Instituto Politécnico de Viseu

Escola Superior de Tecnologia e Gestão de Viseu

Departamento de Informática

Armazenamento e Processamento Analítico de Dados

Proposta de tema para o Projeto

Jesús Betancourt (pv22987)

Leandro Dias (pv23028)

Miguel Batista (pv22976)

Rodrigo Correia (pv23006)

Viseu, 2025

Instituto Politécnico de Viseu

Escola Superior de Tecnologia e Gestão de Viseu

Departamento de Informática

Armazenamento e Processamento Analítico de Dados

Curso de Mestrado em

Engenharia Informática - Sistemas de Informação

Proposta de tema para o Projeto

Ano Letivo 2024/2025

Jesús Betancourt (pv22987)

Leandro Dias (pv23028)

Miguel Batista (pv22976)

Rodrigo Correia (pv23006)

Viseu, 2025

Índice

1. Caso a Tratar

Tem-se a StreamFlix, que não é mais do que uma plataforma de *streaming* que oferece filmes e séries sob demanda aos seus utilizadores. O seu principal objetivo é fornecer conteúdo de qualidade, assim como a personalização das recomendações de acordo com os gostos dos clientes para melhorar a sua experiência e reduzir o *churn rate* (a taxa de cancelamentos).

A área de atuação da empresa é o entretenimento digital, seguindo um modelo SVOD *(Subscription Video on Demand)*, onde os clientes pagam uma assinatura mensal que lhes dá acesso a uma ampla biblioteca de conteúdos (outros exemplos deste modelo seriam a Netflix ou a Disney+).

Quanto ao processo de negócio que se vai modelar, este será a personalização das recomendações, tendo-se em conta os padrões de consumo dos utilizadores. Será assim possível entender quais os conteúdos mais populares, identificar tendências de visualização e minimizar os cancelamentos ao direcionar o conteúdo para o seu público-alvo.

1.1. Problemas Enfrentados

Enfrentam-se desafios na gestão e análise dos dados devido à fragmentação da informação em três sistemas operacionais diferentes. Os registos referentes às sessões de visualização por parte dos utilizadores são armazenados num ficheiro CSV. As informações dos utilizadores em si encontram-se numa base de dados PostgreSQL e ainda os detalhes sobre o conteúdo que é assistido está numa outra base de dados MySQL. Esta dispersão dos dados dificulta a análise do perfil do utilizador face ao conteúdo que assiste, exigindo assim a integração dos dados entre as várias fontes.

2. Sistemas Operacionais

Os dados dos sistemas operacionais fonte serão obtidos recorrendo à ferramenta TPC-H DBGEN, que permite a geração de dados sintéticos. O DBGEN será executado no ambiente Cygwin para que se produzam os conjuntos de dados precisos para os três sistemas operacionais distintos.

2.1. Sistema 1 – Ficheiro 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.

session_id	INT	O identificador da sessão
user_id	INT	O identificador do utilizador que assistiu à sessão
content_id	INT	O identificador do conteúdo que foi assistido
device_type	VARCHAR (20)	Qual o tipo de dispositivo utilizado (e.g. computador, telemóvel, tablet)
app_version	VARCHAR (20)	A versão da aplicação utilizada
os_name	VARCHAR (20)	O nome do sistema operativo onde assistiu (e.g. macOS, Linux, Windows)
time	TIMESTAMP	Quando é que a sessão se realizou
watched_duration	INT	Qual a duração da sessão em minutos

Tabela 1 - Estrutura do CSV

2.2. Sistema 2 – PostgreSQL

Na base de dados PostgreSQL (ver Figura 1) está armazenada a informação relativa a cada um dos utilizadores da plataforma. Na tabela *users*, guardam-se o seu nome, email e data de

inscrição. A mesma relaciona-se ainda com quatro outras tabelas, o que leva a que os utilizadores estejam associados a uma faixa etária (e.g. 10-14, 20-24), género, país e estado da subscrição (e.g. ativa, cancelada ou expirada).

