Slide 1 - Título
"Boa noite, somos o grupo responsável pelo projeto StreamFlix. O nosso objetivo foi simular uma plataforma de streaming realista, com todos os componentes que constituem uma solução completa de Business Intelligence, desde a geração dos dados operacionais até à análise OLAP com o Power BI."

Slide 2 - Contexto e Objetivo
"As plataformas de streaming, como a Netflix ou a Disney+, geram diariamente grandes volumes de dados relacionados com os seus utilizadores, conteúdos assistidos, dispositivos em que assistem e duração assistida.

Estes dados têm enorme valor analítico, mas são muitas vezes dispersos por sistemas distintos, com estruturas diferentes. Por isso, a criação de uma arquitetura analítica bem estruturada torna-se essencial para garantir a integração, consistência e capacidade de exploração destes dados.

Neste projeto, simulou-se esse cenário e desenvolveu-se uma arquitetura de BI composta por quatro sistemas operacionais, um processo ETL, um modelo dimensional com suporte a histórico e um cubo OLAP interligado com dashboards em Power BI."

Slide 3 - Arquitetura Geral
"Este slide apresenta a arquitetura geral da solução.

Inicialmente, estava previsto o uso de três sistemas operacionais distintos: um ficheiro CSV, uma base de dados MySQL e uma PostgreSQL. No entanto, ao analisar a estrutura e função destes sistemas, percebeu-se que se tratava, na prática, de uma base de dados fragmentada, sem sobreposição ou redundância de dados.

Para simular um cenário de integração real, criou-se um quarto sistema operacional denominado “PostgreSQL2”. Este sistema introduz dados e estruturas diferentes, obrigando à integração entre fontes.

Todos os modelos (conceptual, lógico e físico) dos sistemas e do modelo dimensional foram desenhados no PowerDesigner. A partir destes modelos, geraram-se as scripts de criação das tabelas.

Os dados, depois de gerados, são submetidos a um processo ETL desenvolvido em Python, que procede à sua extração, transformação e carregamento para o modelo dimensional, alojado no SQL Server. Sobre este modelo constrói-se um

cubo OLAP, cujos dados são posteriormente analisados e visualizados através de dashboards em Power BI."

Slide 4 a 7 - Geração de Dados
"Para garantir controlo sobre os dados e diversidade de cenários, optou-se por gerar os dados com Python, utilizando a biblioteca Faker.

Como dito anteriormente, existem quatro sistemas:

- **Ficheiro CSV:** guarda sessões de visualização; (Próximo slide)
- **MySQL:** armazena informação sobre conteúdos; (Próximo slide)
- **PostgreSQL1:** contém dados sobre utilizadores; (Próximo slide)
- **PostgreSQL2:** que acaba por ser a junção dos outros 3, com algumas diferenças.

Para certas tabelas, os valores admitidos em determinados campos diferem entre os três primeiros sistemas operacionais e o quarto. Um exemplo é a tabela 'genres', que nos três primeiros sistemas pode conter o valor 'Action', enquanto no quarto sistema, a tabela equivalente 'categories' utiliza 'Adventure' como valor correspondente.

Estas diferenças exigem um tratamento específico durante a fase de transformação, obrigando à normalização e mapeamento dos dados entre sistemas."

Slide 8 - Processo ETL e Tratamento de Alterações
“Na fase de extração, recolhem-se os dados de cada sistema. Na primeira extração, adiciona-se a cada tabela o campo IS_UP_TO_DATE, com valor default 0. Criam-se ainda triggers que alteram esse campo para 0 sempre que ocorre uma atualização nos dados. Após a extração, o campo é atualizado para 1, sinalizando que os dados já foram processados. Assim, em extrações futuras, apenas os registos com IS_UP_TO_DATE = 0 serão considerados na transformação.

Durante a transformação, aplicou-se o tratamento das Slowly Changing Dimensions (SCD). Os atributos foram classificados como tipo 0 caso imutáveis (como identificadores, por exemplo), tipo 1 para correções sem histórico (quando só mudam em casos de erro), tipo 2.1 para duplicação simples (o caso da dimensão devices) e tipo 2.3, em que cada registo tem uma versão ativa e intervalo de validade. Nas dimensões USERS e CONTENTS, foram utilizados os

campos initial_date, final_date e active para gerir o histórico, visto terem atributos classificados com o tipo 2.3.

