1. Sistemas Operacionais

O sistema em estudo acaba por se caracterizar na fragmentação de uma base de dados em três sistemas operacionais distintos, cada um responsável por armazenar uma parte específica da informação. As sessões de visualização dos utilizadores são registadas num ficheiro CSV. Por sua vez, os dados relativos aos conteúdos assistidos estão armazenados na base de dados MySQL e as informações dos utilizadores, como dados pessoais e estado da subscrição, na base de dados PostgreSQL1. Existe ainda um quarto sistema operacional PostgreSQL2 que agrega o conjunto das estruturas armazenadas nos outros três sistemas operacionais, mas com dados diferentes. É ainda importante referir que certos campos admitem apenas um conjunto limitado de valores válidos.

Para possibilitar a deteção de alterações e a atualização dos dados durante o processo de ETL, foi adicionado o campo "is_up_to_date" a todas as tabelas operacionais. Este campo assume o valor 0 por defeito e é atualizado para 1 assim que o registo for processado no ETL. Foram também definidos *triggers* para garantir que qualquer atualização nos dados redefina automaticamente este campo para 0, sinalizando a necessidade de novo processamento. Vale salientar que este campo não é apresentado nas tabelas incluídas neste dicionário.

1.1. CSV

Como dito anteriormente, o ficheiro CSV contém a informação relativa às sessões de visualização de conteúdo por parte dos utilizadores. Encontra-se na Tabela 1 a apresentação e explicação da sua estrutura.

Campo	Tipo	Descrição
session_code	VARCHAR (16)	Identificador da sessão
user_code	VARCHAR (16)	Identificador do utilizador que assistiu à sessão
content_code	VARCHAR (16)	Identificador do conteúdo que foi assistido
time	TIMESTAMP	Quando é que a sessão se realizou
watched_duration	INT	Qual a duração da sessão em minutos
platform	VARCHAR (20)	Categoria geral do dispositivo utilizado para assistir à sessão
device_type	VARCHAR (20)	Categoria física do dispositivo utilizado para assistir à sessão
os_family	VARCHAR (20)	Família a que pertence o sistema operativo do dispositivo
os_name	VARCHAR (20)	O nome do sistema operativo do dispositivo
app_version	VARCHAR (20)	A versão da aplicação utilizada durante a sessão

Tabela 1 - Estrutura do Ficheiro CSV

Domínio de valores admitidos por campo:

- platform: "Mobile", "Computer", "TV", "Tablet" e "Console";
- device_type: "Desktop", "Laptop", "Smartphone", "Tablet", "Smart TV" e "Game Console";
- os_family: "Windows", "macOS", "Linux", "Android", "iOS", "tvOS", "FireOS", "Playstation", "Nintendo" e "Xbox";

- os_name: “Android 13”, “iOS 16”, “iPadOS 15”, “FireOS 6”, “tvOS 14”, “Windows 11”, “macOS Ventura”, “Ubuntu 22.04”, “PS5”, “Switch” e “Xbox”;

- app_version: “v1.0.0”, “v1.2.3”, “v2.0.1”, “v2.1.0” e “v3.0.5”.

1.2. MySQL

Na base de dados MySQL (ver Figura 1) é contida toda a informação sobre o conteúdo disponível na StreamFlix. A tabela “contents” armazena a informação sobre cada conteúdo (ver Tabela 5). Através do relacionamento com as outras tabelas, os conteúdos dispõem também de um tipo (ver Tabela 2), classificação etária (ver Tabela 3), diretor (ver Tabela 4) e ainda de um ou mais géneros (ver Tabela 6 e Tabela 7).

