# **Queries SQL**

Arquivo com as Queries que foram executadas e as devidas observações.

No github também encontra-se um notebook (sql/databricks\_notebook.py), exportado do Databricks, contendo as mesmas queries.

## **Queries de validação:**

Inicialmente fiz algumas queries para validar os dados, de forma bem superficial.

* 1. **Validação Total de consumidores Netflix**

SELECT

    COUNT(DISTINCT customer\_id)

FROM

    movie\_star\_rating WHERE platform = "NFLX"

Resultado: 480189

* 1. **Validação id máximo de consumidor Netflix**

SELECT

    MAX(customer\_id)

FROM

    movie\_star\_rating WHERE platform = "NFLX"

Resultado: 2649429

## **Queries das Questões de Negócio**

* 1. **Quantos filmes estão disponíveis na Amazon? E na Netflix?**

No caso da Netflix foi utilizada a tabela com dados de filme: netflix\_movie\_data. Parti da idéia que poderia ter filmes que não havia sido avaliados.

SELECT

    COUNT(DISTINCT product\_id)

FROM

    netflix\_movie\_data

Resultado: 17770

No caso da Amazon eu fiz a busca pelo título, pois temos o mesmo filme em categorias diferentes (DVD, DOWNLOAD, etc)

SELECT

    COUNT(DISTINCT product\_title)

FROM

    movie\_star\_rating WHERE platform = "AMZ"

Resultado: 407777

Resposta: Na NETFLIX temos 17.700 filmes disponíveis e na AMAZON 407.777

* 1. **Dos filmes disponíveis na Amazon, quantos % estão disponíveis na Netflix?**

WITH same\_movies\_netflix\_amazon AS (

    SELECT

        COUNT(DISTINCT product\_title) AS total

FROM

    movie\_star\_rating

WHERE

    platform = "AMZ" AND

    product\_title IN

    (

        SELECT

            DISTINCT product\_title

        FROM

            netflix\_movie\_data

    )

),

all\_movies\_amazon AS (

    SELECT

        COUNT(DISTINCT product\_title) total

    FROM

        movie\_star\_rating

    WHERE

        platform = "AMZ"

)

SELECT

    ROUND(100\* (SELECT total from same\_movies\_netflix\_amazon) / (SELECT total from all\_movies\_amazon), 2) AS perc\_total

Resultado: 2,75%

* 1. **O quão perto a média das notas dos filmes disponíveis na Amazon está dos filmes disponíveis na Netflix?**

WITH avg\_amazon AS (

    SELECT

        AVG(star\_rating) AS avg

    FROM

        movie\_star\_rating

    WHERE

        platform = "AMZ" AND

        star\_rating IS NOT NULL

),

avg\_netflix AS (

    SELECT

        AVG(star\_rating) AS avg

    FROM

        movie\_star\_rating

    WHERE

        platform = "NFLX" AND

        star\_rating IS NOT NULL

)

SELECT

    ROUND (100\* (SELECT avg FROM avg\_amazon)/ (SELECT avg FROM avg\_netflix)) AS diff\_avg\_rating

Resultado: 118%

* 1. **Qual ano de lançamento possui mais filmes na Netflix?**

WITH tb\_release\_per\_year AS (

  SELECT

    year\_of\_release,

    COUNT(1) AS total

  FROM

    netflix\_movie\_data

  GROUP BY

    year\_of\_release

)

SELECT

  year\_of\_release

FROM

  tb\_release\_per\_year

ORDER BY

  total DESC

LIMIT 1

Resposta: 2004

* 1. **Quais filmes que não estão disponíveis no catálogo da Amazon foram melhor avaliados? (Melhores notas são as 4 e 5)**

Resultado: 6100 títulos, disponíveis no arquivo sql/resultado/questao\_5.csv.

WITH amazon\_data AS (

  SELECT

      DISTINCT product\_title

  FROM

      movie\_star\_rating

  WHERE

      platform = "AMZ"

)

SELECT

    DISTINCT product\_title

 FROM

    movie\_star\_rating

  WHERE

      platform = "NFLX" AND

      product\_title NOT IN (SELECT product\_title FROM amazon\_data) AND

      star\_rating IN (4, 5)

* 1. **Quais filmes que não estão disponíveis no catálogo da Netflix foram \*melhor avaliados? (Melhores notas são as 4 e 5)**

Resultado: 350653 títulos, disponíveis no arquivo sql/resultado/questao\_6.csv.

