CEFET/RJ

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA

CAMPUS MARIA DA GRAÇA

SISTEMAS DE INFORMAÇÃO

BANCO DE DADOS ||

STAR WARS

Rio de Janeiro – RJ

Dezembro-2025

STAR WARS

EQUIPE:

Bianca de Jesus

Jeovanna Picanço

Maria dos Anjos

Matheus Satana

Professor: Diego Cardoso

Sumário

1.Introdução

O trabalho tem como objetivo desenvolver um ambiente completo de banco de dados a partir de uma base pública composta por informações fornecidas por entrevistados sobre a franquia Star Wars. Esses dados abrangem perfis demográficos, opiniões sobre personagens, rankings de filmes e indicadores de preferência, configurando um conjunto amplo e heterogêneo. Assim, a primeira etapa do projeto consistiu em compreender a estrutura original da base, identificar inconsistências e elaborar um dicionário de dados inicial capaz de orientar o processo de modelagem.

Posteriormente, foi realizada uma reestruturação completa do modelo por meio de normalização, definição de entidades e chaves, padronização de tipos e correção de redundâncias. Essa etapa incluiu o desenvolvimento de scripts responsáveis por transportar os dados do modelo original para o modelo ajustado de forma íntegra e confiável. Além disso, foram criados índices destinados a otimizar o desempenho das consultas, especialmente em cenários analíticos e de grande volume de leitura.

O trabalho também contemplou a implementação de automatizações no PostgreSQL como triggers, functions, views e procedures, por fim, a construção de um Data Warehouse baseado em modelagem dimensional. Esse DW foi projetado a partir das principais perguntas de negócio identificadas, possibilitando análises mais robustas e estruturadas. Dessa forma, o projeto integra desde a análise da fonte de dados até a entrega de um ambiente analítico completo, destacando a importância das boas práticas de engenharia de dados.

2.Dicionário de Dados Inicial

Dicionário de dados inicial foi construído a partir da análise direta da base original, sem qualquer modificação estrutural. A tabela principal, denominada star_wars, contém todas as respostas dos entrevistados, reunindo dados demográficos, avaliações e preferências em uma única estrutura. Esse dicionário inicial teve como função compreender completamente o estado bruto da base e servir como ponto de partida para as etapas de normalização e reorganização de dados.

Durante a análise, foram encontradas as seguintes inconsistências estruturais:

- Ausência de chave primária (Primary Key – PK): Não existe uma coluna declarada como identificadora única da tabela. A coluna RespondentID cumpre parcialmente essa função, mas contém valores em ponto flutuante e não está configurada como PK.
- Ausência de chaves estrangeiras (Foreign Keys – FK): Mesmo contendo informações que deveriam ser distribuídas entre diferentes entidades (respondentes, filmes, rankings, personagens), toda a estrutura está agregada em uma única tabela, impossibilitando a existência de relacionamentos formais.
- Colunas sem título: nomeadas automaticamente como Unnamed: X, refletindo falhas no processo de exportação da base.
- Campos agregados: nos quais múltiplas respostas de uma lista são divididas em colunas paralelas, dificultando o tratamento relacional.
- Tipos de dados genéricos: majoritariamente varchar(50), utilizados mesmo quando o conteúdo deveria ser numérico ou categórico bem definido.

A seguir, apresenta-se o detalhamento de cada coluna, com seu nome original, tipo identificado, descrição e observações da tabela star_wars:

Coluna: RespondentID

- Tipo: float4
- Descrição: Identificador numérico do respondente.
- Observações: Não é uma chave primária declarada.

Coluna: Have you seen any of the 6 films in the Star Wars franchise?

- Tipo: varchar(50)

- Descrição: Indica se o respondente já assistiu algum dos seis filmes originais da franquia Star Wars.
- Valores comuns: “Yes”, “No”.
- Observações: Nome da coluna foi renomeado na normalização.

Coluna: Do you consider yourself to be a fan of the Star Wars film franchise?

- Tipo: varchar(50)
- Descrição: Pergunta se o respondente se considera fã da franquia Star Wars.
- Observações: Nome da coluna foi renomeado na normalização.

Coluna: Which of the following Star Wars films have you seen? Please select all that apply.

- Tipo: varchar(50)
- Descrição: Marca se o respondente assistiu determinados filmes específicos da lista.
- Observações: Aqui, apenas o primeiro filme aparece, os demais estão nas colunas Unnamed: 4 a Unnamed: 8.

Colunas: Unnamed: 4, Unnamed: 5, Unnamed: 6, Unnamed: 7, Unnamed: 8

- Tipo: varchar(50)
- Descrição: Cada coluna representa um filme adicional selecionado pelo participante.
- Valores típicos:
    - Episode I – The Phantom Menace
    - Episode II – Attack of the Clones
    - Episode III – Revenge of the Sith
    - Episode IV – A New Hope
    - Episode V – The Empire Strikes Back
    - Episode VI – Return of the Jedi
- Observações: Essas colunas são resultam da lista de filmes selecionados e necessitam ser reorganizadas.

Coluna: Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film.

- Tipo: varchar(50)
- Descrição: Representa a classificação do filme preferido (1) ao menos preferido (6).
- Observações: Como nas colunas anteriores, cada ranking por filme está distribuído nas colunas seguintes: Unnamed: 10 a Unnamed: 14.

Colunas: Unnamed: 10, Unnamed: 11, Unnamed: 12, Unnamed: 13, Unnamed: 14

- Tipo: varchar(50)
- Descrição: Cada coluna corresponde ao ranking atribuído a um filme específico.
- Valores comuns: 1 a 6 (ordem de preferência).
- Observações: Essa identificação está implícita pela posição da coluna.

Colunas: Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her.