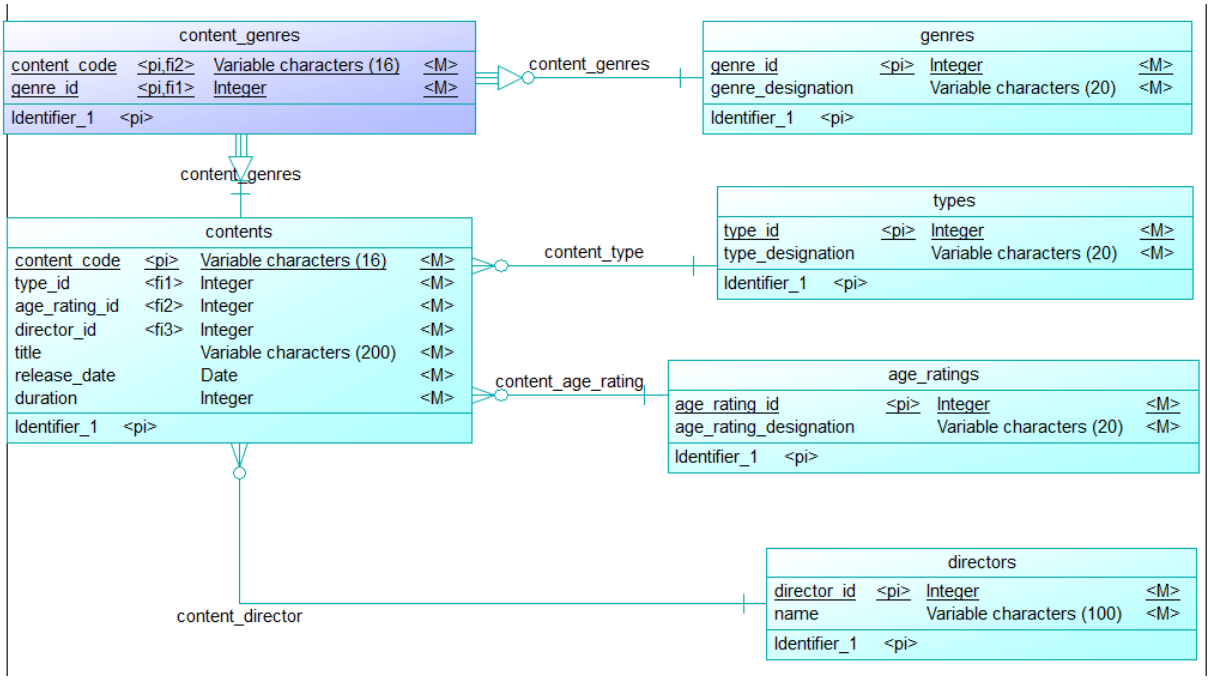

Figura 1 - Modelo Lógico do MySQL

1.2.1. Tabela “types”

Campo	Tipo	Descrição
type_id	INT	Identificador do tipo de conteúdo
type_designation	VARCHAR (20)	Designação textual do tipo de conteúdo

Tabela 2 - Estrutura da Tabela "types" do MySQL

Valores admitidos pelo campo “type_designation”:

- “Movie”, “Series”, “Short Film”, “Documentary” e “Special”.

1.2.2. Tabela “age_ratings”

Campo	Tipo	Descrição
age_rating_id	INT	Identificador da classificação etária
age_rating_designation	VARCHAR (20)	Designação textual da classificação etária

Tabela 3 - Estrutura da Tabela "age_ratings" do MySQL

Valores admitidos pelo campo “age_rating_designation”:

- “G”, “PG”, “PG-13”, “R” e “NC-17”.

