

# Netflix Database

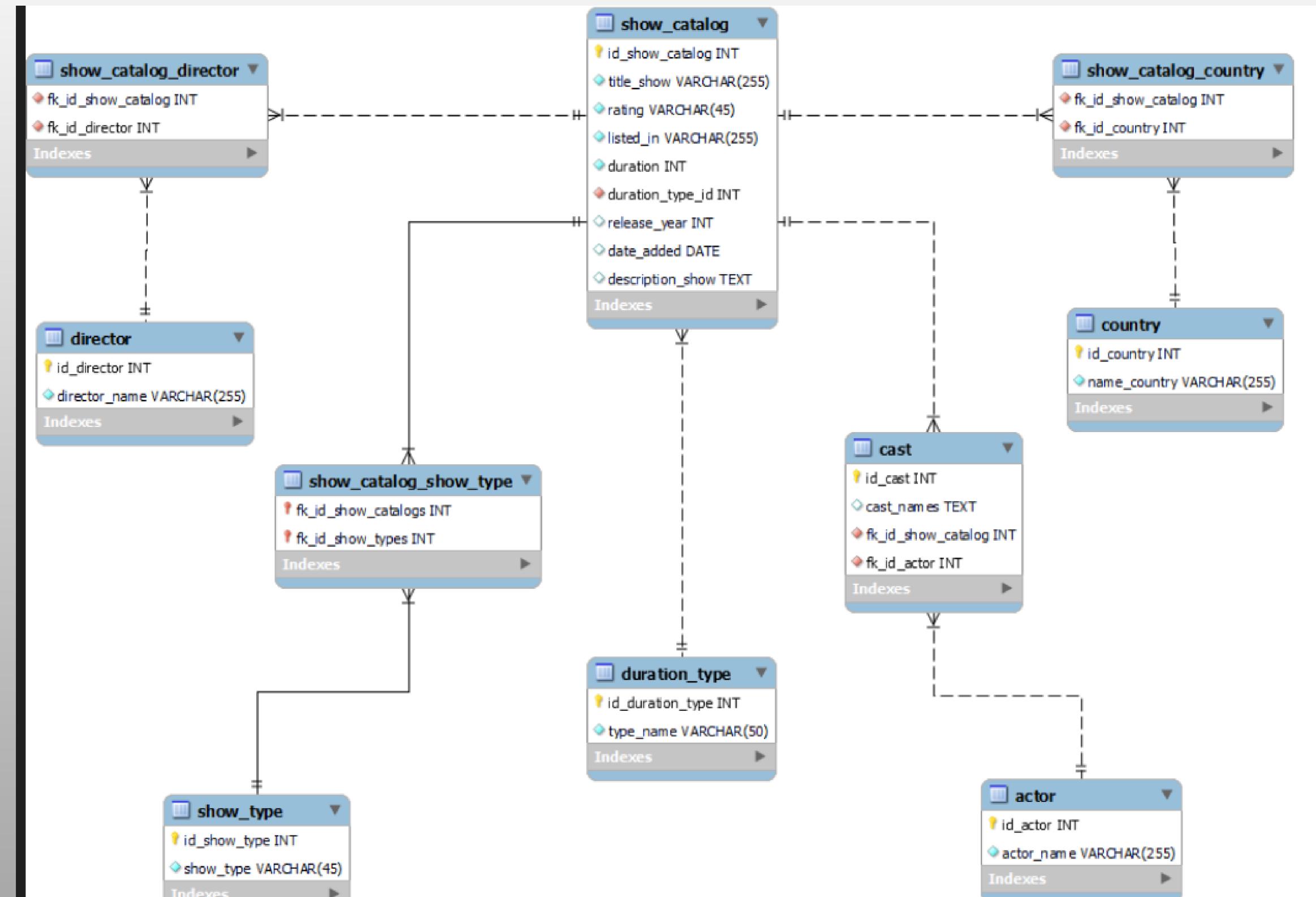
O data set escolhido por nosso grupo foi o Netflix Database, que contém os dados de filmes e séries da Netflix até o ano de 2021. Nele, estão todas as informações necessárias para qualquer busca de pesquisa. Contém dados de diretores, elenco, cidades e países onde o filme foi gravado, duração dos filmes e séries, entre outros dados.