- Tipo: varchar(50)
- Descrição: Representam o início da seção sobre avaliação de personagens. O campo registra a opinião do respondente sobre o personagem.
- Valores comuns: Favorable, Unfavorable, Neither favorable nor unfavorable, Unfamiliar (N/A).

Colunas: Unnamed: 16 a Unnamed: 28

- Tipo: varchar(50)
- Descrição: Cada coluna contém a opinião do respondente sobre um personagem específico.
- Exemplos de personagens mapeados:
  - Han Solo
  - Luke Skywalker
  - Princess Leia
  - Anakin Skywalker
  - Obi-Wan Kenobi

    - Darth Vader
    - Palpatine
    - Yoda
    - Padmé Amidala
    - Jar Jar Binks
    - C-3PO
    - R2-D2
- Observações: Colunas foram separadas em uma tabela na etapa normalização.
- Valores comuns:
    - "Very favorably"
    - "Somewhat favorably"
    - "Neither favorably nor unfavorably"
    - "Unfamiliar (N/A)"

Coluna: Which character shot first?

- Tipo: varchar(50)
- Descrição: Identifica a resposta do participante à pergunta polêmica “Quem atirou primeiro?”
- Valores comuns: “Han”, “Greedo”, “I don't understand this question”.

Coluna: Are you familiar with the Expanded Universe?

- Tipo: varchar(50)
- Descrição: Indica se o respondente conhece o Universo Expandido de Star Wars.

Coluna: Do you consider yourself to be a fan of the Expanded Universe?

- Tipo: varchar(50)
- Descrição: Pergunta se o participante se considera fã do Universo Expandido.

Coluna: Do you consider yourself to be a fan of the Star Trek franchise?

- Tipo: varchar(50)

- Descrição: Pergunta se o participante é fã da franquia Star Trek.

Coluna: Gender

- Tipo: varchar(50)
- Descrição: Gênero do participante.
- Valores comuns: “Male”, “Female”.

Coluna: Age

- Tipo: varchar(50)
- Descrição: Faixa etária do respondente.
- Exemplos: “18–29”, “30–44”, “45–60”.

Coluna: Household Income

- Tipo: varchar(50)
- Descrição: Faixa de renda familiar.
- Exemplos:
  - “$0 – $24,999”
  - “$25,000 – $49,999”
  - “$100,000 – $149,999”

Coluna: Education

- Tipo: varchar(50)
- Descrição: Nível educacional.
- Exemplos:
  - “High school degree”
  - “Some college”
  - “Bachelor degree”
  - “Graduate degree“

Coluna: Location (Census Region)

- Tipo: varchar(50)
- Descrição: Região censitária dos EUA onde o participante reside.
- Exemplos:

- “South Atlantic”
- “Pacific”
- “West North Central”
- “Middle Atlantic”

3.Análise da Base, Ajustes e Indexação

A análise estrutural da tabela star_wars indicou uma série de problemas que comprometiam a integridade lógica do banco e dificultavam consultas e análises. Entre os principais problemas identificados, destacam-se a ausência normalização, colunas nas sem nomenclatura adequada, repetição de grupos de atributos, tipos incorretos, nenhuma chave primária e ausência de relacionamento. Com base nos problemas identificados, foi elaborado um processo de normalização, reorganizando a base em entidades distintas. O objetivo foi corrigir inconsistências, preservar os dados existentes e proporcionar maior integridade e desempenho.

DDL:

Tabela: region

- Descrição Geral: Armazena as regiões geográficas utilizadas para categorizar os respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único da região. Chave primária.
  - name (VARCHAR(100)): nome da região. Valor único.
- Chaves:
  - PK: id
- Justificativa: Centralizar as regiões evita inconsistências em nomes e permite relacionamentos eficientes com os respondentes.

Tabela: gender

- Descrição Geral: Armazena os gêneros possíveis dos respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único do gênero. Chave primária.
  - description (VARCHAR(50)): descrição do gênero, única.
- Chaves:
  - PK: id
- Justificativa: Normalização dos gêneros evita variações textuais e facilita consultas demográficas.

Tabela: age_group

- Descrição Geral: Define as faixas etárias utilizadas para classificar os respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único da faixa etária. Chave primária.
  - age_range_start (INT): início da faixa etária.
  - age_range_end (INT): fim da faixa etária.
- Chaves:
  - PK: id
- Justificativa: Facilita análises agrupadas por idade e mantém consistência na classificação etária.

Tabela: household_income

- Descrição Geral: Armazena as faixas de renda familiar dos respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único da faixa de renda. Chave primária.
  - income_range_start (INT): início da faixa de renda.
  - income_range_end (INT): fim da faixa de renda.
- Chaves:
  - PK: id
- Justificativa: Permite análises por renda sem inconsistências de formato textual.

Tabela: education_level

- Descrição Geral: Contém os níveis educacionais possíveis para os respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único do nível educacional. Chave primária.
  - name (VARCHAR(100)): descrição do nível educacional. Valor único.

- Chaves:
  - PK: id
- Justificativa: Normaliza os dados educacionais para padronizar relatórios e consultas.

Tabela: respondent

- Descrição Geral: Armazena os respondentes da pesquisa com referências para seus dados demográficos.
- Colunas:
  - id (BIGINT): identificador do respondente, chave primária.
  - gender_id (INT): referência para gender(id).
  - age_group_id (INT): referência para age_group(id).
  - household_income_id (INT): referência para household_income(id).
  - education_level_id (INT): referência para education_level(id).
  - region_id (INT): referência para region(id).
- Chaves:
  - PK: id
  - FK: gender_id → gender.id
  - FK: age_group_id → age_group.id
  - FK: household_income_id → household_income.id
  - FK: education_level_id → education_level.id
  - FK: region_id → region.id
- Justificativa: Centraliza os dados demográficos do respondente e padroniza relacionamentos para todas as respostas.