1.2.3. Tabela “directors”

Campo	Tipo	Descrição
director_id	INT	Identificador do diretor
name	VARCHAR (100)	Nome do diretor

Tabela 4 - Estrutura da Tabela "directors" do MySQL

1.2.4. Tabela “contents”

Campo	Tipo	Descrição
content_code	VARCHAR (16)	Identificador do conteúdo
type_id	INT	Identificador do tipo de conteúdo associado
age_rating_id	INT	Identificador da classificação etária associada
director_id	INT	Identificador do diretor associado
title	VARCHAR (200)	Título do conteúdo
release_date	DATE	Data de lançamento do conteúdo
duration	INT	Duração do conteúdo (em minutos)

Tabela 5 - Estrutura da Tabela "contents" do MySQL

1.2.5. Tabela “genres”

Campo	Tipo	Descrição
genre_id	INT	Identificador do género de conteúdo
genre _designation	VARCHAR (20)	Designação textual do género de conteúdo

Tabela 6 - Estrutura da Tabela "genres" do MySQL

Valores admitidos pelo campo “genre _designation”:

- “Action”, “Comedy”, “Drama”, “Thriller”, “Horror”, “Romance”, “Documentary”, “Animation”, “Sci-Fi” e “Fantasy”.

1.2.6. Tabela “content_genres”

Campo	Tipo	Descrição
content_code	VARCHAR (16)	Identificador do conteúdo
genre_id	INT	Identificador do género

Tabela 7 - Estrutura da Tabela "content_genres" do MySQL

1.3. PostgreSQL1

Quanto à base de dados PostgreSQL1 (ver Figura 2), está armazenada a informação relativa a cada um dos utilizadores da plataforma, na tabela “users” (ver Tabela 12). A mesma relaciona-se com quatro outras tabelas, o que leva a que os utilizadores estejam associados a um grupo etário (ver Tabela 8), sexo (ver Tabela 9), país (ver Tabela 10) e estado da subscrição (ver Tabela 11).

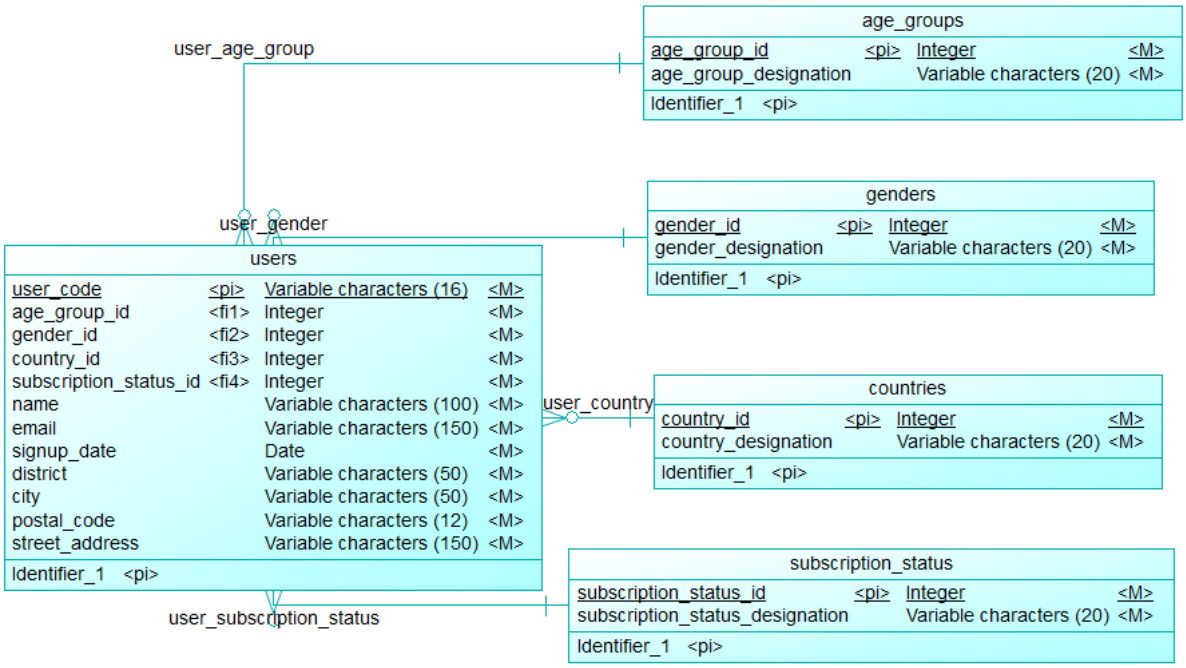

Figura 2 - Modelo Lógico do PostgreSQL1

1.3.1. Tabela “age_groups”

Campo	Tipo	Descrição
age_group_id	INT	Identificador do grupo etário
age_group_designation	VARCHAR (20)	Designação textual do grupo etário