Adicionalmente, optou-se por substituir os identificadores alfanuméricos originais por chaves substitutas inteiras, reduzindo o espaço ocupado na tabela de factos e melhorando o desempenho das consultas.

Para garantir consistência entre os sistemas operacionais, procedeu-se também ao mapeamento dos valores admitidos em campos com diferenças. Optou-se por preservar os valores dos primeiros sistemas. Por exemplo, o valor 'Adventure' do PostgreSQL2 é mapeado para 'Action'. Para se saber a origem dos registos, os mesmos são etiquetados com o campo source, que assume 'postgresql1' para os três primeiros sistemas e 'postgresql2' para o quarto."

Slide 9 - Modelo Dimensional
"Desenvolveu-se depois um modelo dimensional para o Data Mart da StreamFlix, baseado numa estrutura em estrela, pensada para permitir a análise de padrões de consumo de conteúdo.

A tabela de factos 'SESSIONS', armazena informações como a duração das sessões e a percentagem visualizada. Esta está ligada a quatro dimensões:

- **USERS:** inclui grupo etário, género, localização de forma hierárquica, data de inscrição e estado da subscrição. Permitindo esta analisar o perfil dos utilizadores e a relação entre uso da plataforma e cancelamentos;
- **CONTENTS:** contém dados como géneros, tipo, duração, data de lançamento e realizador. Que permite estudar o comportamento face a diferentes tipos de conteúdo;
- **DEVICES:** regista os dispositivos utilizados para assistir ao conteúdo, sendo armazenada a sua hierarquia completa, desde a plataforma até à versão da aplicação, permitindo estudar padrões de consumo por dispositivo;
- **TIMES:** permite análise temporal com hierarquia até ao minuto.

Slide 10 a 12 - Testes e Validação
"Para validar o comportamento do processo ETL, executaram-se testes de inserção e atualização nos sistemas operacionais, simulando cenários reais.

Para um dos testes de inserção, um novo registo foi adicionado ao ficheiro CSV. Após a execução do processo ETL, verifica-se que o registo é refletido corretamente no modelo dimensional. (Próximo slide)

Quanto ao teste de atualização realizado no sistema MySQL, modifica-se um registo existente. (Próximo slide) O processo ETL identifica a alteração e cria uma nova versão do conteúdo no modelo dimensional, preservando a versão anterior, em conformidade com o tratamento de SCD do tipo 2.3.

Comprova-se então que tanto as inserções como as atualizações são tratadas corretamente e refletidas no Data Mart."

Slide 13 - Cubo OLAP
"Com os dados carregados no modelo dimensional, procedeu-se à criação de um cubo OLAP com o Microsoft SQL Server Analysis Services (SSAS), recorrendo ao Visual Studio 2022.

O cubo permite análise multidimensional e drill-through, assim como slicing conforme os filtros aplicados. A estrutura suporta cruzamentos de dados por utilizador, conteúdo, data, dispositivo e tempo.**"**

Slide 14 a 15 - Dashboards Power BI
"A criação do cubo possibilitou o desenvolvimento de duas dashboards no Power BI.

A primeira é uma dashboard geral, que permite acompanhar métricas globais da plataforma, como número total de sessões, conteúdos mais vistos, dispositivos preferidos e evolução diária da atividade. (Próximo slide)

A segunda é individual, permitindo analisar o comportamento detalhado de um utilizador específico. Através de drill-throughs e segmentações, é possível entender os conteúdos mais assistidos por esse utilizador, o seu tempo total de visualização, dispositivos mais utilizados e preferências de género.

Estas dashboards simulam casos de uso que suportam decisões analíticas numa plataforma de streaming."

Slide 16 - Trabalho Futuro
"Como trabalho futuro, existem várias direções possíveis para evolução do projeto:

- Automatizar todo o pipeline de ETL com ferramentas de instrumentação como o Apache Airflow, que permitirá execuções periódicas, gestão de dependências e alertas;
- Aumentar o volume de dados, de modo a testar escalabilidade e desempenho da arquitetura;
- Incorporar fontes de dados externas, como redes sociais ou perfis de utilizador.

Slide 17 - Fim da Apresentação
"E com isto, dá-se como terminada a apresentação. Agradecemos toda a atenção e ficamos disponíveis para esclarecer qualquer questão."