Tabela: question

- Descrição Geral: Armazena todas as perguntas da pesquisa.
- Colunas:
  - id (BIGINT GENERATED ALWAYS AS IDENTITY): identificador único da pergunta.
  - statement (VARCHAR(255)): texto da pergunta.
- Chaves:
  - PK: id

- Justificativa: Normaliza as perguntas para evitar redundâncias e possibilitar relacionamentos com respostas e opções.

Tabela: answer_option

- Descrição Geral: Contém as opções de resposta para cada pergunta.
- Colunas:
  - id (BIGINT GENERATED ALWAYS AS IDENTITY): identificador único da opção.
  - question_id (BIGINT): referência para a pergunta correspondente.
  - code (VARCHAR(30)): código único da opção dentro da pergunta.
  - label (VARCHAR(255)): descrição da opção.
- Chaves:
  - PK: id
  - FK: question_id → question.id
  - UNIQUE: (question_id, code)
- Justificativa: Permite múltiplas opções por pergunta sem redundância e garante consistência na codificação.

Tabela: answer

- Descrição Geral: Registra a resposta de cada respondente para cada pergunta.
- Colunas:
  - respondent_id (BIGINT): identificador do respondente.
  - question_id (BIGINT): identificador da pergunta.
  - option_id (BIGINT): identificador da opção selecionada.
- Chaves:
  - PK: (respondent_id, question_id)
  - FK: respondent_id → respondent.id
  - FK: question_id → question.id
  - FK: option_id → answer_option.id
- Justificativa: Estrutura completamente normalizada que evita colunas repetidas e permite análises detalhadas por pergunta e respondente.

Tabela: film

- Descrição Geral: Catálogo dos filmes avaliados pelos respondentes.
- Colunas:
  - id (INT GENERATED ALWAYS AS IDENTITY): identificador único do filme.
  - name (VARCHAR(100)): nome do filme, valor único.
- Chaves:
  - PK: id
- Justificativa: Centraliza os filmes para referência em visualizações e rankings, evitando redundância.

Tabela: film_seen

- Descrição Geral: Registra quais filmes foram assistidos por cada respondente.
- Colunas:
  - respondent_id (BIGINT): identificador do respondente.
  - film_id (INT): identificador do filme.
- Chaves:
  - PK: (respondent_id, film_id)
  - FK: respondent_id → respondent.id
  - FK: film_id → film.id
- Justificativa: Transformação de múltiplas colunas de filmes vistos em uma tabela relacional, garantindo normalização.

Tabela: film_ranking

- Descrição Geral: Armazena a classificação atribuída por cada respondente aos filmes.
- Colunas:
  - respondent_id (BIGINT): identificador do respondente.
  - film_id (INT): identificador do filme.
  - ranking (INT): posição do filme na preferência do respondente.
- Chaves:
  - PK: (respondent_id, film_id)

    - FK: respondent_id → respondent.id
    - FK: film_id → film.id
- Justificativa: Permite análises comparativas e rankings individuais de forma normalizada.

Tabela: character

- Descrição Geral: Catálogo de personagens avaliados pelos respondentes.
- Colunas:
    - id (INT GENERATED ALWAYS AS IDENTITY): identificador único do personagem.
    - name (VARCHAR(100)): nome do personagem. Valor único.
- Chaves:
- PK: id
    - Justificativa: Normaliza a lista de personagens para relacionar opiniões de forma consistente.

Tabela: character_opinion

- Descrição Geral: Armazena a opinião de cada respondente sobre cada personagem.
- Colunas:
    - respondent_id (BIGINT): identificador do respondente.
    - character_id (INT): identificador do personagem.
    - option_id (BIGINT): identificador da opção de opinião selecionada.
- Chaves:
    - PK: (respondent_id, character_id)
    - FK: respondent_id → respondent.id
    - FK: character_id → character.id
    - FK: option_id → answer_option.id
- Justificativa: Substitui múltiplas colunas de opinião em uma estrutura simples e totalmente normalizada, permitindo análises consistentes.

DML:

Inserção em gender

- Operação: Carregamento da tabela gender com todos os gêneros distintos presentes na base bruta star_wars.
- Descrição: Seleciona todos os valores distintos da coluna "gender" da base original, removendo espaços em branco e valores inválidos (NULL, '', Response).
- Justificativa: Normaliza os dados de gênero, evitando inconsistências de digitação. Permite referência consistente na tabela respondent.

Inserção em region

- Operação: Carregamento da tabela region com as regiões censitárias dos respondentes.
- Descrição: Seleciona valores distintos da coluna "Location (Census Region)", removendo espaços em branco e valores inválidos (NULL, '', Response).
- Justificativa: Padroniza nomes de regiões, permitindo consultas demográficas precisas e relacionamentos corretos com a tabela respondent.

Inserção em education_level

- Operação: Carregamento da tabela education_level com níveis educacionais distintos.
- Descrição: Seleciona valores distintos da coluna "education" da base bruta, eliminando espaços e valores inválidos (NULL, '', Response).
- Justificativa: Normaliza os níveis educacionais, evitando redundância e garantindo consistência na análise demográfica.

Inserção em age_group

- Operação: População da tabela age_group com faixas etárias predefinidas.
- Descrição: Insere intervalos de idade (18-29, 30-44, 45-60, 61-200) caso ainda não existam na tabela.
- Justificativa: Garante que todas as faixas etárias relevantes estejam presentes, permitindo vinculação consistente com os respondentes e análise agregada.

Inserção em household_income

- Operação: População da tabela household_income com faixas de renda.
- Descrição: Insere intervalos de renda (0–24.999, 25.000–49.999, 50.000–99.999, 100.000–149.999, 150.000–9.999.999) caso ainda não existam.
- Justificativa: Padroniza a classificação de renda, permitindo agrupamentos consistentes e análises econômicas.