Tabela 8 - Estrutura da Tabela "age_groups" do PostgreSQL1

Valores admitidos pelo campo “age_group_designation”:

- “0-9”, “10-14”, “15-19”, “20-24”, “25-34”, “35-44”, “45-54”, “55-64” e “65+”.

1.3.2. Tabela “genders”

Campo	Tipo	Descrição
gender_id	INT	Identificador do sexo do utilizador
gender_designation	VARCHAR (20)	Designação textual do sexo do utilizador

Tabela 9 - Estrutura da Tabela "genders" do PostgreSQL1

Valores admitidos pelo campo “gender_designation”:

- “Male”, “Female” e “Other”.

1.3.3. Tabela “countries”

Campo	Tipo	Descrição
country_id	INT	Identificador do país
country_designation	VARCHAR (20)	Designação textual do país

Tabela 10 - Estrutura da Tabela "countries" do PostgreSQL1

Valores admitidos pelo campo “country_designation”:

- “Portugal”, “Spain”, “France”, “Germany”, “Italy”, “Netherlands”, “United Kingdom”, “United States”, “Brazil”, “Canada” e “Venezuela”.

1.3.4. Tabela “subscription_status”

Campo	Tipo	Descrição
subscription_status_id	INT	Identificador do estado da subscrição
subscription_status_designation	VARCHAR (20)	Designação textual do estado da subscrição

Tabela 11 - Estrutura da Tabela "subscription_status" do PostgreSQL1

Valores admitidos pelo campo “subscription_status_designation”:

- “Active”, “Cancelled” e “Expired”.

1.3.5. Tabela “users"

Campo	Tipo	Descrição
user_code	VARCHAR (16)	Identificador do utilizador
age_group_id	INT	Identificador do grupo etário associado
gender_id	INT	Identificador do sexo associado
country_id	INT	Identificador do país associado
subscription_status_id	INT	Identificador do estado da subscrição associado
name	VARCHAR (100)	Nome do utilizador
email	VARCHAR (150)	Email do utilizador
signup_date	DATE	Data de inscrição na plataforma
district	VARCHAR (50)	Distrito do utilizador
city	VARCHAR (50)	Cidade do utilizador
postal_code	VARCHAR (12)	Código Postal do utilizador
street_address	VARCHAR (150)	Endereço da rua do utilizador

Tabela 12 - Estrutura da Tabela "users" do PostgreSQL1

1.4. PostgreSQL2

A base de dados PostgreSQL2 (ver Figura 3) representa a agregação das estruturas dos três sistemas operacionais anteriores (CSV, MySQL e PostgreSQL1), concentrando num único sistema toda a informação sobre sessões, conteúdos e utilizadores. No entanto, existem algumas

diferenças de nomenclatura. A tabela “genres” é substituída pela “categories”, o que por sua vez conduz à renomeação de “content_genres” para “content_categories”. Da mesma forma, a tabela “age_ratings”, passa a designar-se “age_restrictions”. Além disso, certos campos admitem conjuntos de valores diferentes face aos anteriores.