	title	original_title	year	rating	genre	plot	imdb_id
3	s2,TV Show,Blood & Water,,,"Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, F						
4	s3,TV Show,Ganglands,Julien Leclercq,"Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farihi, Geert Van Rampelberg, Bakary Diombe						
5	s4,TV Show,Jailbirds New Orleans,,,,"September 24, 2021",2021,TV-MA,1 Season,"Docuseries, Reality TV","Feuds, flirtations and toilet talk go down among the incarcerated women at the O						
6	s5,TV Show,Kota Factory,,,"Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, Revathi Pillai, Urvi Singh, Arun Kumar",India,"September 24, 2021",2021,TV-MA,2 Seasons,"Int						
7	s6,TV Show,Midnight Mass,Mike Flanagan,"Kate Siegel, Zach Gilford, Hamish Linklater, Henry Thomas, Kristin Lehman, Samantha Sloyan, Igby Rigney, Rahul Kohli, Annarah Cymone, Annabeth C						
8	s7,Movie,My Little Pony: A New Generation,"Robert Cullen, José Luis Ucha",Vanessa Hudgens, Kimiko Glenn, James Marsden, Sofia Carson, Liza Koshy, Ken Jeong, Elizabeth Perkins, Jane Ki						
9	s8,Movie,Sankofa,Haile Gerima,"Kofi Ghanaba, Oyafunmike Ogunlana, Alexandra Duah, Nick Medley, Mutabaruka, Afemo Omilami, Reggie Carter, Mzuri",United States, Ghana, Burkina Faso, Un						
10	s9,TV Show,The Great British Baking Show,Andy Devonshire,"Mel Giedroyc, Sue Perkins, Mary Berry, Paul Hollywood",United Kingdom,"September 24, 2021",2021,TV-14,9 Seasons,"British TV Sh						
11	s10,Movie,The Starling,Theodore Melfi,"Melissa McCarthy, Chris O'Dowd, Kevin Kline, Timothy Olyphant, Daveed Diggs, Skyler Gisondo, Laura Harrier, Rosalind Chao, Kimberly Quinn, Loret						
12	s11,TV Show,Vendetta: Truth, Lies and The Mafia,,,,"September 24, 2021",2021,TV-MA,1 Season,"Crime TV Shows, Docuseries, International TV Shows","Sicily boasts a bold ""Anti-Mafia""						
13	s12,TV Show,Bangkok Breaking,Kongkiat Komesiri,"Sukollawat Kanarot, Sushar Manaying, Pavarit Mongkolpisit, Sahajak Boonthanakit, Suthipongse Thatphithakkul, Bhasaworn Bawronkirati, Da						
14	s13,Movie,Je Suis Karl,Christian Schwochow,"Luna Wedler, Jannis Niewöhner, Milan Peschel, Edin Hasanović, Anna Fialová, Marlon Boess, Victor Boccard, Fleur Geffrier, Aziz Dyab, Mélanie						
15	s14,Movie,Confessions of an Invisible Girl,Bruno Garotti,"Klara Castanho, Lucca Picon, Júlia Gomes, Marcus Bessa, Kiria Malheiros, Fernanda Concon, Gabriel Lima, Caio Cabral, Leonardo						
16	s15,TV Show,Crime Stories: India Detectives,,,,"September 22, 2021",2021,TV-MA,1 Season,"British TV Shows, Crime TV Shows, Docuseries",Camaras following Bengaluru police on the job of						
17	s16,TV Show,Dear White People,,,"Logan Browning, Brandon P. Bell, DeRon Horton, Antoinette Robertson, John Patrick Amedori, Ashley Blaine Featherson, Marque Richardson, Giancarlo Espos						
18	s17,Movie,Europe's Most Dangerous Man: Otto Skorzeny in Spain,"Pedro de Echave Garcia, Pablo Azorin Williams",,"September 22, 2021",2020,TV-MA,67 min,"Documentaries, International Mo						
19	s18,TV Show,Falsa identidad,,,"Luis Ernesto Franco, Camila Sodi, Sergio Goyri, Samadhi Zendejas, Eduardo Yáñez, Sonya Smith, Alejandro Camacho, Azela Robinson, Uriel del Toro, Géraldin						
20	s19,Movie,Intrusion,Adam Salky,"Freida Pinto, Logan Marshall-Green, Robert John Burke, Megan Elisabeth Kelly, Sarah Minnich, Hayes Hargrove, Mark Sivertsen, Brandon Fierro, Antonio Val						