Inserção em respondent

- Operação: Carregamento da tabela respondent com todos os participantes da base original.
- Descrição: Converte "respondentid" de float para BIGINT para preservar precisão. Faz join com gender, age_group, household_income, education_level e region para obter referências normalizadas. Ignora registros com "respondentid" nulo.
- Justificativa: Cria um identificador interno consistente, eliminando redundâncias e padronizando relações com respostas, filmes e opiniões sobre personagens.

Inserção em film

- Operação: Carregamento da tabela film com todos os títulos de filmes Star Wars.
- Descrição: Unifica títulos espalhados em várias colunas (Which of the following…, Unnamed: 4–8). Remove duplicidades e valores nulos ou em branco.
- Justificativa: Normaliza títulos de filmes para referência única em film_seen e film_ranking, evitando inconsistência textual e redundância.

Inserção em film_seen

- Operação: Transformação de colunas de filmes assistidos em formato relacional.
- Descrição:
- Para cada respondente, insere registros correspondentes a cada filme assistido, cruzando com a tabela film.

- Justificativa: Converte múltiplas colunas redundantes em estrutura 1:N, garantindo integridade e facilidade de consultas sobre hábitos de visualização.

Inserção em film_ranking

- Operação: Carregamento das classificações atribuídas aos filmes pelos respondentes.
- Descrição: Converte os valores de ranking (originais em colunas "Unnamed: 10–14") para INT. Associa cada ranking ao filme correto e ao respondente. Desconsidera valores nulos ou em branco.
- Justificativa: Normaliza rankings espalhados em múltiplas colunas, permitindo cálculos de média, comparação entre filmes e análise demográfica.

Inserção em character

- Operação: Carregamento da tabela character com todos os personagens avaliados.
- Descrição: Insere manualmente os nomes dos personagens (Han Solo, Luke Skywalker, etc.), garantindo IDs consistentes.
- Justificativa: Substitui colunas sem título da base original, centralizando todos os personagens em um catálogo único para uso em character_opinion.

Inserção em question

- Operação: Carregamento da tabela question com todas as perguntas da pesquisa.
- Descrição: Insere perguntas fixas, garantindo que não haja duplicidade com registros existentes.
- Justificativa: Normaliza todas as perguntas em uma tabela dedicada, permitindo relacionamentos consistentes com answer e answer_option.

Inserção em answer_option

- Operação: Carregamento das opções de resposta para cada pergunta.

- Descrição: Insere opções binárias (Yes, No) para perguntas de sim/não. Insere opções específicas para perguntas de personagens (Han, Greedo, I don’t understand this question). Insere opções de avaliação de personagem (Very favorably, Somewhat favorably, etc.).
- Justificativa: Normaliza as respostas possíveis, evitando inconsistência textual e possibilitando relacionamento seguro com a tabela answer.

Inserção em answer

- Operação: Carregamento das respostas dos respondentes para cada pergunta.
- Descrição: Associa cada respondente à opção de resposta correspondente na tabela answer_option. Filtra valores nulos, em branco ou inválidos (Response).
- Justificativa: Converte respostas espalhadas em múltiplas colunas em estrutura relacional 1:N, garantindo integridade, normalização e facilidade de análise.

Inserção em character_opinion

- Operação: Carregamento das opiniões dos respondentes sobre cada personagem.
- Descrição: Cruza cada personagem com cada respondente e associa à opção correta na tabela answer_option. Considera apenas valores válidos (não nulos e diferentes de Response).
- Justificativa: Substitui 14 colunas redundantes de opiniões por uma estrutura relacional totalmente normalizada, permitindo análises consistentes por personagem e respondente.

Índices:

- Índice: idx_age_group_start_end
- Tabela: age_group
- Colunas: age_range_start, age_range_end
- Descrição: Índice composto criado sobre o início e fim da faixa etária.

- Justificativa: Acelera consultas que filtram ou agrupam respondentes por faixa etária, especialmente para junções com a tabela respondent.

Índice: idx_household_income_start_end

- Tabela: household_income
- Colunas: income_range_start, income_range_end
- Descrição: Índice composto sobre os intervalos de renda familiar.
- Justificativa: Facilita filtragem e agregação por faixas de renda em análises demográficas, melhorando performance em junções com respondent.

Índice: idx_respondent_gender

- Tabela: respondent
- Coluna: gender_id
- Descrição: Índice sobre o gênero do respondente.
- Justificativa: Acelera consultas e filtros que segmentam respondentes por gênero, suportando agregações e análises estatísticas.

Índice: idx_respondent_age_group

- Tabela: respondent
- Coluna: age_group_id
- Descrição: Índice sobre a faixa etária do respondente.
- Justificativa: Facilita filtragem por idade e junções com a tabela age_group, otimizando relatórios e análises demográficas.

Índice: idx_respondent_income

- Tabela: respondent
- Coluna: household_income_id
- Descrição: Índice sobre a faixa de renda do respondente.
- Justificativa: Permite consultas rápidas por faixa de renda e melhora performance de relatórios analíticos agregados.

Índice: idx_respondent_education

- Tabela: respondent

- Coluna: education_level_id
- Descrição: Índice sobre o nível educacional do respondente.
- Justificativa: Acelera filtros e agregações baseadas em educação, essencial para cruzamentos demográficos e estatísticos.

Índice: idx_respondent_region

- Tabela: respondent
- Coluna: region_id
- Descrição: Índice sobre a região do respondente.
- Justificativa: Facilita pesquisas e relatórios por região, melhorando desempenho em análises geográficas e demográficas.

Índice: idx_answer_option_question

- Tabela: answer_option
- Coluna: question_id
- Descrição: Índice sobre a pergunta associada à opção de resposta.
- Justificativa: Acelera junções e filtragens de respostas por pergunta, importante em consultas de avaliação de dados de pesquisa.