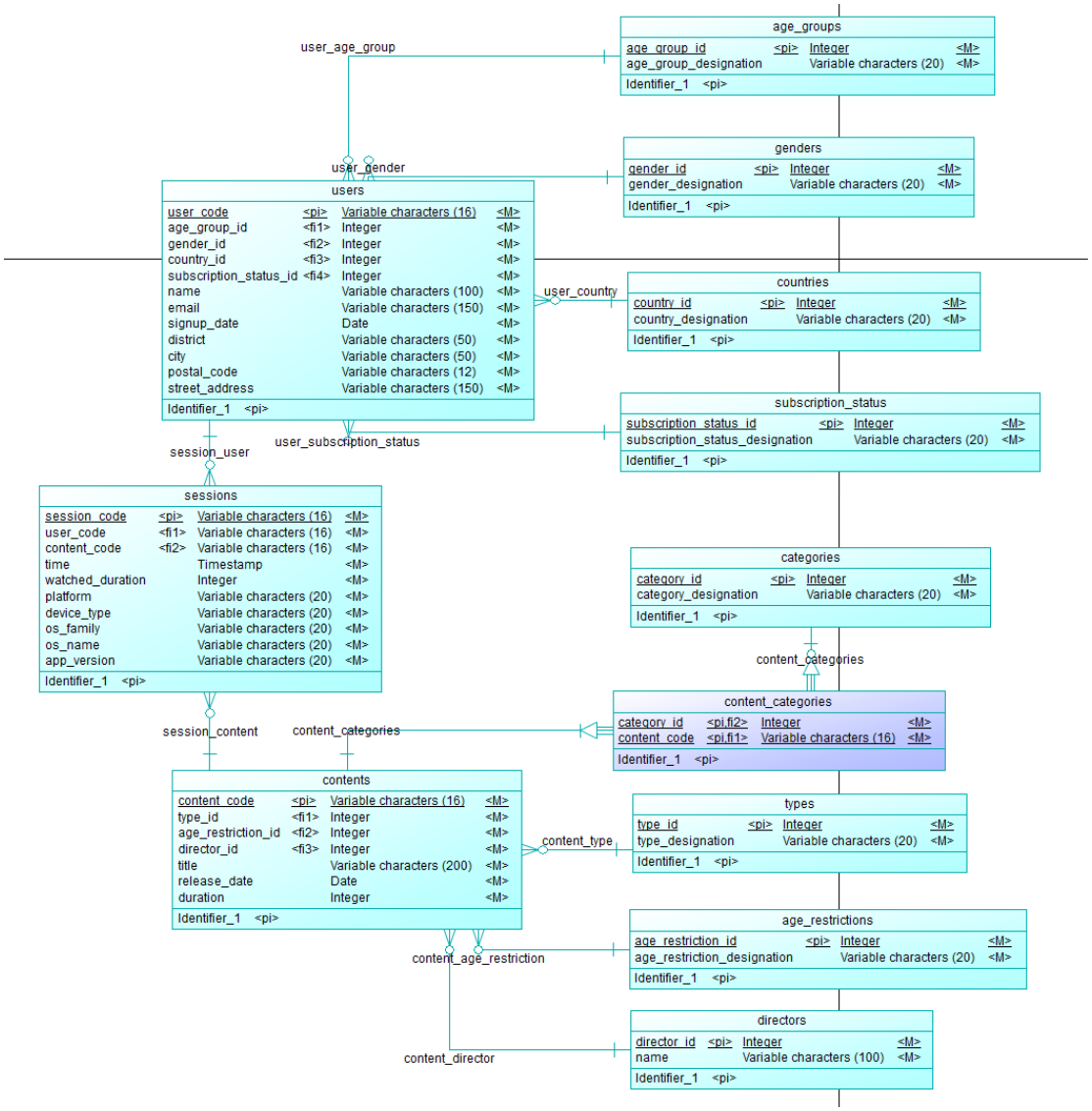

Figura 3 - Modelo Lógico do PostgreSQL2

1.4.1. Tabela “genders”

Valores admitidos pelo campo “gender_designation”:

- “Man”, “Woman” e “Prefer not to say”.

1.4.2. Tabela “subscription_status”

Valores admitidos pelo campo “subscription_status_designation”:

- “Subscribed”, “Terminated” e “Lapsed”.

