

INSTITUTO INFNET
PÓS EM ENGENHARIA DE DADOS

Juliana Opis Mikosz

Hadoop

2023

Introdução:

Este trabalho tem como objetivo realizar uma análise de dados em dois conjuntos de dados distintos: Filmes da Netflix e Transações de Compras.

Para a execução desse trabalho foi necessária a criação de um cluster no dataproc, em seguida a importação dos datasets escolhidos para um bucket dentro do GCP. Após, foi feita a cópia desses arquivos para o HDFS, criação de databases e tabelas no Hive através do Beeline. Os datasets armazenados no HDFS foram então injetados nas tabelas correspondentes no Hive. Por último algumas consultas para a análise de dados e obtenção de insights relevantes.

Conjuntos de dados:

Conjunto de Dados - Filmes da Netflix:

O primeiro conjunto de dados foi obtido a partir de fontes públicas no Kaggle. Ele inclui informações sobre filmes disponíveis na plataforma Netflix, como o ano de lançamento e as avaliações dos filmes. Estes dados serão utilizados para analisar e explorar a evolução das avaliações dos filmes ao longo do tempo, bem como para identificar padrões de qualidade com base nas avaliações dos usuários da plataforma.

Fonte: <https://www.kaggle.com/datasets/rishitjavia/netflix-movie-rating-dataset>

Conjunto de Dados - Transações de Compras:

O segundo conjunto de dados foi extraído do banco de dados da nossa empresa, mas todas as informações sensíveis foram anonimizadas e substituídas por dados fictícios para garantir a privacidade e a segurança dos dados. Este conjunto de dados contém informações sobre transações de compras realizadas pela empresa, incluindo detalhes como identificação da loja, cliente, valor da compra e datas das transações. Será realizada uma análise detalhada dessas transações para identificar tendências de compra, comportamentos dos clientes e padrões sazonais.

Passos do projeto:

1. Exportação dos arquivos do Bucket para o HDFS:

pos_engdados			
Local	Classe de armazenamento	Acesso público	
us (várias regiões nos Estados Unidos)	Standard	Não público	
OBJETOS	CONFIGURAÇÃO	PERMISSÕES	PROTEÇÃO
Intervalos > pos_engdados			
FAZER UPLOAD DE ARQUIVOS CARREGAR PASTA CRIAR PASTA TR			
Filtrar apenas pelo prefixo do nome ▾ Filtro Filtrar objetos e pastas			
<input type="checkbox"/>	Nome	Tamanho	Tipo
<input type="checkbox"/>	 CLIENTES.csv	17,8 MB	text/csv
<input type="checkbox"/>	 LOJAS.csv	103,9 KB	text/csv
<input type="checkbox"/>	 Netflix_Dataset_Movie.csv	558,1 KB	text/csv
<input type="checkbox"/>	 Netflix_Dataset_Rating.csv	237,3 MB	text/csv
<input type="checkbox"/>	 TRANSACOES.csv	37,2 MB	text/csv

2. Criação da pasta datasets:

Comando: `mkdir datasets`

3. Importação dos arquivos do bucket:

Comando: `gcsfuse pos_engdados datasets`

Datasets criados na pasta /datasets:

```
juliana_mikosz@cluster-laa7-m:~/datasets$ ls -la
total 299981
-rw-r--r-- 1 juliana_mikosz juliana_mikosz 18657607 Aug 31 21:45 CLIENTES.csv
-rw-r--r-- 1 juliana_mikosz juliana_mikosz 106443 Aug 31 21:45 LOJAS.csv
-rw-r--r-- 1 juliana_mikosz juliana_mikosz 571494 Aug 31 21:45 Netflix_Dataset_Movie.csv
-rw-r--r-- 1 juliana_mikosz juliana_mikosz 248836313 Aug 31 21:46 Netflix_Dataset_Rating.csv
-rw-r--r-- 1 juliana_mikosz juliana_mikosz 39007358 Aug 31 21:45 TRANSACOES.csv
```