21	s20,TV Show,Jaguar,,,"Blanca Suárez, Iván Marcos, Óscar Casas, Adrián Lastra, Francesc Garrido, Stefan Weinert, Julia Möller, Alicia Chojnowski",,"September 22, 2021",2021,TV-MA,1 Season						
22	s21,TV Show,Monsters Inside: The 24 Faces of Billy Milligan,Olivier Megaton,,,,"September 22, 2021",2021,TV-14,1 Season,"Crime TV Shows, Docuseries, International TV Shows","In the late						
23	s22,TV Show,Resurrection: Ertugrul,,,"Engin Altan Düzyatan, Serdar Gökan, Hülya Darcan, Kaan Taşaner, Esra Bilgiç, Osman Soykut, Serdar Deniz, Cengiz Coşkun, Reshad Strik, Hande Subaş						
24	s23,Movie,Avvai Shanmugi,K.S. Ravikumar,"Kamal Hassan, Meena, Gemini Ganesan, Heera Rajgopal, Nassar, S.P. Balasubrahmanyam",,"September 21, 2021",1996,TV-PG,161 min,"Comedies, Intern						
25	s24,Movie,Go! Go! Cory Carson: Chrissy Takes the Wheel,"Alex Woo, Stanley Moore",Maisie Benson, Paul Killam, Kerry Godliman, AC Lim",,"September 21, 2021",2021,TV-Y,61 min,Children						
26	s25,Movie,Jeans,S. Shankar,"Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi, Nassar",India,"September 21, 2021",1998,TV-14,166 min,"Comedies, International Movies, Romantic Movies","Who						
27	s26,TV Show,Love on the Spectrum,,Brooke Satchwell,Australia,"September 21, 2021",2021,TV-14,2 Seasons,"Docuseries, International TV Shows, Reality TV","Finding love can be hard for a						
28	s27,Movie,Minsara Kanavu,Rajiv Menon,"Arvind Swamy, Kajol, Prabhu Deva, Nassar, S.P. Balasubrahmanyam, Girish Karnad",,"September 21, 2021",1997,TV-PG,147 min,"Comedies, International						
29	s28,Movie,Grown Ups,Dennis Dugan,"Adam Sandler, Kevin James, Chris Rock, David Spade, Rob Schneider, Salma Hayek, Maria Bello, Maya Rudolph, Colin Quinn, Tim Meadows, Joyce Van Patten						
30	s29,Movie,Dark Skies,Scott Stewart,"Keri Russell, Josh Hamilton, J.K. Simmons, Dakota Goyo, Kadan Rockett, L.J. Benét, Rich Hutchman, Myndy Crist, Annie Thurman, Jake Brennan",United						
31	s30,Movie,Paranoia,Robert Luketic,"Liam Hemsworth, Gary Oldman, Amber Heard, Harrison Ford, Lucas Till, Embeth Davidtz, Julian McMahon, Josh Holloway, Richard Dreyfuss, Angela Sarafyan						
32	s31,Movie,Ankahi Kahaniya,"Ashwini Iyer Tiwari, Abhishek Chaubey, Saket Chaudhary",Abhishek Banerjee, Rinku Rajguru, Delzad Hiwale, Kunal Kapoor, Zoya Hussain, Nikhil Dwivedi, Palomi						
33	s32,TV Show,Chicago Party Aunt,,,"Lauren Ash, Rory O'Malley, RuPaul Charles, Jill Talley, Ike Barinholtz, Jon Barinholtz, Matthew Craig, Bob Odenkirk, Mike Hagerty, Katie Rich, Chris W						
34	s33,TV Show,Sex Education,,,"Asa Butterfield, Gillian Anderson, Ncuti Gatwa, Emma Mackey, Connor Swindells, Kedar Williams-Stirling, Alistair Petrie",United Kingdom,"September 17, 2021",2020,TV-14,1 Season						
35	s34,TV Show,Squid Game,,,"Lee Jung-jae, Park Hae-soo, Wi Ha-jun, Oh Young-soo, Jung Ho-yeon, Heo Sung-tae, Kim Joo-ryoung, Tripathi Anupam, You Seong-joo, Lee You-mi",,"September 17, 2021",2020,TV-Y,61 min,Children						
36	s35,TV Show,Tayo and Little Wizards,,,"Dami Lee, Jason Lee, Bonnie Catherine Han, Jennifer Waescher, Nancy Kim",,"September 17, 2021",2020,TV-Y,1 Season,Kids' TV,Tayo speeds into an ad						
37	s36,Movie,The Father Who Moves Mountains,Daniel Sandu,"Adrian Titieni, Elena Purea, Judith State, Valeriu Andriută, Tudor Smoleanu, Virgil Aioanei, Radu Botar, Petronela Grigorescu, Bo						