WITH netflix\_data AS (

  SELECT

      DISTINCT product\_title

  FROM

      movie\_star\_rating

  WHERE

      platform = "NFLX"

)

SELECT

    DISTINCT product\_title

 FROM

    movie\_star\_rating

  WHERE

      platform = "AMZ" AND

      product\_title NOT IN (SELECT product\_title FROM netflix\_data) AND

      star\_rating IN (4, 5)

* 1. **Quais os 10 filmes que possuem mais avaliações nas plataformas?**

SELECT

  \*

FROM

    (

      WITH total\_rating AS (

        SELECT

          product\_title,

          COUNT(star\_rating) AS total,

          platform

        FROM

            movie\_star\_rating

        WHERE

            star\_rating is NOT NULL

        GROUP BY

          platform, product\_title

      )

      SELECT

        \*,

        row\_number() over (partition by platform order by total desc) as movie\_rank

      FROM

        total\_rating

    )

WHERE

    movie\_rank <= 10

ORDER BY

  platform, movie\_rank

Resultado: disponível no arquivo sql/resultado/questao\_7.csv

* 1. **Quais são os 5 clientes que mais avaliaram filmes na Amazon e quantos produtos diferentes eles avaliaram? E na Netflix?**

WITH customer\_total\_rating AS (

  SELECT

    customer\_id,

    COUNT(DISTINCT product\_title) AS total

  FROM

      movie\_star\_rating

  WHERE

      star\_rating is NOT NULL AND

      platform = "AMZ"

  GROUP BY

    customer\_id

)

SELECT

    \*

FROM

    customer\_total\_rating

ORDER BY

  total DESC

LIMIT 5

Resultado: disponível no arquivo sql/resultado/questao\_8\_amazon.csv

WITH customer\_total\_rating AS (

  SELECT

    customer\_id,

    COUNT(DISTINCT product\_title) AS total

  FROM

      movie\_star\_rating

  WHERE

      star\_rating is NOT NULL AND

      platform = "NFLX"

  GROUP BY

    customer\_id

)

SELECT

    \*

FROM

    customer\_total\_rating

ORDER BY

  total DESC

LIMIT 5

Resultado: disponível no arquivo sql/resultado/questao\_8\_netflix.csv

* 1. **Quantos filmes foram avaliados na data de avaliação mais recente na Amazon?**

SELECT

  COUNT(DISTINCT product\_title) AS total,

  rating\_date

FROM

  movie\_star\_rating

WHERE

  platform = "AMZ" AND

  star\_rating is NOT NULL AND

  rating\_date  IN (

    SELECT

      DISTINCT MAX(rating\_date)

    FROM

      movie\_star\_rating

    WHERE

      platform = "AMZ"

  )

GROUP BY ALL

Resultado: 6640 filmes em 2015-08-31

* 1. **Quais os 10 filmes mais bem avaliados nesta data?**

%sql

WITH best\_movies AS (

  SELECT

     product\_title,

     star\_rating,

     rating\_date

  FROM

    movie\_star\_rating

  WHERE

    platform = "AMZ" AND

    rating\_date  IN (

      SELECT

        DISTINCT MAX(rating\_date)

      FROM

        movie\_star\_rating

      WHERE

        platform = "AMZ"

    )

  ORDER BY star\_rating DESC

)

SELECT

  \*

FROM

  best\_movies

ORDER BY

  star\_rating DESC

LIMIT 10

Resultado: disponível no arquivo sql/resultado/questao\_10.csv. Importante: O resultado sempre vai ser diferente pois existem vários filmes com a mesma classificação. Seria necessário ranquear por outro indicador adicional.

* 1. **Quantos filmes foram avaliados na data de avaliação mais recente na Netflix?**

SELECT

  COUNT(DISTINCT product\_title) AS total,

  rating\_date

FROM

  movie\_star\_rating

WHERE

  platform = "NFLX" AND

  star\_rating is NOT NULL AND

  rating\_date  IN (

    SELECT

      DISTINCT MAX(rating\_date)

    FROM

      movie\_star\_rating

    WHERE

      platform = "NFLX"

  )

GROUP BY ALL

Resultado: 6848 filmes em 2005-12-31

* 1. **Quais os 10 filmes mais bem avaliados nesta data?**

WITH best\_movies AS (

  SELECT

     product\_title,

     star\_rating,

     rating\_date

  FROM

    movie\_star\_rating

  WHERE

    platform = "NFLX" AND

    rating\_date  IN (

      SELECT

        DISTINCT MAX(rating\_date)

      FROM

        movie\_star\_rating

      WHERE

        platform = "NFLX"

    )

  ORDER BY star\_rating DESC

)

SELECT

  \*

FROM

  best\_movies

ORDER BY

  star\_rating DESC

LIMIT 10

Resultado: disponível no arquivo sql/resultado/questao\_12.csv. Importante: O resultado sempre vai ser diferente pois existem vários filmes com a mesma classificação. Seria necessário ranquear por outro indicador adicional.