Índice: idx_answer_option_id

- Tabela: answer
- Coluna: option_id
- Descrição: Índice sobre a opção selecionada pelo respondente.
- Justificativa: Facilita consultas e agregações por opção de resposta, suportando análises de distribuição e preferências.

Índice: idx_answer_question_id

- Tabela: answer
- Coluna: question_id
- Descrição: Índice sobre a pergunta respondida.
- Justificativa: Acelera filtragem por pergunta, crucial em relatórios de questionário e cruzamentos com answer_option.

Índice: idx_film_seen_respondent

- Tabela: film_seen
- Coluna: respondent_id
- Descrição: Índice sobre o respondente que assistiu ao filme.
- Justificativa: Facilita consultas que listam filmes vistos por um participante específico, otimizando análise individual e agregada.

Índice: idx_film_seen_film

- Tabela: film_seen
- Coluna: film_id
- Descrição: Índice sobre o filme assistido.
- Justificativa: Acelera consultas e contagens de espectadores por filme, suportando métricas de popularidade.

Índice: idx_film_ranking_respondent

- Tabela: film_ranking
- Coluna: respondent_id
- Descrição: Índice sobre o respondente que atribuiu ranking ao filme.
- Justificativa: Facilita filtragem e junção em consultas sobre rankings individuais ou análises demográficas.

Índice: idx_film_ranking_film

- Tabela: film_ranking
- Coluna: film_id
- Descrição: Índice sobre o filme ranqueado.
- Justificativa: Otimiza consultas agregadas e cálculo de médias de ranking, permitindo análises rápidas de preferência por filme.

Índice: idx_character_opinion_respondent

- Tabela: character_opinion
- Coluna: respondent_id
- Descrição: Índice sobre o respondente que expressou opinião sobre o personagem.

- Justificativa: Facilita consultas e relatórios que analisam opiniões por respondente.

Índice: idx_character_opinion_character

- Tabela: character_opinion
- Coluna: character_id
- Descrição: Índice sobre o personagem avaliado.
- Justificativa: Permite consultas rápidas de opinião agregada por personagem, usado em rankings de popularidade ou análise de percepção.

Índice: idx_character_opinion_option

- Tabela: character_opinion
- Coluna: option_id
- Descrição: Índice sobre a opção de opinião escolhida.
- Justificativa: Acelera análises e relatórios que filtram respondentes por tipo de avaliação (favorável, neutra, desfavorável).

4.Criação de Automatizações no PostgreSQL

FUNCTIONS:

contar_filmes_vistos(p_respondent_id BIGINT)

- Descrição: Retorna o total de filmes vistos por um respondente, verificando antes se o respondente existe na tabela respondent.
- Parâmetros: p_respondent_id (BIGINT): identificador do respondente cujo histórico de filmes será contado.
- Retorno: INT: número de filmes vistos pelo respondente.
- Importância: Permite medir engajamento real do participante. Usada em relatórios analíticos, validações internas e cruzamentos com outras métricas de participação.

obter_ranking_medio_filme(p_film_id INT)

- Descrição: Calcula o ranking médio atribuído a um filme pelos respondentes, desconsiderando valores nulos.
- Parâmetros: p_film_id (INT): identificador do filme cuja média de ranking será calculada.
- Retorno: NUMERIC: média dos rankings atribuídos ao filme. Retorna 0 se não houver avaliações.
- Importância: Fornece estatísticas fundamentais para dashboards e relatórios de preferência. Permite comparações entre filmes e análise de popularidade.

eh_fan_star_wars(p_respondent_id BIGINT)

- Descrição: Verifica se um respondente declarou ser fã da franquia Star Wars, considerando as respostas registradas na tabela answer e answer_option.
- Parâmetros: p_respondent_id (BIGINT): identificador do respondente a ser verificado.
- Retorno: BOOLEAN: TRUE se o respondente é fã, FALSE caso contrário ou se a pergunta não existir.
- Importância: Segmenta análises entre fãs e não-fãs. Crucial para entender padrões de opinião, comportamento e engajamento com a franquia.

PROCEDURES:

inserir_respondente_com_validacao()

- Descrição: Insere um respondente na base garantindo que ele exista na tabela principal (respondent) e que todos os dados associados estejam consistentes.
- Importância: Evita registros órfãos ou inconsistentes. Uniformiza o processo de cadastro de novos respondentes, garantindo integridade referencial e confiabilidade dos dados.

Procedure: atualizar_opiniao_personagem_lote()

- Descrição: Atualiza em lote as opiniões de personagens já registradas, substituindo valores antigos por novos, conforme necessidade de correção ou padronização.
- Importância: Facilita ajustes massivos de dados sem a necessidade de atualizações manuais individualizadas. Mantém a consistência das informações e agiliza processos de correção de registros históricos.

Procedure: limpar_respondente()

- Descrição: Remove todas as respostas associadas a um respondente específico, limpando registros de tabelas relacionadas.
- Importância: Permite exclusão segura de dados pessoais mediante requisição, garantindo auditoria e rastreabilidade do processo.

TRIGGERS:

Trigger: trg_validar_ranking

- Descrição: Impede a inserção ou atualização de um ranking de filme fora do intervalo válido (1 a 6) na tabela film_ranking.
- Importância: Garante integridade dos dados relacionados a avaliações de filmes. Evita distorções estatísticas ou registros inválidos em análises de ranking.

Trigger: trg_validar_answer_option

- Descrição: Valida que o option_id informado na tabela answer pertença à pergunta correspondente (question_id). Bloqueia inserções ou atualizações caso haja inconsistência.
- Importância: Mantém consistência entre respostas e opções válidas. Evita registros que poderiam corromper relatórios ou análises de respostas individuais.