<https://www.kaggle.com/datasets/syedmubarak/netflix-dataset-latest-2021>

# Modelo Lógico

Com esse conjunto de dados,  
montamos nosso modelo lógico já  
normalizado.



# Pré-tratamento de dados

Com o modelo lógico em mãos, pré-tratamos os dados.

```
1 """
2 a coluna duration tinha numeros e letras dentro dela e era dividida em minutos e seasons.
3 tiramos todos os caracteres que nao fossem numericos de dentro dela e transformamos ela para INT
4
5 df['duration'] = df['duration'].replace(r'\D', '', regex=True)
6
7 df['duration'] = df['duration'].replace('', pd.NA)
8
9 df['duration'] = df['duration'].fillna(0)
10
11 df['duration'] = df['duration'].astype(int)
12
13 """
```

```
1 """
2 a coluna "date_added" estava no formato "August 4, 2017" quando deveria estar nesse %B, %d, %Y para ser um do tipo DATE.
3
4
5 meses_numeros = {
6     'January': 1, 'February': 2, 'March': 3, 'April': 4,
7     'May': 5, 'June': 6, 'July': 7, 'August': 8,
8     'September': 9, 'October': 10, 'November': 11, 'December': 12
9 }
10
11
12 def converter(date_str):
13     if pd.isnull(date_str):
14         return None
15
16     partes = date_str.split()
17
18     mes_nome = partes[0]
19     dia = int(partes[1][:-1])
20
21     ano = partes[2]
22
23     mes_numero = meses_numeros[mes_nome]
24
25     data_formatada = f'{mes_numero:02d}/{dia:02d}/{ano}'
26
27     return data_formatada
28
29 df['date_added'] = df['date_added'].apply(converter) # chamando a função
30
31 df['date_added'] = pd.to_datetime(df['date_added']) # converter para o tipo DATE
32
33
34 """
```

```
"""
1 Coluna "release_year" estava como DATE na V_1.0 do modelo físico, foi substituído para INT.
2
3 release_year INT NOT NULL
4
5 """
```

```
1 """
2 A coluna "show_id" tinha um "s" antes de todo o numero do ID. Então o primeiro passo foi retirar esse "s" e deixar apenas o número para mudarmos o tipo da coluna para INT.
3
4 #df['show_id'] = df['show_id'].str.replace('s', '') # tirar o s
5 #df['show_id'] = df['show_id'].astype(int) # mudar o tipo da coluna para INT
6
7 """
```

```
1 """
2 A coluna "title" que tinha OBJECT como tipo, foi substituído para STRING.
3
4 df['title'] = df['title'].astype(str) # mudar o tipo da coluna para string
5
6 """
```

```
1 """
2 Mudança de nome na coluna TYPE para TYPE_SHOW, porque "type" é uma palavra reservada do mysql, além de mudar seu tipo para STRING.
3
4 df.rename(columns={'type': 'type_show'}, inplace=True) # mudar nome da coluna
5 df['type_show'] = df['type_show'].astype(str) # mudar tipo da coluna para STRING
6
7
8 """
```

# Dados tratados

# Inserção de dados

```
conexao_mysql.py Ingestao X
💡 Click here to ask Blackbox to help you code faster
1 import pandas as pd
2 import mysql.connector
3
4 arq = 'Arquivo CSV/Arquivo Refined/netflix_refined.csv'
5 df = pd.read_csv(arq)
6
7 config = {
8     'user': 'root',
9     'password': 'root',
10    'host': 'localhost',
11    'port': 3306,
12    'database': 'ufc'
13 }
14
15 con = mysql.connector.connect(**config)
16 cursor = con.cursor()
17
18 for index, row in df.iterrows():
19     values = [None if pd.isna(value) else value for value in row.values] # coloca NONE no lugar NAN
20
21     sql = """
22     INSERT INTO show_temporaria
23         (show_id, type_show, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description_show)
24     VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s)
25     """
26
27     cursor.execute(sql, values)
28     con.commit() # da commit na transação toda vez que insere um dado
29
30 cursor.close()
31 con.close()
```



```
show_temporaria.sql Tabela_temporaria X
💡 Click here to ask Blackbox to help you code faster
1
2 CREATE TABLE show_temporaria (
3     show_id VARCHAR(255) PRIMARY KEY,
4     type_show VARCHAR(100),
5     title VARCHAR(255),
6     director VARCHAR(255),
7     cast TEXT,
8     country VARCHAR(255),
9     date_added DATE,
10    release_year INT,
11    rating VARCHAR(50),
12    duration VARCHAR(100),
13    listed_in TEXT,
14    description_show TEXT,
15    INDEX (release_year),
16    INDEX (type_show),
17    INDEX (rating)
18 );
19
```