4. Copia os arquivos locais para o HDFS:

Comando: `hdfs dfs -put *.csv /user/juliana_mikosz/datasets`

5. Conferência se os arquivos foram copiados:

Comando: `hdfs dfs -ls /user/juliana_mikosz/datasets`

Resultado:

```
juliana_mikosz@cluster-1aa7-m:~$ hdfs dfs -ls /user/juliana_mikosz/datasets
Found 5 items
-rw-r--r--  1 juliana_mikosz hadoop    18657607 2023-08-31 21:56 /user/juliana_mikosz/datasets
/CLIENTES.csv
-rw-r--r--  1 juliana_mikosz hadoop     106443 2023-08-31 21:56 /user/juliana_mikosz/datasets
/IOJAS.csv
-rw-r--r--  1 juliana_mikosz hadoop     571494 2023-08-31 21:56 /user/juliana_mikosz/datasets
/Netflix_Dataset_Movie.csv
-rw-r--r--  1 juliana_mikosz hadoop  248836313 2023-08-31 21:56 /user/juliana_mikosz/datasets
/Netflix_Dataset_Rating.csv
-rw-r--r--  1 juliana_mikosz hadoop   39007358 2023-08-31 21:56 /user/juliana_mikosz/datasets
/TRANSACOES.csv
```

Hive:

1. Conexão beeline:

Comando: *beeline -u jdbc:hive2://localhost:10000/default -n juliana_mikosz@cluster-1aa7-m -d org.apache.hive.jdbc.HiveDriver*

Conexão realizada:

```
juliana_mikosz@cluster-1aa7-m:~$ beeline -u jdbc:hive2://localhost:10000/default -n juliana_mi
kosz@cluster-1aa7-m -d org.apache.hive.jdbc.HiveDriver
Connecting to jdbc:hive2://localhost:10000/default
Connected to: Apache Hive (version 3.1.3)
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://localhost:10000/default> □
```

2. Criação dos databases e tabelas

Comando: *create database if not exists vendas comment "Database com informações de clientes, vendas e lojas";*

Comando: *create database if not exists netflix comment "Database com informações de filmes da Netflix";*

Databases criados:

```

0: jdbc:hive2://localhost:10000/default> create database if not exists vendas comment "Database com informações de clientes, vendas e lojas";
INFO : Compiling command(queryId=hive_20230831234118_255f6fb8-e8af-4e88-a7b6-2be1e2687d4e): create database if not exists vendas comment "Database com informações de clientes, vendas e lojas"
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20230831234118_255f6fb8-e8af-4e88-a7b6-2be1e2687d4e); Time taken: 0.019 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230831234118_255f6fb8-e8af-4e88-a7b6-2be1e2687d4e): create database if not exists vendas comment "Database com informações de clientes, vendas e lojas"
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230831234118_255f6fb8-e8af-4e88-a7b6-2be1e2687d4e); Time taken: 0.033 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.068 seconds)
0: jdbc:hive2://localhost:10000/default> create database if not exists netflix comment "Database com informações de filmes da Netflix";
INFO : Compiling command(queryId=hive_20230831234139_c6034228-8c9f-475c-9020-7ba547d159a4): create database if not exists netflix comment "Database com informações de filmes da Netflix"
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20230831234139_c6034228-8c9f-475c-9020-7ba547d159a4); Time taken: 0.024 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230831234139_c6034228-8c9f-475c-9020-7ba547d159a4): create database if not exists netflix comment "Database com informações de filmes da Netflix"
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230831234139_c6034228-8c9f-475c-9020-7ba547d159a4); Time taken: 0.043 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.083 seconds)

```

```

0: jdbc:hive2://localhost:10000/default> show databases;
INFO : Compiling command(queryId=hive_20230831234220_a422e91d-1928-4238-a930-8a607412e0cf): show databases
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20230831234220_a422e91d-1928-4238-a930-8a607412e0cf); Time taken: 0.018 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20230831234220_a422e91d-1928-4238-a930-8a607412e0cf): show databases
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20230831234220_a422e91d-1928-4238-a930-8a607412e0cf); Time taken: 0.005 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| database_name |
+-----+
| default      |
| netflix      |
| testejuliana |
| vendas       |
+-----+

```

Dicionário de dados tabelas | Database Vendas:

1. Tabela clientes

id (INT): Um identificador único para o cliente.
nome (STRING): O nome do cliente.
ativo (INT): Indica se o cliente está ativo ou não.
genero (STRING): O gênero do cliente.
city (STRING): A cidade em que o cliente reside.