Trigger: trg_validar_character_opinion

- Descrição: Garante que somente opções da pergunta “Character opinion” sejam inseridas na tabela character_opinion. Bloqueia registros que não correspondam à pergunta específica.
- Importância: Evita inconsistências e registros inválidos sobre opiniões de personagens. Mantém integridade do relacionamento entre respondentes, personagens e respostas categorizadas.

VIEWS:

v_respondentes_por_regiao

- Descrição: Apresenta estatísticas de respondentes agrupados por região, incluindo contagem total e percentual em relação ao total de respondentes. Valores nulos ou não informados são tratados como “Não informado”.
- Importância: Fundamental para estudos de distribuição geográfica e representatividade da pesquisa. Facilita dashboards e relatórios de segmentação regional.

View: v_ranking_medio_filmes

- Descrição: Combina dados do catálogo de filmes com rankings fornecidos pelos respondentes. Mostra total de rankings, média, melhor e pior classificação para cada filme.
- Importância: Permite avaliações comparativas e insights sobre preferências de filmes. Base para análises de popularidade, relatórios e decisões estratégicas sobre a franquia.

View: v_fans_vs_nao_fans

- Descrição: Compara respondentes que se declaram fãs ou não fãs de Star Wars, relacionando com a visualização de filmes e interesse em Star Trek. Inclui total de respondentes, percentual, quantos viram algum filme e quantos também são fãs de Star Trek.
- Importância: Crucial para análises comportamentais e segmentação do público. Suporta relatórios estratégicos, campanhas de marketing e estudos de correlação entre franquias.

5.Modelagem do Data Warehouse (DW)

Nesta modelagem, os dados são organizados em uma tabela fato, que centraliza os eventos mensuráveis, e em tabelas dimensão, que fornecem contexto descritivo e categórico para esses eventos. O tema do Data Warehouse é a opinião e comportamento dos respondentes em relação a filmes e personagens, além da identificação de fãs de determinados universos, como Star Wars e Star Trek. Assim, o Data Warehouse é capaz de representar e analisar informações relacionadas a preferências, hábitos de consumo de mídia e percepções de personagens.

DDL:

Tipo Enumerado: action_type

- Descrição: Criação de um tipo enumerado (I, U, D) para registrar operações de inclusão, atualização e exclusão.
- Justificativa: Permite rastrear transformações e controlar a qualidade dos dados carregados no DW, garantindo auditoria de mudanças em dimensões e fatos.

Dimensão Respondente (dim_respondent)

- Descrição: Contém informações demográficas dos respondentes, incluindo gênero (gender), faixa etária (age_group), renda familiar (household_income), escolaridade (education) e região (region).
- Chaves: id (PK), respondent_id (único).
- Justificativa: Permite segmentar análises por características sociais, essenciais para interpretar preferências, comportamentos e padrões de consumo.

Dimensão Filme (dim_film)

- Descrição: Lista de filmes avaliados pelos respondentes, com identificador único (film_id) e nome (film_name).
- Chaves: id (PK), film_id (único).
- Justificativa: Estrutura fundamental para análises de popularidade, consumo e comparação entre títulos.

Dimensão Personagem (dim_character)

- Descrição: Armazena personagens avaliados pelos respondentes, incluindo identificador (character_id) e nome (character_name).
- Chaves: id (PK), character_id (único).
- Justificativa: Permite mensurar aprovação, rejeição e padrões de percepção do público sobre personagens específicos.

Dimensão Questão (dim_question)

- Descrição: Contém as questões aplicadas aos respondentes, com identificador (question_id) e enunciado (statement).
- Chaves: id (PK), question_id (único).
- Justificativa: Estrutura para análise de respostas, permitindo relacionar preferências ou opiniões com características demográficas e outros fatos.

Dimensão Opção de Resposta (dim_answer_option)

- Descrição: Armazena opções de resposta para cada questão, incluindo identificador (option_id), código (code), label (label) e referência à questão (question_id).
- Chaves: id (PK), option_id (único).
- Justificativa: Estrutura essencial para interpretar respostas, facilitar cruzamentos com fatos e possibilitar análises detalhadas de comportamento e preferência.

Tabela Fato Respostas (fact_response)

- Descrição: Registra interações entre respondentes, filmes, personagens, questões e opções de resposta, incluindo indicadores como: seen (assistiu), ranking (classificação), opiniões ou escolhas de fãs.
- Chaves: id (PK).
- Relacionamentos:
  - respondent_id → dim_respondent
  - question_id → dim_question
  - option_id → dim_answer_option
  - film_id → dim_film

    - character_id → dim_character
- Justificativa: Consolida todas as interações analisáveis, servindo como base central para cruzamento das dimensões, análises de comportamento e padrões de consumo.

Índices da Tabela Fato

- Colunas Indexadas: respondent_id, question_id, option_id, film_id, character_id, seen, ranking.
- Justificativa: Melhoram significativamente o desempenho das consultas analíticas, principalmente filtros e agregações frequentes em relatórios de BI.

Tabela de Controle de ETL (etl_execution)

- Descrição: Registra o nome do processo e a data/hora da última execução do ETL.
- Chaves: process (PK).
- Justificativa: Suporte operacional, garantindo integridade e rastreabilidade das cargas de dados no DW.

DML:

Carga da Dimensão Respondente (dim_respondent)

- Descrição: Selecionam-se respondentes únicos do sistema fonte, extraindo dados demográficos, como gênero, faixa etária, renda familiar, escolaridade e região.
- Transformações:
    - Faixa etária combinada no formato start-end.
    - Renda familiar combinada no formato start-end.
    - Gênero, nível educacional e região associados aos identificadores originais.
- Tratamento de duplicidade: Registros já existentes não são reinseridos (ON CONFLICT DO NOTHING).
- Função: Garantir que cada respondente possua descrição única e consistente, permitindo segmentações confiáveis para análise.