# Criação de tabelas

```
scripts_tabelas.sql DDLs X
💡 Click here to ask Blackbox to help you code faster
1 CREATE DATABASE IF NOT EXISTS ufc;
2 USE ufc;
3
4
5 CREATE TABLE IF NOT EXISTS duration_type (
6     id_duration_type INT NOT NULL AUTO_INCREMENT,
7     type_name VARCHAR(50) NOT NULL,
8     PRIMARY KEY (id_duration_type)
9 );
10
11 CREATE TABLE IF NOT EXISTS show_catalog (
12     id_show_catalog INT NOT NULL AUTO_INCREMENT,
13     title_show VARCHAR(255) NOT NULL,
14     rating VARCHAR(45) NOT NULL,
15     listed_in VARCHAR(255) NOT NULL,
16     duration INT NOT NULL,
17     duration_type_id INT NOT NULL,
18     release_year INT,
19     date_added DATE,
20     description_show TEXT,
21     PRIMARY KEY (id_show_catalog),
22     FOREIGN KEY (duration_type_id) REFERENCES duration_type(id_duration_type)
23 );
24
25
26
27
28 CREATE TABLE IF NOT EXISTS actor (
29     id_actor INT NOT NULL AUTO_INCREMENT,
30     actor_name VARCHAR(255) NOT NULL,
31     PRIMARY KEY (id_actor)
32 );
33
```

```
35 CREATE TABLE IF NOT EXISTS cast (
36     id_cast INT NOT NULL AUTO_INCREMENT,
37     cast_names TEXT,
38     fk_id_show_catalog INT NOT NULL,
39     fk_id_actor INT NOT NULL,
40     PRIMARY KEY (id_cast),
41     INDEX fk_id_actor_idx (fk_id_actor ASC),
42     INDEX fk_id_show_catalog_idx (fk_id_show_catalog ASC),
43     CONSTRAINT fk_id_show_catalog
44         FOREIGN KEY (fk_id_show_catalog)
45             REFERENCES show_catalog (id_show_catalog)
46             ON DELETE NO ACTION
47             ON UPDATE NO ACTION,
48     CONSTRAINT fk_id_actor
49         FOREIGN KEY (fk_id_actor)
50             REFERENCES actor (id_actor)
51             ON DELETE NO ACTION
52             ON UPDATE NO ACTION
53 );
54
55
56 CREATE TABLE IF NOT EXISTS country (
57     id_country INT NOT NULL AUTO_INCREMENT,
58     name_country VARCHAR(255) NOT NULL,
59     PRIMARY KEY (id_country)
60 );
61
62
63 CREATE TABLE IF NOT EXISTS director (
64     id_director INT NOT NULL AUTO_INCREMENT,
65     director_name VARCHAR(255) NOT NULL,
```

# Criação de tabelas

```
71 CREATE TABLE IF NOT EXISTS show_catalog_country (
72     fk_id_show_catalog INT NOT NULL,
73     fk_id_country INT NOT NULL,
74     INDEX fk_id_country_idx (fk_id_country ASC),
75     CONSTRAINT fk_id_country_catalog
76         FOREIGN KEY (fk_id_country)
77             REFERENCES country (id_country),
78     CONSTRAINT fk_id_show_catalog_country
79         FOREIGN KEY (fk_id_show_catalog)
80             REFERENCES show_catalog (id_show_catalog)
81 );
82
83
84 CREATE TABLE IF NOT EXISTS show_catalog_director (
85     fk_id_show_catalog INT NOT NULL,
86     fk_id_director INT NOT NULL,
87     INDEX fk_id_director_idx (fk_id_director ASC),
88     CONSTRAINT fk_id_director_show_catalog
89         FOREIGN KEY (fk_id_director)
90             REFERENCES director (id_director),
91     CONSTRAINT fk_id_show_catalog_director
92         FOREIGN KEY (fk_id_show_catalog)
93             REFERENCES show_catalog (id_show_catalog)
94 );
95
96
97 CREATE TABLE IF NOT EXISTS show_type (
98     id_show_type INT NOT NULL AUTO_INCREMENT,
99     show_type VARCHAR(45) NOT NULL,
100    PRIMARY KEY (id_show_type)
101 );
```

```
104 CREATE TABLE IF NOT EXISTS show_catalog_show_type (
105     fk_id_show_catalogs INT NOT NULL,
106     fk_id_show_types INT NOT NULL,
107     CONSTRAINT fk_id_show_catalogs
108         FOREIGN KEY (fk_id_show_catalogs)
109             REFERENCES show_catalog (id_show_catalog),
110     CONSTRAINT fk_id_show_types
111         FOREIGN KEY (fk_id_show_types)
112             REFERENCES show_type (id_show_type),
113     PRIMARY KEY (fk_id_show_catalogs, fk_id_show_types)
114 );
115
```

# Inserir os dados nas tabelas através da tabela temporária.

```
scripts_tabelas.sql DDLs | scripts_inserts.sql DMLs X
Click here to ask Blackbox to help you code faster

1 INSERT INTO duration_type (type_name) VALUES
2 ('minutes'),
3 ('seasons');

4

5

6 INSERT INTO ufc.show_catalog (title_show, rating, listed_in, duration, duration_type_id, release_year, date_added, description_show)
7 SELECT st.title,
8       st.rating,
9       st.listed_in,
10      st.duration,
11      CASE
12        WHEN st.type_show = 'Movie' THEN 1
13        WHEN st.type_show = 'TV Show' THEN 2
14        ELSE NULL
15      END AS duration_type_id,
16      st.release_year,
17      st.date_added,
18      st.description_show
19 FROM ufc.show_temporaria st;
20
21
22
23 INSERT INTO ufc.actor (actor_name)
24 SELECT DISTINCT TRIM(REGEXP_SUBSTR(cast, '[^,]+', 1, n)) AS actor_name
25 FROM ufc.show_temporaria
26 JOIN (
27   SELECT 1 AS n UNION ALL SELECT 2 UNION ALL SELECT 3 UNION ALL SELECT 4
28   UNION ALL SELECT 5 UNION ALL SELECT 6 UNION ALL SELECT 7 UNION ALL SELECT 8
29   UNION ALL SELECT 9 UNION ALL SELECT 10
30 ) AS numbers
31 ON CHAR_LENGTH(cast) - CHAR_LENGTH(REPLACE(cast, ',', '')) ≥ n - 1;
32
33
```