2. Tabela transacoes

Id (INT): Um identificador único para a transação.
loja (INT): O identificador único da loja onde a transação ocorreu.
cliente (INT): O identificador único do cliente que fez a transação.
valor (INT): O valor da transação.
data (DATE): A data e hora em que a transação ocorreu.

3. Tabela lojas

Id (INT): Um identificador único para a loja.
ativo (INT): Indica se a loja está ativa ou não.
cidade (STRING): A cidade onde a loja está localizada.
data_instalacao (DATE): A data de instalação da loja.

Criação das tabelas | Database vendas:

1. Criação da tabela clientes:

```
CREATE TABLE clientes (id INT, nome STRING, ativo INT, genero STRING, city STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS TEXTFILE;
```

2. Ingestão dos dados na tabela criada:

```
LOAD DATA INPATH '/user/juliana_mikosz/datasets/CLIENTES.csv' overwrite into table
clientes;
```

3. Criação da tabela transacoes:

```
CREATE TABLE transacoes (Id INT, loja INT, cliente INT, valor INT, datahora DATE)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS TEXTFILE;
```

4. Ingestão dos dados na tabela criada:

```
LOAD DATA INPATH '/user/juliana_mikosz/datasets/TRANSACOES.csv' overwrite into table
transacoes;
```

5. Criação da tabela lojas:

```
CREATE TABLE lojas (Id INT, ativo INT, cidade STRING, data_instalacao DATE)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS TEXTFILE;
```

6. Ingestão dos dados na tabela criada:

```
LOAD DATA INPATH '/user/juliana_mikosz/datasets/LOJAS.csv' overwrite into table lojas;
```

Dicionário de dados tabelas | Database Vendas:

1. Tabela filme

Id (INT): Um identificador único para o filme.

year (INT): O ano de lançamento do filme.

name (STRING): O nome do filme.

2. Tabela avaliacao

Id (INT): Um identificador único para a avaliação.

rating (INT): A classificação atribuída ao filme na avaliação.

movie_id (INT): O identificador único do filme ao qual esta avaliação está associada.

Criação das tabelas | Database netflix:

1. Criação da tabela filme:

```
CREATE TABLE filme (movie_id INT, year INT, name STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS TEXTFILE;
```

2. Ingestão dos dados na tabela criada:

```
LOAD DATA INPATH '/user/juliana_mikosz/datasets/Netflix_Dataset_Movie.csv' overwrite
into table filme;
```

3. Criação da tabela avaliacao:

```
CREATE TABLE avaliacao (user_id INT, rating INT, movie_id INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS TEXTFILE;
```

4. Ingestão dos dados na tabela criada:

```
LOAD DATA INPATH '/user/juliana_mikosz/datasets//Netflix_Dataset_Rating.csv' overwrite
into table avaliacao;
```

Consultas Hive:

1. Database Vendas:

a) Qual o total de transações por cliente?

Consulta realizada:

```
SELECT c.id, c.nome, COUNT(*) AS total_transacoes
FROM clientes c
LEFT JOIN transacoes t ON c.id = t.cliente
GROUP BY c.id, c.nome
ORDER BY total_transacoes DESC
LIMIT 20;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	3	3	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 23.04 s								

Resposta:

c.id	c.nome	total_transacoes
450405	Gabriel	2182
505132	Manuela	1097
444612	Manuela	1007
405123	Afonso	786
432171	Pedro	747
450426	Alice	716
419959	Lucas	664
525734	Guilherme	605
559536	Laura	585
601902	Marina	538
438969	Gabriel	456
498439	Caio	435
512916	Muriel	340
280302	Bianca	316
455609	Isaac	253
513277	Miguel	247
284187	Gabriel	224
517632	Eduarda	212
405468	Alice	206
462861	Felipe	204

b) Qual o total de vendas por cidade?

Consulta realizada:

```
SELECT l.cidade, COUNT(t.id) AS total_vendas
FROM lojas l
LEFT JOIN transacoes t ON l.id = t.loja
GROUP BY l.cidade
ORDER BY total_vendas DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 14.08 s								