1.4.3. Tabela “types”

Valores admitidos pelo campo “type_designation”:

- “Movie”, “TV Show”, “Mini Movie”, “Docuseries” e “One-Off”.

1.4.4. Tabela “age_restrictions”

Valores admitidos pelo campo “age_restriction_designation”:

- “6+”, “12+”, “16+” e “18+”.

1.4.5. Tabela “categories”

Valores admitidos pelo campo “category_designation”:

- “Adventure”, “Humor”, “Melodrama”, “Suspense”, “Terror”, “Love Story”, “Nonfiction”, “Cartoon”, “Science Fiction” e “Fiction”.

1.4.6. Tabela “sessions”

Valores admitidos pelo campo “app_version”:

- “version 1.0.0”, “version 1.2.3”, “version 2.0.1”, “version 2.1.0” e “version 3.0.5”.

2. Modelo Dimensional

Desenvolveu-se um modelo dimensional para o Data Mart da StreamFlix (ver Figura 4). Este segue um modelo em estrela, tendo a sua estrutura sido pensada de forma a atender à análise dos padrões de consumo de conteúdo. Nele, a tabela "sessions" (a tabela de factos) guarda as sessões de visualização, com dados como a duração da sessão em minutos e a percentagem do conteúdo visualizada. A mesma encontra-se ainda ligada a quatro tabelas de dimensão. De forma a otimizar o armazenamento e o desempenho das consultas, optou-se por substituir os antigos identificadores dos utilizadores, conteúdos e sessões (anteriormente definidos como VARCHAR (16)) por chaves substitutas do tipo INT. Esta alteração reduz significativamente o espaço ocupado pelos registos na tabela de factos.

A dimensão "users" (ver Tabela 13) armazena informações sobre os utilizadores, como o seu grupo etário, sexo, hierarquia de localização (país → distrito → cidade → código postal → endereço da rua), data de inscrição e estado da subscrição, permitindo assim a análise de conteúdos por perfil ou a relação entre o tempo de utilização da plataforma e a taxa de cancelamento de assinaturas.

Por sua vez, a dimensão "contents" (ver Tabela 14), contém detalhes sobre os conteúdos, como os seus géneros, data de lançamento, tipo, duração e diretor. Permite-se assim entender quais os géneros e tipos de conteúdo que mais captam a atenção, assim como os que são mais propensos ao abandono antes da sua conclusão.

Na dimensão "devices" (ver Tabela 15), encontram-se registados os dispositivos utilizados para assistir ao conteúdo. É armazenada a sua hierarquia (plataforma → tipo de dispositivo → família do sistema operativo → nome do sistema operativo → versão da aplicação), o que permite compreender como varia o consumo com a plataforma utilizada. Por exemplo, assistir a conteúdos mais longos na televisão e conteúdos mais curtos no telemóvel.

Na última dimensão, "times" (ver Tabela 16), permite-se a análise temporal do consumo, contendo a hierarquia temporal (ano → mês → semana → dia → hora → minuto). Esta permite estudar padrões ao longo do tempo, como horários de pico de visualização ou variações sazonais no consumo.

Voltando às dimensões "users" e "contents", estas incorporam os campos "initial_date", "final_date" e "active", resultantes do tratamento das *Slowly Changing Dimensions* (SCD). Nas várias tabelas, classificaram-se os campos como tipo 0 quando os valores dos atributos nunca são alterados. Tipo 1 quando a alteração só faz sentido em caso de erro (não se querendo então guardar histórico). Tipo 2.1 quando basta criar um novo registo na dimensão e tipo 2.3, em que para além de se adicionar um registo, adiciona-se uma *flag* e um intervalo de validade para identificar o mais atual.