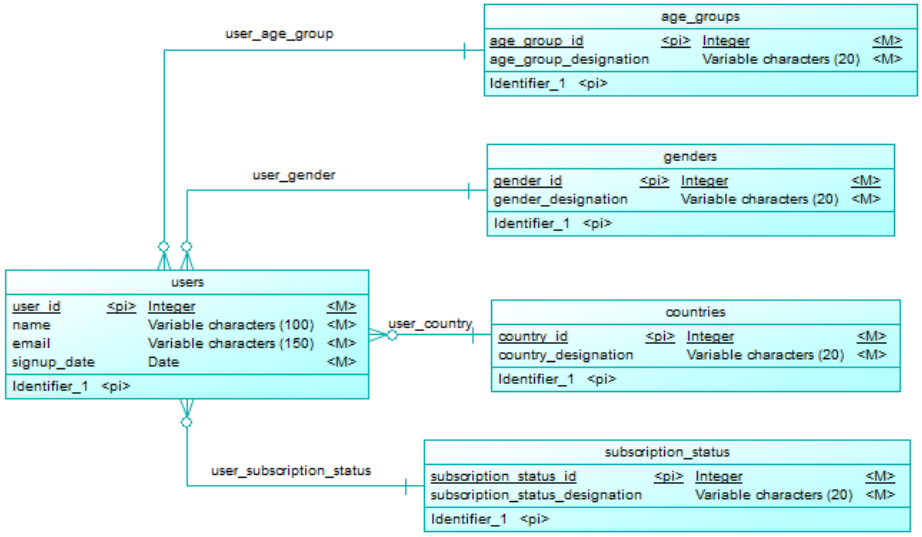

Figura 1 - Modelo E-R PostgreSQL

2.3. Sistema 3 – MySQL

Quanto à base de dados MySQL, é nela contida toda a informação sobre o conteúdo disponível na StreamFlix. A tabela *contents* armazena o título do conteúdo, a sua data de lançamento e a sua duração em minutos. Através de relacionamento com as outras tabelas, os conteúdos dispõem também de um tipo (e.g. filme ou série), classificação etária, diretor e ainda de um ou mais géneros.

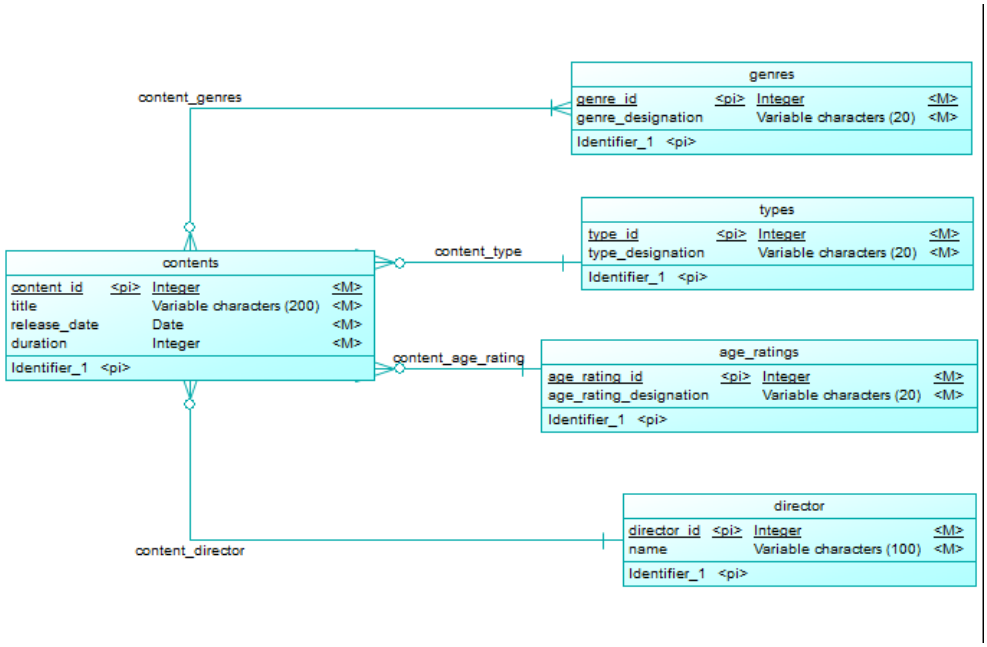

Figura 2 - Modelo E-R MySQL

3. Modelo Dimensional

Desenvolveu-se um esboço do modelo dimensional para o Data Mart da StreamFlix (ver Figura 3). 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.