# Inserir os dados nas tabelas através da tabela temporária.

```
34
35
36 INSERT INTO cast (cast_names, fk_id_show_catalog, fk_id_actor)
37 SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(st.cast, ',', n.digit+1), ',', -1)) AS actor_name,
38       sc.id_show_catalog,
39       a.id_actor
40 FROM show_temporaria st
41 JOIN
42 (
43     SELECT 0 AS digit UNION ALL SELECT 1 UNION ALL SELECT 2 UNION ALL SELECT 3 UNION ALL SELECT 4
44 ) n
45 ON LENGTH(REPLACE(st.cast, ',', '')) ≤ LENGTH(st.cast)-n.digit
46 JOIN actor a ON TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(st.cast, ',', n.digit+1), ',', -1)) = a.actor_name
47 JOIN show_catalog sc ON st.title = sc.title_show;
48
49
50
51 INSERT INTO ufc.country (name_country)
52 SELECT DISTINCT country
53 FROM ufc.show_temporaria;
54
55
56 INSERT INTO ufc.director (director_name)
57 SELECT DISTINCT director
58 FROM ufc.show_temporaria;
59
60
61
62
63 INSERT INTO ufc.show_type (show_type) VALUES
64 ('movie'),
65 ('TV series');
```

```
67
68 INSERT INTO ufc.show_catalog_country (fk_id_show_catalog, fk_id_country)
69 SELECT sc.id_show_catalog, c.id_country
70 FROM ufc.show_catalog sc
71 JOIN ufc.show_temporaria st ON sc.title_show = st.title
72 JOIN ufc.country c ON st.country = c.name_country;
73
74
75
76 INSERT INTO ufc.show_catalog_director (fk_id_show_catalog, fk_id_director)
77 SELECT sc.show_id, d.id_director
78 FROM ufc.show_temporaria sc
79 JOIN ufc.director d ON sc.director = d.director_name;
80
81
82 INSERT INTO ufc.show_catalog_show_type (fk_id_show_catalogs, fk_id_show_types)
83 SELECT sc.show_id, st.id_show_type
84 FROM ufc.show_temporaria sc
85 JOIN ufc.show_type st ON sc.type_show = st.show_type;
86
```

# Extração dos dados

Através das tabelas normalizadas e já populadas, podemos fazer a extração, começando pela desnormalização. Pegamos as tabelas normalizadas e as transformamos na tabela temporária, realizando uma engenharia reversa para criar a tabela 'show\_extract'.

```
💡 Click here to ask Blackbox to help you code faster
1 ALTER TABLE show_catalog
2 ADD COLUMN type_show VARCHAR(100);      -- precisa ser feito para evitar valores nulos na coluna type_show
3 UPDATE show_catalog sc
4 JOIN show_type st ON sc.duration_type_id = st.id_show_type
5 SET sc.type_show = st.show_type;
6
7
8 INSERT IGNORE INTO show_extract (      -- em versões anteriores precisava ignorar para evitar valores duplicados
9     show_id,
10    type_show,
11    title,
12    director,
13    cast,
14    country,
15    date_added,
16    release_year,
17    rating,
18    duration,
19    listed_in,
20    description_show
21 )
22 SELECT
23     sc.id_show_catalog AS show_id,
24     sc.type_show AS type_show,
25     sc.title_show AS title,
26     GROUP_CONCAT(DISTINCT d.director_name) AS director,
27     GROUP_CONCAT(DISTINCT a.actor_name) AS cast,
28     GROUP_CONCAT(DISTINCT c.name_country) AS country,
29     sc.date_added AS date_added,
30     sc.release_year AS release_year,
31     sc.rating AS rating,
32     CONCAT(sc.duration, ' ', dt.type_name) AS duration,
33     sc.listed_in AS listed_in,
34     sc.description_show AS description_show
```

```
35   FROM
36     show_catalog sc
37   LEFT JOIN
38     show_catalog_director scd ON sc.id_show_catalog = scd.fk_id_show_catalog
39   LEFT JOIN
40     director d ON scd.fk_id_director = d.id_director
41   LEFT JOIN
42     cast ca ON sc.id_show_catalog = ca.fk_id_show_catalog
43   LEFT JOIN
44     actor a ON ca.fk_id_actor = a.id_actor
45   LEFT JOIN
46     show_catalog_country scc ON sc.id_show_catalog = scc.fk_id_show_catalog
47   LEFT JOIN
48     country c ON scc.fk_id_country = c.id_country
49   LEFT JOIN
50     duration_type dt ON sc.duration_type_id = dt.id_duration_type
51   GROUP BY
52     sc.id_show_catalog, sc.title_show, sc.date_added, sc.release_year, sc.rating, sc.listed_in, sc.description_show;
53
```