Carga da Dimensão Filme (dim_film)

- Descrição: Todos os filmes do sistema original são listados e inseridos na dimensão.
- Tratamento de duplicidade: Evita duplicações usando ON CONFLICT DO NOTHING sobre film_id.
- Função: Padronizar a lista oficial de filmes, servindo como referência central para análises de popularidade, consumo e comparações entre títulos.

Carga da Dimensão Personagem (dim_character)

- Descrição: Personagens avaliados pelos respondentes são extraídos do sistema fonte e adicionados à dimensão.
- Tratamento de duplicidade: Mantém unicidade via ON CONFLICT DO NOTHING sobre character_id.
- Função: Garantir correspondência correta entre opiniões e personagens, permitindo análises detalhadas de aprovação, rejeição e percepção do público.

Carga da Dimensão Questão (dim_question)

- Descrição: Questões aplicadas aos respondentes são carregadas para o DW, mantendo identificador único e enunciado.
- Tratamento de duplicidade: Registros duplicados são ignorados (ON CONFLICT DO NOTHING).
- Função: Estrutura essencial para relacionar respostas com respondentes, filmes e personagens nas análises multidimensionais.

Carga da Dimensão Opção de Resposta (dim_answer_option)

- Descrição: Todas as opções de resposta para cada questão são extraídas e inseridas na dimensão.
- Transformações: Mantida relação com question_id e armazenados código e label da opção.
- Tratamento de duplicidade: ON CONFLICT DO NOTHING sobre option_id.

- Função: Facilitar cruzamentos de respostas, padronizando opções para análises detalhadas de comportamento e opinião.

Carga da Tabela Fato Respostas (fact_response)

- Descrição: Integra dados operacionais das tabelas de respondentes, filmes assistidos, rankings e opiniões de personagens.
- Transformações aplicadas:
  - Conversão de respostas textuais em booleanas (seen).
  - Associação de ranking ao filme correto.
  - Consolidação de opiniões sobre personagens via option_id.
- Tratamento de duplicidade: Evita inserções repetidas com ON CONFLICT DO NOTHING.
- Resultado: Cada linha representa uma interação completa entre respondente, filme e personagem, pronta para análises multidimensionais, permitindo cruzamentos entre todas as dimensões e métricas do DW.

Perguntas de Negócio Respondidas pelo DW:

Codificação de Valores e Categorias:

a) Opiniões (opinion)

- VF = Very Favorably (Muito favorável)
- SF = Somewhat Favorably (Favorável)
- N = Neutral (Neutro)
- SU = Somewhat Unfavorably (Desfavorável)
- VU = Very Unfavorably (Muito desfavorável)

b) Categoria de Idade (age_group)

- YA = 18–29 anos
- AD = 30–44 anos
- MA = 45–60 anos
- SR = >60 anos

c) Renda Familiar (household_income)

- L = Menos de $25,000

- LM = $25,000 – $49,999
- M = $50,000 – $99,999
- UM = $100,000 – $149,999
- H = $150,000+

d) Escolaridade (education)

- LHS = Less than High School (Menos que Ensino Médio)
- HS = High School (Ensino Médio)
- AS = Associate's Degree (Curso técnico ou tecnólogo)
- BA = Bachelor's Degree (Graduação)
- GR = Graduate Degree (Pós-graduação)
- 5) Gênero (gender)
- M = Male (Masculino)
- F = Female (Feminino)

1) Quais personagens têm as opiniões mais positivas?

- Origem: fact_response.opinion + dim_character.character_name
- Insight: Contabiliza elogios por personagem (VF + SF). Permite comparar aprovação por faixa etária, gênero ou renda. Identifica personagens com maior potencial comercial e narrativo.

2) Personagens mais rejeitados

- Origem: fact_response.opinion (SU + VU)
- Insight: Identifica rejeições e padrões negativos de percepção por personagem.

3) Personagens mais desconhecidos

- Origem: fact_response.opinion = 'U'
- Insight: Avalia familiaridade do público com cada personagem.

4) Personagem mais amado por faixa etária

- Origem: fact_response + dim_respondent.age_group
- Insight: Analisa preferências específicas por grupo etário, permitindo segmentações estratégicas.

5) Personagem mais rejeitado entre fãs de Star Wars

- Origem: fact_response.fan_star_wars = TRUE
- Insight: Entende rejeições específicas em segmentos de fãs, auxiliando em campanhas direcionadas.

6) Diferença de opinião por gênero

- Origem: fact_response + dim_respondent.gender
- Insight: Permite analisar polarização de opiniões entre homens e mulheres.

7) Escolaridade influencia opinião sobre Darth Vader?

- Origem: fact_response.opinion + dim_respondent.education
- Insight: Avalia impacto da escolaridade na percepção de personagens icônicos.

8) Filme mais assistido

- Origem: fact_response.seen = TRUE
- Insight: Identifica os títulos com maior audiência.

9) Ranking médio e mediano dos filmes

- Origem: fact_response.ranking
- Insight: Mede popularidade relativa e consistência de classificação dos filmes.

10) Filme favorito por faixa etária

- Origem: fact_response.ranking = 1 + dim_respondent.age_group
- Insight: Segmenta preferência de títulos por idade.

11) Popularidade dos filmes por região

- Origem: fact_response.seen = TRUE + dim_respondent.region
- Insight: Analisa popularidade regional, útil para estratégias de marketing localizadas.

12) Correlação entre opinião de personagens e ranking dos filmes

- Origem: fact_response.opinion transformada em score (-2 a 2)
- Insight: Identifica se avaliações de personagens influenciam a classificação dos filmes.

13) Filme mais popular entre nível de escolaridade específico

- Origem: dim_respondent.education + fact_response.seen
- Insight: Descobre preferências por faixa educacional.

14) Distribuição de respondentes por renda

- Origem: dim_respondent.household_income
- Insight: Entende composição socioeconômica do público.