A dimensão dos *users* armazena informações sobre os utilizadores, como a sua faixa etária, género, país, 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.

Já a dimensão dos *contents*, 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 dos *devices*, encontram-se registados os dispositivos utilizados para assistir ao conteúdo. São armazenadas informações como o seu tipo, versão da aplicação e nome do seu sistema operativo, 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.

Por fim, a dimensão de *times* permite a análise temporal do consumo, contendo informação como o dia, semana, mês, hora, etc. Esta dimensão permite estudar padrões ao longo do tempo, como horários de pico de visualização ou variações sazonais no consumo.

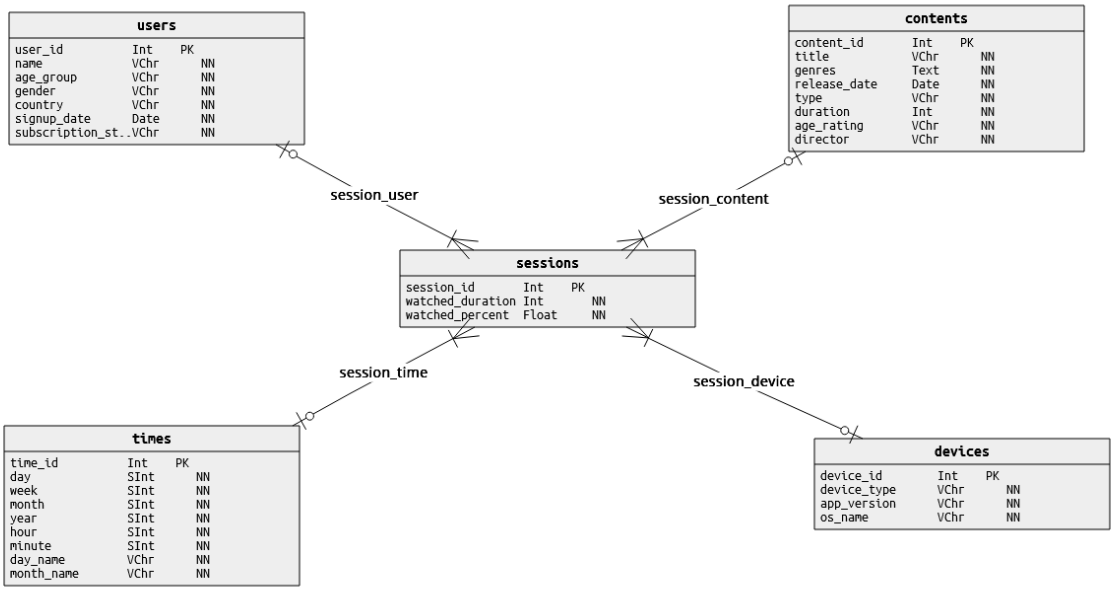

Figura 3 - Modelo Dimensional

4. Ferramenta OLAP

Escolheu-se a ferramenta Microsoft SQL Server Analysis Services (SSAS) para a solução OLAP. A mesma permite a criação de cubos de dados e modelos tabulares, o que possibilita então a realização de análises cruzadas, operações *drill-down, roll-up, slice and dice* e também a separação de dados em várias dimensões. Dispõe ainda de integração nativa com o Power BI e o Excel, tornando assim mais fácil a visualização dos dados para efeito analítico dos padrões de consumo e comportamento dos utilizadores da plataforma StreamFlix.