# Extração dos dados

**Pegamos os dados das tabelas normalizadas e os inserimos na tabela 'show\_extract' no formato que devem ficar.**

```
💡 Click here to ask Blackbox to help you code faster
1 CREATE TABLE show_extract (
2     show_id VARCHAR(255) PRIMARY KEY,
3     type_show VARCHAR(100),
4     title VARCHAR(255),
5     director VARCHAR(255),
6     cast TEXT,
7     country VARCHAR(255),
8     date_added DATE,
9     release_year INT,
10    rating VARCHAR(50),
11    duration VARCHAR(100),
12    listed_in TEXT,
13    description_show TEXT,
14    INDEX (release_year),
15    INDEX (type_show),
16    INDEX (rating)
17 );
18
19 -- mesma estrutura da tabela show_temporaria no inicio
```

# Extração dos dados

Com os dados inseridos na tabela 'show\_extract', agora podemos extrair esses dados como SQL.

Com este arquivo, definimos o campo de pasta que será acessado e definimos a query SQL que irá extrair todos os dados e trazê-los para um arquivo CSV.

```
1 ALTER TABLE show_catalog
2 ADD COLUMN type_show VARCHAR(100);      -- precisa ser feito para evitar valores nulos na coluna type_show
3 UPDATE show_catalog sc
4 JOIN show_type st ON sc.duration_type_id = st.id_show_type
5 SET sc.type_show = st.show_type;
6
7
8 INSERT IGNORE INTO show_extract (      -- em versões anteriores precisava ignorar para evitar valores duplicados
9     show_id,
10    type_show,
11    title,
12    director,
13    cast,
14    country,
15    date_added,
16    release_year,
17    rating,
18    duration,
19    listed_in,
20    description_show
21 )
22 SELECT
23     sc.id_show_catalog AS show_id,
24     sc.type_show AS type_show,
25     sc.title_show AS title,
26     GROUP_CONCAT(DISTINCT d.director_name) AS director,
27     GROUP_CONCAT(DISTINCT a.actor_name) AS cast,
28     GROUP_CONCAT(DISTINCT c.name_country) AS country,
29     sc.date_added AS date_added,
30     sc.release_year AS release_year,
31     sc.rating AS rating,
32     CONCAT(sc.duration, ' ', dt.type_name) AS duration,
33     sc.listed_in AS listed_in,
34     sc.description_show AS description_show
35 FROM
36     show_catalog sc
37 LEFT JOIN
38     show_catalog_director scd ON sc.id_show_catalog = scd.fk_id_show_catalog
39 LEFT JOIN
40     director d ON scd.fk_id_director = d.id_director
41 LEFT JOIN
42     cast ca ON sc.id_show_catalog = ca.fk_id_show_catalog
43 LEFT JOIN
44     actor a ON ca.fk_id_actor = a.id_actor
45 LEFT JOIN
46     show_catalog_country scc ON sc.id_show_catalog = scc.fk_id_show_catalog
47 LEFT JOIN
48     country c ON scc.fk_id_country = c.id_country
49 LEFT JOIN
50     duration_type dt ON sc.duration_type_id = dt.id_duration_type
51 GROUP BY
52     sc.id_show_catalog, sc.title_show, sc.date_added, sc.release_year, sc.rating, sc.listed_in, sc.description_show;
```

# Ingestão no Neo4j

```
💡 Click here to ask Blackbox to help you code faster
1 CREATE INDEX FOR (a:Actor) ON (a.actor_name);      //criar os index antes fez o tempo total de criacao cair de 17 min para 5 min
2
3 CREATE INDEX FOR (c:Country) ON (c.name_country);
4
5 CREATE INDEX FOR (d:Director) ON (d.director_name);
6
7
8 LOAD CSV WITH HEADERS FROM 'file:///show_extract.csv' AS row
9 CREATE (:ShowCatalog {
10   id: toInteger(row.show_id),
11   type_show: row.type_show,
12   title: row.title,
13   director: row.director,
14   cast: row.cast,
15   country: row.country,
16   date_added: row.date_added,
17   release_year: toInteger(row.release_year),
18   rating: row.rating,
19   duration: row.duration,
20   listed_in: row.listed_in,
21   description_show: row.description_show
22 });
23
24
25 LOAD CSV WITH HEADERS FROM 'file:///show_extract.csv' AS row
26 MATCH (sc:ShowCatalog {id: toInteger(row.show_id)})
27 UNWIND split(row.cast, ',') AS actor_name
28 MERGE (a:Actor {actor_name: trim(actor_name)})
29 CREATE (sc)-[:HAS_CAST]→(a);
30
31 LOAD CSV WITH HEADERS FROM 'file:///show_extract.csv' AS row
32 MATCH (sc:ShowCatalog {id: toInteger(row.show_id)})
33 MERGE (c:Country {name_country: row.country})
34 CREATE (sc)-[:LOCATED_IN]→(c);
35
36 LOAD CSV WITH HEADERS FROM 'file:///show_extract.csv' AS row
37 MATCH (sc:ShowCatalog {id: toInteger(row.show_id)})
38 MERGE (d:Director {director_name: row.director})
39 CREATE (sc)-[:DIRECTED_BY]→(d);
40
```