15) Fãs de Star Wars e Star Trek

- Origem: fact_response.fan_star_wars / fact_response.fan_star_trek
- Insight: Classifica respondentes em quatro grupos: fãs de ambos, de um só ou de nenhum. Quantifica sobreposição entre fandoms, auxiliando em estratégias de mercado e conteúdo cruzado.

16) Influência de renda e faixa etária em consumo

- Origem: fact_response.seen + dim_respondent.age_group / household_income
- Insight: Avalia quais grupos veem mais filmes e identificam padrões de comportamento.

17) Personagens favoritos e rejeitados por segmentos

- Origem: fact_response.opinion + dim_respondent.household_income / fandom
- Insight: Permite análises segmentadas de aprovação, rejeição e controvérsia, auxiliando na estratégia de merchandising e marketing.

TRIGGERS:

dim_answer_option

- Função: fn_dim_answer_option

- Tabela de origem: public.answer_option
- Tabela de destino: dw.dim_answer_option
- Operações capturadas: INSERT, UPDATE, DELETE
- Justificativa: Permite acompanhar alterações nas opções de resposta, mantendo histórico e integridade da dimensão. Garante que cada opção seja atualizada ou marcada como excluída (action = D) quando necessário.
- Perguntas de negócio atendidas:
  - Quais respostas têm maior adesão ou rejeição?
  - Há alterações na popularidade ou uso das opções ao longo do tempo?

dim_character

- Função: fn_dim_character
- Tabela de origem: public.character
- Tabela de destino: dw.dim_character
- Operações capturadas: INSERT, UPDATE, DELETE
- Justificativa: Permite acompanhar alterações nos personagens avaliados. Mantém histórico de inclusão (I), atualização (U) e exclusão (D) na dimensão.
- Perguntas de negócio atendidas:
  - Quais personagens têm as opiniões mais positivas ou negativas?
  - Há mudanças na percepção do público ao longo do tempo?

dim_film

- Função: fn_dim_film
- Tabela de origem: public.film
- Tabela de destino: dw.dim_film
- Operações capturadas: INSERT, UPDATE, DELETE
- Justificativa: Garante que todos os filmes avaliados estejam corretamente registrados. Mantém integridade e histórico da dimensão de filmes.
- Perguntas de negócio atendidas:
  - Quais filmes são mais populares por faixa etária ou região?
  - Como varia a classificação média e mediana dos filmes ao longo do tempo?

dim_question

- Função: fn_dim_question
- Tabela de origem: public.question
- Tabela de destino: dw.dim_question
- Operações capturadas: INSERT, UPDATE, DELETE
- Justificativa: Atualiza a dimensão de perguntas garantindo que alterações sejam refletidas. Mantém histórico de mudanças nas questões avaliadas.
- Perguntas de negócio atendidas:
    - Quais perguntas possuem maior engajamento?
    - Há alterações nas respostas ao longo do tempo?

dim_respondent

- Função: fn_dim_respondent
- Tabela de origem: public.respondent
- Tabela de destino: dw.dim_respondent
- Operações capturadas: INSERT, UPDATE, DELETE
- Justificativa: Permite segmentar análises por características demográficas: idade, gênero, renda, escolaridade e região. Mantém histórico de inclusão, atualização e exclusão dos respondentes.
- Perguntas de negócio atendidas:
- Como o comportamento dos respondentes varia conforme idade, gênero ou região?
- Existe correlação entre características demográficas e fandom (Star Wars/Star Trek)?

fact_response

- Função: fn_fact_response
- Tabelas de origem: public.answer, public.film_seen, public.film_ranking, public.character_opinion
- Tabela de destino: dw.fact_response
- Operações capturadas: INSERT, UPDATE, DELETE

- Justificativa: Consolida todas as interações entre respondentes, filmes e personagens em uma tabela fato única. Mantém histórico das respostas, rankings e opiniões, permitindo rastreabilidade (action = I/U/D).
- Perguntas de negócio atendidas:
- Quais filmes são mais assistidos por faixa etária?
- Quantos fãs de Star Wars também são fãs de Star Trek?
- Quais personagens possuem maior aprovação ou rejeição?
- Correlação entre opinião sobre personagens e ranking de filmes.

6.Considerações finais

O desenvolvimento do projeto permitiu compreender de forma aprofundada a importância da modelagem correta de um banco de dados e dos riscos associados ao uso de fontes não normalizadas. A organização da base inicial, que apresentava inconsistências e estruturas pouco adequadas, possibilitou a criação de um modelo mais eficiente, coerente e alinhado às boas práticas de sistemas relacionais. Essa transformação garantiu maior clareza sobre o domínio e facilitou a posterior aplicação de rotinas de análise e processamento.

A migração dos dados preservou integralmente as informações relevantes e restabeleceu relações consistentes entre respondentes, filmes e personagens. Ademais, a criação de índices otimizou o desempenho das consultas, especialmente em operações de cruzamentos estatísticos e exploração analítica. O uso de automatizações no PostgreSQL contribuiu para a integridade do sistema, a padronização de operações e a automação de tarefas rotineiras.

De forma geral, o trabalho evidencia a relevância da normalização, da definição rigorosa de chaves, da documentação técnica e da modelagem adequada para assegurar qualidade e eficiência em projetos de dados. Demonstra também como uma base simples, mas mal estruturada, pode ser transformada em um ambiente robusto, preparado para análises avançadas, construção de Data Warehouses e integração com pipelines de ETL. Assim, reforça-se a importância da engenharia de dados como disciplina essencial para a geração de valor por meio da informação.

7.Referências Bibliográficas

KAGGLE. Official Crime Data – São Paulo State (Brazil) – SSP. Disponível em: https://www.kaggle.com/datasets/dbwaller/official-crime-data-sao-paulo-statebrazil-ssp/data. Acesso em: 30 nov. 2025.