Para garantir a consistência entre os diferentes sistemas operacionais, foram mapeados os valores admitidos em campos que apresentam discrepâncias. Quando os três primeiros sistemas apresentam um conjunto de valores admitidos diferentes do quarto sistema, opta-se por preservar os valores dos primeiros. Assim, valores não coincidentes como, por exemplo, "Adventure", são mapeados para valores como "Action". Para rastrear a sua origem, os registos passam a conter um valor de "source": "postgresql1" caso provenha de um dos três primeiros sistemas operacionais e "postgresql2", caso a origem seja o sistema PostgreSQL2.

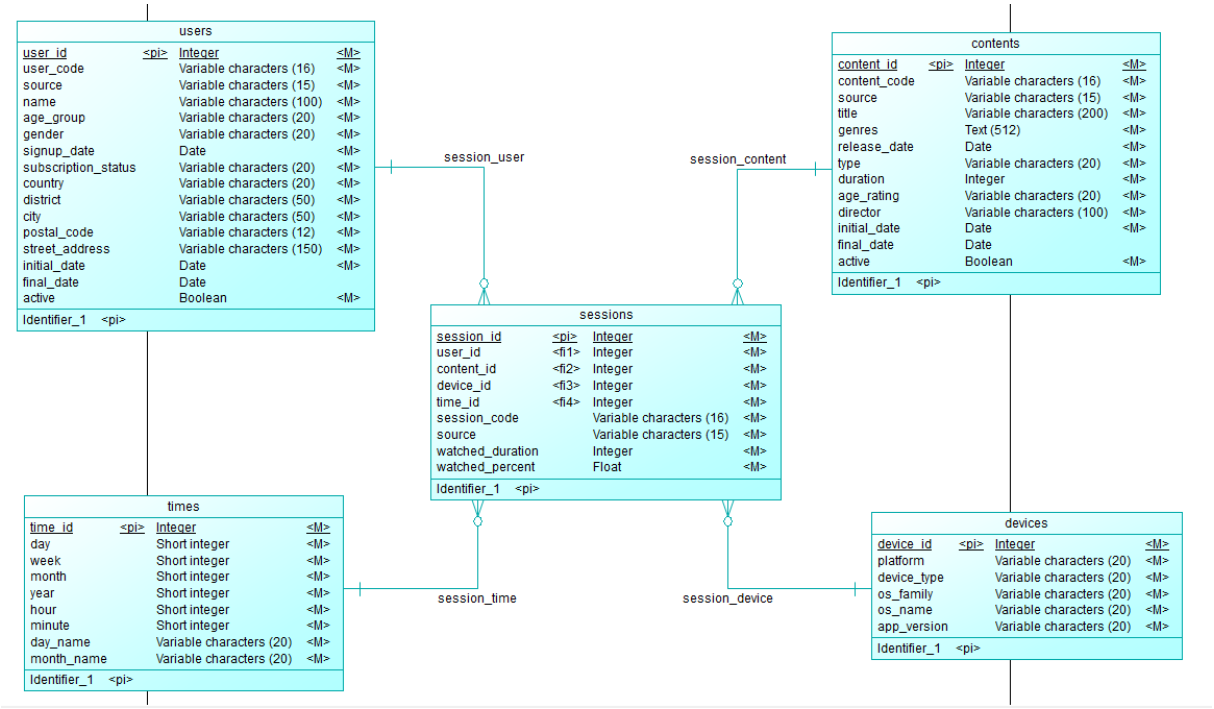

Figura 4 - Modelo Lógico do Modelo Dimensional

2.1.1. Dimensão “users”

Campo	Tipo	Descrição	Tipo SCD
user_id	INT	Novo identificador do utilizador (chave substituta)	0
user_code	VARCHAR (16)	Antigo identificador do utilizador	0
source	VARCHAR (15)	Qual a origem do registo	0
name	VARCHAR (100)	Nome do utilizador	2.3
age_group	VARCHAR (20)	Grupo etário associado	2.3
gender	VARCHAR (20)	Sexo associado	2.3
signup_date	DATE	Data de inscrição na plataforma	0
subscription_status	VARCHAR (20)	Estado da subscrição do utilizador	2.3
country	VARCHAR (20)	País associado	2.3
district	VARCHAR (50)	Distrito do utilizador	2.3
city	VARCHAR (50)	Cidade associada	2.3
postal_code	VARCHAR (12)	Código Postal associado	2.3
street_address	VARCHAR (150)	Endereço da rua do utilizador	2.3
initial_date	DATE	A data em que o registo foi criado	N/A
final_date	DATE	A data em que o registo ficou desatualizado	N/A
active	Boolean	Identifica se o registo é o mais recente	N/A