44

```
1 from neo4j import GraphDatabase
2 import csv
3
4 neo4j_url = "bolt://localhost:7687"
5 neo4j_username = "neo4j"
6 neo4j_password = "wr99582435"
7 arq = 'Arquivo CSV/Arquivo_transicao/show_extract.csv'
8
9 def criar_inserir_neo4j(neo4j_url, neo4j_username, neo4j_password, csv_file):
10
11   conexao = GraphDatabase.driver(neo4j_url, auth=(neo4j_username, neo4j_password))
12
13   # Criando indices, acelera no processo de insercao de dados, sem a criação previa demorou 5
14   # minutos
15   with conexao.session() as session:
16     session.run("CREATE INDEX IF NOT EXISTS FOR (a:Actor) ON (a.actor_name)")
17     session.run("CREATE INDEX IF NOT EXISTS FOR (c:Country) ON (c.name_country)")
18     session.run("CREATE INDEX IF NOT EXISTS FOR (d:Director) ON (d.director_name)")
19
20   with conexao.session() as session:
21     with open(csv_file, 'r', encoding='utf-8') as file:      # Inserindo nós e relacionamentos
22       reader = csv.DictReader(file)
23       for row in reader:
24         # Cria nó ShowCatalog
25         query_show = (
26           "CREATE (:ShowCatalog {id: toInteger($show_id), type_show: $type_show, title: $title, director: $director, cast: $cast, "
27           "country: $country, date_added: $date_added, release_year: toInteger($release_year), rating: $rating, duration: $duration, "
28           "listed_in: $listed_in, description_show: $description_show})"
29         )
30         session.run(query_show, row)
31
32         # Cria relacionamento entre ShowCatalog e Actor
33         query_actor = (
34           "MATCH (sc:ShowCatalog {id: toInteger($show_id)}) "
35           "UNWIND split($cast, ',') AS actor_name "
36           "MERGE (a:Actor {actor_name: trim(actor_name)}) "
37           "CREATE (sc)-[:HAS_CAST]→(a)"
38         )
39         session.run(query_actor, row)
40
41         # Cria relacionamento entre ShowCatalog e Country
42         query_country = (
43           "MATCH (sc:ShowCatalog {id: toInteger($show_id)}) "
44           "MERGE (c:Country {name_country: $country}) "
45           "CREATE (sc)-[:LOCATED_IN]→(c)"
46         )
47         session.run(query_country, row)
48
49         # Cria relacionamento entre ShowCatalog e Director
50         query_director = (
51           "MATCH (sc:ShowCatalog {id: toInteger($show_id)}) "
52           "MERGE (d:Director {director_name: $director}) "
53           "CREATE (sc)-[:DIRECTED_BY]→(d)"
54         )
55         session.run(query_director, row)
56
57   conexao.close()
58
59
60 criar_inserir_neo4j(neo4j_url, neo4j_username, neo4j_password, arq)
```

# Queries de testes

```
▼ Click here to ask Blackbox to help you code faster  
1 MATCH (d:Director)  
2 RETURN d.director_name AS director  
3
```

```
1 MATCH (sc:ShowCatalog)  
2 WHERE sc.release_year = $ano      //definir o ano que voce quer  
3 RETURN sc.title AS title, sc.release_year AS release_year
```

```
1 MATCH (sc:ShowCatalog)  
2 RETURN sc.title AS title, sc.rating AS rating, sc.duration AS duration
```

```
1 MATCH (c:Country)←[:LOCATED_IN]-(sc:ShowCatalog)  
2 RETURN c.name_country AS country, count(sc) AS num_movies
```

# Database Application - Projeto UFC

## Integrantes:

Emanuel Ernesto - 01614951

Guilherme Barros - 01615925

Lucas Gomes - 01631591

Thiago Pessoa - 01089643

Wesley Ruan - 01555915

## Tecnologia:

Neo4j

link:

<https://neo4j.com/>

[https://www.canva.com/design/DAGEc7wTOSs/hNhZx5gr-IIKxJkqdgGHog/edit?  
utm\\_content=DAGEc7wTOSs&utm\\_campaign=designshare&utm\\_medium=link2&utm\\_source=sharebutton](https://www.canva.com/design/DAGEc7wTOSs/hNhZx5gr-IIKxJkqdgGHog/edit?utm_content=DAGEc7wTOSs&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

[wesleyruanwr/projeto UFC](#)