Concluído

**Resultados (1)**

# ▼	num_filmes
1	4499

## 1.2 Qual é o nome dos 5 filmes com melhor média de avaliação?


Consulta 11 : X | Consulta 12 : X | Consulta 5 : X | **Consulta 6 : X** | Consulta 8 : X | Consulta 9 : X | Consulta 10 : X

```

1 -- 1. Seleciona as colunas 'Title' e as médias das avaliações da coluna 'Rating' da tabela 'movies_cloud_comp' e renomeia as médias para 'avg_rating';
2 -- 2. Agrupa por 'Title' e ordena por 'avg_rating'.
3
4 SELECT title, AVG(rating) AS avg_rating
5 FROM movies_cloud_comp
6 GROUP BY title
7 ORDER BY avg_rating DESC
8 LIMIT 5;

```

SQL Ln 1, Col 151

**Executar novamente** Explicar  Cancelar Limpar Criar ▼

**Resultados da consulta** Estatísticas da consulta

✔ Concluído

**Resultados (5)**

Q Linhas de pesquisa

#	title
1	Lost: Season 1
2	Ghost in the Shell: Stand Alone Complex: 2nd Gig
3	The Simpsons: Season 6
4	Inu-Yasha
5	Lord of the Rings: The Return of the King: Extended Edition: Bonus Material


## 1.3 Quais os 9 anos com menos lançamentos de filmes?

```

1 SELECT Year, COUNT(*) AS num_movies
2 FROM movies_cloud_comp
3 GROUP BY Year
4 ORDER BY num_movies
5 LIMIT 9;

```

SQL Ln 5, Col 9

**Executar novamente** Explicar  Cancelar Limpar Criar ▼

**Resultados da consulta** Estatísticas da consulta

✔ Concluído

**Resultados (9)**

Q Linhas de pesquisa

#	Year	num_movies
1	1926	107
2	1915	127
3	1917	138
4	1922	196
5	1928	498
6	1921	650
7	1918	832
8	1937	890
9	1948	959

1.4 Quantos filmes que possuem avaliação maior ou igual a 4.7, considerando apenas os filmes avaliados na última data de avaliação do dataset?

Consulta 5

Consulta 6

Consulta 8

Consulta 9

Consulta 10

```

1 -- 1. Faz a contagem do título de filmes únicos (num_movies) da tabela 'movies_cloud_comp', onde os dados da tabela estão sob
2 -- a condição (WHERE) de possuir avaliação na última data da coluna 'date' e (AND) possuir 'Rating' maior que 4.7.
3
4 SELECT COUNT(DISTINCT Movie_Id) AS num_movies
5 FROM movies_cloud_comp
6 WHERE Date = (SELECT MAX(Date) FROM movies_cloud_comp)
7 AND Rating >= 4.7;

```

SQL Ln 1, Col 126

Executar

Explicar

Cancelar

Limpar

Criar

Resultados da consulta

Estatísticas da consulta

Concluído

Resultados (1)

Linhas de pesquisa

#	num_movies
1	780

1.5 Quais os id's dos 5 customers que mais avaliaram filmes e quantas avaliações cada um fez?

Consulta 5

Consulta 6

Consulta 8

Consulta 9

Consulta 10

```

1 -- 1. Faz a contagem do nº de avaliações feitas (num_reviews) por cada id único de usuário (Cust_id)
2 -- presente na tabela e ordena por 'num_reviews'.
3
4 SELECT Cust_Id, COUNT(*) AS num_reviews
5 FROM movies_cloud_comp
6 GROUP BY Cust_Id
7 ORDER BY num_reviews DESC
8 LIMIT 5;

```

SQL Ln 1, Col 102

Executar

Explicar

Cancelar

Limpar

Criar

Resultados da consulta

Estatísticas da consulta

Concluído

Resultados (5)

Linhas de pesquisa

#	Cust_Id	num_reviews
1	305344	4467
2	387418	4422
3	2439493	4195
4	1664010	4019
5	2118461	3769