Resposta:

l.cidade	total_vendas
São Paulo	346458
Curitiba	113490
Rio de Janeiro	35798
Belo Horizonte	26875
Guarulhos	26062
Blumenau	23599
Brasília	22371
Salvador	18830
Porto Alegre	18027
Londrina	16353
Maceió	13871
Jundiaí	12941
Vitória	12396
Osasco	11622
Santo André	10918
São José dos Campos	10560
Cuiabá	10104
Campo Grande	10019
Santos	9823
Teresina	8531
Vila Velha	8412
Uberlândia	8358
Barueri	8282
Florianópolis	8247
São José	8160
Novo Hamburgo	8030
Cascavel	7588
Nova Lima	6783
João Pessoa	6462
Bauru	5829
Joinville	5700
Araçatuba	5667
Diadema	5345
Niterói	5331
Campinas	5276
Maringá	5263
Recife	5073

c) Qual o valor total de compras por cliente?

Consulta realizada:

```
SELECT c.id, c.nome, SUM(t.valor) AS valor_total_compras
FROM clientes c
LEFT JOIN transacoes t ON c.id = t.cliente
GROUP BY c.id, c.nome
ORDER BY valor_total_compras DESC
LIMIT 20;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	3	3	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 28.03 s								

Resposta:

c.id	c.nome	valor_total_compras
450405	Gabriel	194545
505132	Manuela	99883
444612	Manuela	88969
405123	Afonso	68857
432171	Pedro	63592
450426	Alice	61619
419959	Lucas	59134
525734	Guilherme	55120
559536	Laura	53434
601902	Marina	49801
438969	Gabriel	40300
498439	Caio	37255
512916	Muriel	30425
280302	Bianca	27993
455609	Isaac	22594
513277	Miguel	20546
284187	Gabriel	19302
405468	Alice	18211
462861	Felipe	17911
517632	Eduarda	17306

d) Qual a quantidade de clientes ativos por loja?

Consulta realizada:

```
SELECT l.id, l.cidade, COUNT(DISTINCT c.id) AS clientes_ativos
FROM lojas l
LEFT JOIN transacoes t ON l.id = t.loja
LEFT JOIN clientes c ON t.cliente = c.id
WHERE c.ativo = 1
GROUP BY l.id, l.cidade
ORDER BY clientes_ativos DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 2	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
Reducer 5	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 21.75 s								

Resposta:

l.id	l.cidade	clientes_ativos
3225	São Paulo	1263
1555	Guarulhos	819
2245	Recife	801
1821	São José	764
1001	Curitiba	740
1741	Santos	732
1707	Salvador	696
1823	Londrina	660
2432	São Paulo	641
2655	São Paulo	638
1043	Curitiba	599
1411	Jundiaí	589
1112	Campo Grande	589
3252	Curitiba	588
2565	São Paulo	563
2868	São Paulo	547
3212	Brasília	540
1039	Curitiba	519
1014	Curitiba	516
1231	Belo Horizonte	514
1862	Londrina	512
3337	Curitiba	495
3249	Juiz de Fora	495
2985	Brasília	491
1893	Votorantim	483
1746	Sorocaba	463
1182	Curitiba	456
2209	São Paulo	454
1121	São Bernardo do Campo	438
2968	Vitória	436
2407	Salvador	422
1052	Curitiba	417
2175	Teresina	411
3105	Barueri	409
1095	Curitiba	409
2073	São Paulo	408
2767	São Paulo	408
2364	Jundiaí	407
1315	Nova Lima	403
2132	Porto Alegre	400

e) **Qual a quantidade de clientes ativos por cidade?**

Consulta realizada:

```
SELECT l.cidade, COUNT(DISTINCT c.id) AS clientes_ativos
FROM lojas l
LEFT JOIN transacoes t ON l.id = t.loja
LEFT JOIN clientes c ON t.cliente = c.id
WHERE c.ativo = 1
GROUP BY l.cidade
ORDER BY clientes_ativos DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 2	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
Reducer 5	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 21.69 s								

Resposta:

l.cidade	clientes_ativos
São Paulo	61377
Curitiba	22951
Rio de Janeiro	6145
Guarulhos	5520
Belo Horizonte	5433
Brasília	5390
Salvador	4718
Blumenau	3380
Londrina	3248
Santos	2846
Cuiabá	2796
Porto Alegre	2694
Maceió	2686
Jundiaí	2594
Santo André	2518
Vitória	2335
Osasco	2316
Campo Grande	2273
Teresina	2142
São José dos Campos	1898
Cascavel	1841
Barueri	1720
Uberlândia	1677
Nova Lima	1508
Niterói	1395
São José	1334
Novo Hamburgo	1304
Mogi das Cruzes	1282
Vila Velha	1281
Florianópolis	1240
Recife	1224
Diadema	1218
Bauru	1186
Araçatuba	1179
Campinas	1176
João Pessoa	1166
Biguaçu	1139
Caruaru	1110
São Bernardo do Campo	1070

f) Quais são as 3 lojas com o maior valor de venda em cada mês de 2023?

Consulta realizada:

```

WITH VendasPorLojaPorMes AS (
  SELECT
    l.id AS id_loja,
    l.cidade AS cidade,
    EXTRACT(MONTH FROM t.datahora) AS mes,
    SUM(t.valor) AS total_vendas,
    DENSE_RANK() OVER (PARTITION BY EXTRACT(MONTH FROM t.datahora)
      ORDER BY SUM(t.valor) DESC) AS ranking
  FROM lojas l
  LEFT JOIN transacoes t ON l.id = t.loja
  WHERE l.id IS NOT NULL AND EXTRACT(MONTH FROM t.datahora) IS NOT
  NULL
  GROUP BY EXTRACT(MONTH FROM t.datahora), l.id, l.cidade
)
SELECT id_loja, cidade, mes, total_vendas
FROM VendasPorLojaPorMes

```

WHERE ranking <= 3
 AND mes IS NOT NULL
 ORDER BY mes, ranking;

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 5	container	SUCCEEDED	1	1	0	0	0	0
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Reducer 2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	container	SUCCEEDED	2	2	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 14.95 s								

Resposta:

id_loja	cidade	mes	total_vendas
1555	Guarulhos	1	240193
2432	São Paulo	1	232895
3248	São Paulo	1	224702
3248	São Paulo	2	175888
1821	São José	2	160414
2767	São Paulo	2	138667
1043	Curitiba	3	7160
2320	Suzano	3	7098
2655	São Paulo	3	6559
2655	São Paulo	4	4888
2013	Brasília	4	3592
2432	São Paulo	4	3547
2407	Salvador	5	2663
1707	Salvador	5	2593
1043	Curitiba	5	2561
1741	Santos	6	2389
2013	Brasília	6	2057
2407	Salvador	6	2029
1741	Santos	7	3113
2655	São Paulo	7	2238
2013	Brasília	7	2213
2132	Porto Alegre	8	9468
2655	São Paulo	8	4455
1741	Santos	8	3558

g) Qual a quantidade de clientes femininos X masculinos por cidade?

Consulta realizada:

```
SELECT l.cidade, c.genero, COUNT(DISTINCT c.id) AS total_clientes
FROM lojas l
LEFT JOIN transacoes t ON l.id = t.loja
LEFT JOIN clientes c ON t.cliente = c.id
WHERE c.genero IS NOT NULL AND c.genero IN ('F', 'M')
GROUP BY l.cidade, c.genero;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 2	container	SUCCEEDED	1	1	0	0	0	0
Map 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 04/04 [=====>>] 100% ELAPSED TIME: 21.34 s								

Resposta:

l.cidade	c.genero	total_clientes
Agudos	F	30
Agudos	M	39
Amparo	F	190
Amparo	M	293
Ananindeua	F	44
Ananindeua	M	44
Anápolis	F	46
Anápolis	M	88
Apiúna	F	86
Apiúna	M	79
Aracaju	F	1
Araraquara	F	2
Araraquara	M	3
Araras	F	96
Araras	M	86
Araucária	F	1
Araçatuba	F	597
Araçatuba	M	605
Balneário Camboriú	M	3
Barueri	F	970
Barueri	M	777
Bauru	F	613
Bauru	M	573
Belo Horizonte	F	2679
Belo Horizonte	M	2824
Belém	F	424
Belém	M	497
Bento Gonçalves	F	257
Bento Gonçalves	M	258
Betim	F	276
Betim	M	305
Biguaçu	F	563
Biguaçu	M	576
Birigüi	F	82
Birigüi	M	99
Blumenau	F	1947
Blumenau	M	1461
Brasília	F	2729
Brasília	M	2720
Cabedelo	F	242
Cabedelo	M	315

2. Database Netflix:

a) Qual a média de avaliação por filme?