Tabela 13 - Estrutura da Dimensão "users"

2.1.2. Dimensão “contents”

Campo	Tipo	Descrição	Tipo SCD
content_id	INT	Novo identificador do conteúdo (chave substituta)	0

content_code	VARCHAR (16)	Antigo identificador do conteúdo		0
source	VARCHAR (15)	Qual a origem do registo		0
title	VARCHAR (200)	Título do conteúdo	2.3	
genres	Text (512)	Os géneros do conteúdo (separados por ';')	2.3	
release_date	DATE	Data de lançamento do conteúdo		1
type	VARCHAR (20)	O tipo do conteúdo	2.3	
duration	INT	A duração do conteúdo (em minutos)		1
age_rating	VARCHAR (20)	Classificação etária do conteúdo	2.3	
director	VARCHAR (100)	Diretor do conteúdo		1
initial_date	DATE	A data em que o registo foi criado	N/A	
final_date	DATE	A data em que o registo ficou desatualizado	N/A	
active	Boolean	Identifica se o registo é o mais recente	N/A	

Tabela 14 - Estrutura da Dimensão "contents"

2.1.3. Dimensão “devices”

Campo	Tipo	Descrição	Tipo SCD
device_id	INT	Novo identificador do dispositivo (chave substituta)	2.1
platform	VARCHAR (20)	Categoria geral do dispositivo utilizado para assistir à sessão	2.1
device_type	VARCHAR (20)	Categoria física do dispositivo utilizado para assistir à sessão	2.1
os_family	VARCHAR (20)	Família a que pertence o sistema operativo do dispositivo	2.1
os_name	VARCHAR (20)	O nome do sistema operativo do dispositivo	2.1
app_version	VARCHAR (20)	A versão da aplicação utilizada durante a sessão	2.1

Tabela 15 - Estrutura da Dimensão "devices"

2.1.4. Dimensão “times”

Campo	Tipo	Descrição	Tipo SCD
time_id	INT	Identificador do tempo	0
day	SHORT INT	Dia do mês (1 a 31)	0
week	SHORT INT	Número da semana no ano (1 a 52)	0
month	SHORT INT	Número do mês (1 a 12)	0
year	SHORT INT	Ano do registo (ex: 2023, 2024, etc.)	0
hour	SHORT INT	Hora do dia (0 a 23)	0
minute	SHORT INT	Minuto da hora (0 a 59)	0
day_name	VARCHAR (20)	Nome do dia da semana	0
month_name	VARCHAR (20)	Nome do mês	0

Tabela 16 - Estrutura da Dimensão "times"

2.1.5. Tabela de Factos “sessions”

Campo	Tipo	Descrição	Tipo SCD
session_id	INT	Novo identificador da sessão (chave substituta)	0
user_id	INT	Identificador do utilizador que assistiu à sessão	1

content_id	INT	Identificador do conteúdo que foi assistido	1
device_id	INT	Identificador do dispositivo em que foi assistido	1
time_id	INT	Identificador do tempo em que se realizou	1
session_code	VARCHAR (16)	Antigo identificador da sessão	0
source	VARCHAR (15)	Qual a origem do registo	0
watched_duration	INT	Qual a duração da sessão em minutos	1
watched_percent	FLOAT	Qual a percentagem assistida (em relação à duração do conteúdo)	1

Tabela 17 - Estrutura da Tabela de Factos "sessions"