Consulta realizada:

```
SELECT f.name AS filme, AVG(a.rating) AS media_avaliacao
FROM filme f
LEFT JOIN avaliacao a ON f.movie_id = a.movie_id
WHERE a.rating IS NOT NULL
GROUP BY f.name
ORDER BY media_avaliacao DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 2	container	SUCCEEDED	5	5	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 04/04 [=====>>] 100% ELAPSED TIME: 26.12 s								

Resposta:

filme	media_avaliacao
Lost: Season 1	4.665432098765432
The Simpsons: Season 6	4.589824034920202
Family Guy: Freakin' Sweet Collection	4.520766378244747
Six Feet Under: Season 4	4.461601211979955
Inu-Yasha	4.457773512476008
Stargate SG-1: Season 8	4.456026058631922
The Best of Friends: Vol. 4	4.449167996352861
The West Wing: Season 3	4.43625843780135
Lord of the Rings: The Fellowship of the Ring	4.43148917942777
Gilmore Girls: Season 3	4.428942582488959
Firefly	4.410771903698188
The Simpsons: Season 3	4.408041422306757
Finding Nemo (Widescreen)	4.395903857377832
Samurai Champloo	4.395629238884703
The Simpsons: Treehouse of Horror	4.395329752312742
The Godfather	4.392642600755082
CSI: Season 1	4.375887066005341
Buffy the Vampire Slayer: Season 6	4.37184159378037
Stargate SG-1: Season 7	4.3612002264578225
Friends: Season 6	4.359458908299912
The Simpsons: Bart Wars	4.338631465517241
Alias: Season 1	4.319803600654664
Nip/Tuck: Season 2	4.318832283915284
Futurama: Monster Robot Maniac Fun Collection	4.315998237108858
The Sixth Sense	4.3114694850355555
Angel: Season 4	4.310205515545406
Farscape: The Peacekeeper Wars	4.304888888888889
The Silence of the Lambs	4.303809453673214
The Simpsons: Season 1	4.297105894161602
Curb Your Enthusiasm: Season 3	4.291043897846299
Pride and Prejudice	4.2896135424184765
Sex and the City: Season 4	4.284771634137615
The Best of Friends: Season 1	4.283065908073899
The Blues Brothers: Extended Cut	4.282422123331214
Star Trek: The Next Generation: Season 6	4.281840380413828
Stargate SG-1: Season 2	4.267909715407262
Braveheart	4.265251032764311
The Best of Friends: Season 2	4.256434882827507
Star Trek: The Next Generation: Season 7	4.250601765177855
Stargate SG-1: Season 3	4.247497578301582
Batman Begins	4.244315458757122

b) Quais os filmes com a classificação acima de 4.5?

Consulta realizada:

```
SELECT f.name AS filme, AVG(a.rating) AS media_avaliacao
```

FROM filme f

LEFT JOIN avaliacao a ON f.movie_id = a.movie_id

GROUP BY f.name

HAVING AVG(a.rating) >= 4.5

ORDER BY media_avaliacao DESC;

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	5	5	0	0	0	0
Reducer 2	container	SUCCEEDED	10	10	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0

VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 42.13 s

Resposta:

```
+-----+-----+
|               filme               | media_avaliacao |
+-----+-----+
| Lost: Season 1                     | 4.665432098765432 |
| The Simpsons: Season 6             | 4.589824034920202 |
| Family Guy: Freakin' Sweet Collection | 4.520766378244747 |
+-----+-----+
```

c) Qual a melhor avaliação por filme lançado em 2005?

Consulta realizada:

```
SELECT f.name AS filme, f.year AS ano_lancamento, MAX(a.rating) AS rating
```

FROM filme f

LEFT JOIN avaliacao a ON f.movie_id = a.movie_id

WHERE f.year = 2005

GROUP BY f.name, f.year

```
ORDER BY rating DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	5	5	0	0	0	0
Reducer 2	container	SUCCEEDED	10	10	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>] 100% ELAPSED TIME: 37.51 s								

Resposta:

filme	ano_lancamento	rating
7 Seconds	2005	5
Alias: Season 4	2005	5
Coach Carter	2005	5
11:14	2005	5
Batman Begins	2005	5
Hostage	2005	5
Saving Face	2005	5
Empire Falls	2005	5
King's Ransom	2005	5
Beauty Shop	2005	5
The Ballad of Jack and Rose	2005	5
The Sandlot 2	2005	5
Dead Birds	2005	5
Pooh's Heffalump Movie	2005	5
Kicking & Screaming	2005	5
Look at Me	2005	5
Nobody Knows	2005	5
The Pacifier	2005	5
The Amityville Horror	2005	5
The Hitchhiker's Guide to the Galaxy	2005	5
The L Word: Season 2	2005	5
Unleashed	2005	5

d) Qual o número de avaliações por filme?

Consulta realizada:

```
SELECT f.name AS filme, COUNT(a.rating) AS num_avaliacoes
```

FROM filme f

LEFT JOIN avaliacao a ON f.movie_id = a.movie_id

GROUP BY f.name

HAVING COUNT(a.rating) > 0

```
ORDER BY num_avaliacoes DESC;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 5	container	SUCCEEDED	5	5	0	0	0	0
Reducer 2	container	SUCCEEDED	10	10	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 05/05 [=====>>>] 100% ELAPSED TIME: 41.94 s								

Resposta:

filme	num_avaliacoes
Pirates of the Caribbean: The Curse of the Black Pearl	117075
Lord of the Rings: The Fellowship of the Ring	102721
The Sixth Sense	102376
American Beauty	101450
Bruce Almighty	98545
The Silence of the Lambs	95053
Finding Nemo (Widescreen)	94235
The Italian Job	93886
Shrek 2	92893
Braveheart	91502
Ghost	87082
What Women Want	86756
The Last Samurai	86354
50 First Dates	85605
The Bourne Supremacy	85247
A Beautiful Mind	82347
Men in Black II	81371
The Matrix: Reloaded	79504
Speed	79476
Sleepless in Seattle	78996
Something's Gotta Give	77502
Man on Fire	77447
Kill Bill: Vol. 2	77312
Lethal Weapon	76147
Road to Perdition	74652
The Wedding Planner	74461
X2: X-Men United	73684
Signs	71405
Napoleon Dynamite	71117
Being John Malkovich	70208
The Recruit	69635
Patch Adams	69461
Eternal Sunshine of the Spotless Mind	69342
Liar Liar	69105
The Mummy	68783
Sideways	68756

e) Qual a média de avaliação por ano de lançamento?

Consulta realizada:

```
SELECT f.year AS ano_lancamento, AVG(a.rating) AS media_avaliacao
FROM filme f
LEFT JOIN avaliacao a ON f.movie_id = a.movie_id
WHERE a.rating IS NOT NULL
GROUP BY f.year
HAVING AVG(a.rating) IS NOT NULL
ORDER BY f.year;
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	1	1	0	0	0	0
Map 2	container	SUCCEEDED	5	5	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	0	0	0
Reducer 4	container	SUCCEEDED	1	1	0	0	0	0
VERTICES: 04/04 [=====>>] 100% ELAPSED TIME: 23.77 s								

Resposta:

ano_lancamento	media_avaliacao
1920	3.3920122887864825
1925	3.8184647302904566
1929	3.6844629559510285
1930	3.8130872204207202
1931	3.75057368053477
1934	3.917944578961528
1935	3.927168416958382
1936	3.916519347868788
1938	4.065483476132191
1939	4.107672595089205
1940	3.8154218162278344
1941	3.9408571751667987
1942	3.774837511606314
1943	3.727823691460055
1944	3.7837876006213813
1946	3.9582454192115493
1947	3.765226756012923
1949	3.956489275047516
1951	3.7986543100165036
1952	3.972993883126617
1953	3.7395563954644917
1954	4.082166977493708
1955	3.865171413951902
1956	3.7712713696207505
1957	3.739489926633283
1958	3.627589063794532
1959	4.02670545071843
1960	3.721403967217009
1961	3.9251369757796097
1962	3.8925190194420964
1963	3.8642983664632946
1964	3.8438571973065683
1965	3.892512491894572
1966	3.5613019891500906
1967	3.6814814814814816
1968	3.616407931492421
1969	3.6984473835537663