Relatório do Trabalho Final de Banco de Dados

Gustavo de Mendonça Freire
123102270

João Victor Lopez Pereira
123317370

William Victor Quintela Paixão
123089993

Yuri Rocha de Albuquerque
123166143

19 de outubro de 2024

Rio de Janeiro - RJ

Sumário

Capítulo 1

Introdução

O tema do nosso trabalho é a análise de tendências no *YouTube*, com foco em compreender quais tipos de vídeos estão em alta na plataforma, identificar os canais mais influentes e observar o engajamento que esses vídeos recebem, medido em termos de curtidas, comentários e visualizações. Através dessa análise, buscamos extrair *insights* que permitam caracterizar o que define um vídeo popular, os padrões de conteúdo que se destacam, e a relação entre a popularidade e o país de origem dos vídeos. Para isso, estamos nos guiando através das seguintes principais questões: "Quais categorias de vídeos estão em alta?", "Quais tipos de canais produzem o conteúdo mais relevante e visualizado?", "Como as tendências variam de acordo com as regiões?" e "Qual a relação entre quantidade de comentários e *likes* e a classificação de um vídeo?". Essas perguntas, além de norteadoras, nos motivaram a escolher esse tema. Por meio desta análise, buscamos não apenas identificar padrões regionais e globais, mas também entender como os algoritmos da plataforma podem influenciar o comportamento do público e o sucesso de certos tipos de conteúdo.

Os conjuntos de dados selecionados são dos sites *GitHub* e *Social Blade* - embora não tenhamos utilizado diretamente este último para a modelagem, apenas servindo como uma alternativa para pegar mais informações no futuro e comparar alguns dados - que podem ser encontrados através dos links:

- *Kaggle* : Site principal utilizado para a coleta de dados. Nele, o autor disponibilizou a *API* (através de um link que dá acesso a um repositório no *GitHub*), usada para a coleta dos dados disponíveis no site, a qual utilizamos diretamente para a coleta de dados mais recentes para a nossa aplicação. Esse site apresenta um *CSV* que contém informações gerais dos vídeos em alta separados por país, condensando todo o tipo de dado que possa ser relevante para a análise. Nossa aplicação utiliza essas informações - sobretudo o ranking do vídeo e a quantidade de visualizações - para mapear que tipo de conteúdo domina o topo das tendências nos diferentes países, além de fazer comparações entre as regiões.
- *Social Blade*: site que utilizamos para obter informações extras e comparar alguns dados obtidos pela *API*. Há uma quantidade muito grande de informações disponibilizadas pelo site. De maneira geral, ele fornece, em tempo real, todas as informações referentes a qualquer canal registrado no *YouTube* (como quantidade de inscritos, ganho estimado, classificação, etc). Utilizamos este site como um suporte para a análise dos canais que publicaram os vídeos presente no "em alta"

Capítulo 2

Datasets

Como informado na seção anterior, foram considerados, principalmente, dois conjuntos de dados: um que contém um registro, gerado por *script* pelo usuário, dos vídeos em alta no *YouTube* em alguns países; e o da plataforma *Social Blade*, que mantém uma imensa variedade de métricas de diferentes redes sociais para análise estratégica dos criadores de conteúdo.

1. Trending YouTube Video Statistics: Este é o principal conjunto de dados utilizado para a especificação do modelo conceitual, e, consequentemente, de todo o projeto do Banco de Dados. Dele buscamos adquirir a maior parte dos dados (ou todos eles) para o povoamento de nosso sistema. A postagem acerca do *dataset*, explicando seu intuito, conteúdo e possibilidades de uso, foi feita no *Kaggle*. O *dataset* em si consiste de arquivos *CSV* gerados pelo *script* disponibilizado pelo usuário Mitchell Jolly no repositório *Trending-YouTube-Scraper* de seu *GitHub*. O *script* foi feito em *Python* e faz uso da *YouTube Data API v3* para a consulta online dos dados. Cada arquivo CSV tem informações de até 200 vídeos mais em alta do *YouTube* em um dado país. Os países inclusos para análise na versão do autor são: EUA, Grã-Bretanha, Índia, Alemanha, Canadá, França, Coreia do Sul, Rússia, Japão, Brasil e México (embora o Brasil não tenha aparecido no conjunto de dados exibido como exemplo no *Kaggle*). O usuário do *script*, entretanto, tem liberdade para determinar seus países de interesse, bastando, para isso, a edição de um arquivo de texto contido na pasta em que o *script* se encontra. Além disso, o usuário deve disponibilizar uma chave da API, que deve ser obtida e ativada no *Google Cloud Console*. De posse de uma chave válida, só resta colá-la em um outro arquivo *.txt* na pasta e, finalmente, executar o programa pelo console. O código foi escrito para gerar tabelas com até 200 registros de vídeos em alta, para cada país, com os seguintes campos:

    - video_id: contém, em cada registro, o *ID* do vídeo gerado pelo próprio *YouTube*. Esse *ID* é utilizado, além de mera caracterização de vídeos únicos na plataforma, para a geração de suas *URL*s, por exemplo. Cada *ID* é uma sequência alfanumérica de símbolos, sendo mais apropriado modelá-lo como cadeia de caracteres (*string*) em nosso sistema;
    - title: refere-se ao título do vídeo em sua língua original (em que foi publicado). Suporta caracteres de *UTF-8*. É, naturalmente, do tipo textual, sendo, portanto, claramente representável por valores do domínio de *strings*;
    - publishedAt: corresponde à data e hora de publicação do vídeo (com relação ao Tempo Universal Coordenado *UTC*). O valores desse atributo são resgatados no formato *ISO 8601* e podem ser mapeados para o domínio *DATETIME* do *SGBD*;

- channelId: análogo ao *ID* gerado para o vídeo, este campo armazena o *ID* criado pelo site para cada canal, neste caso, o canal que publicou o vídeo. É uma cadeia alfanumérica e, consequentemente, está no domínio de *strings*;
- channelTitle: também análogo ao título do vídeo, é relativo ao título do canal que postou o determinado vídeo. É do tipo textual e mapeado no tipo *string*;
- categoryId: corresponde ao identificador da categoria atribuída ao vídeo pelo canal que o publicou. No guia de referências da *API* consta que é um dado do tipo *string*, porém, após coleta da listagem de categorias disponíveis, percebemos que todas se tratam de inteiros de 1 a 44 (pulando alguns). Portanto, consideramos que a atribuição ao domínio dos inteiros seja a melhor alternativa;
- trending_date: contém a data em que o vídeo está em alta, ou seja, equivale à data em que o arquivo *CSV* foi criado. Está em um formato diferente do utilizado pela *API* do *YouTube*, pois é uma informação adicionada, na força bruta, no *script*. Consiste do ano, dia e mês de coleta dos dados, nesta ordem, separados por ponto final (AA.DD.MM). É naturalmente associável ao domínio *DATE*;
- tags: corresponde a uma lista de *tags* (etiquetas) atribuídas ao vídeo, usadas como auxílio para o algoritmo de recomendações da plataforma. São pequenas porções de texto (ocorrências de *tag*) separadas por barras verticais (|). É permitido que a *tag* contenha espaços. Cada *tag* encaixa-se no domínio *string*;
- view_count: é o número total de visualizações que o vídeo teve desde sua publicação até o instante da captura dos dados. Consiste de um inteiro não-negativo (logicamente) e, por isso, pode ser modelado por um inteiro longo sem sinal;
- likes: é uma métrica numérica, assim como a contagem de visualizações, referente ao número de curtidas (*likes*) que o vídeo obteve até a coleta dos dados. É inteiro longo maior ou igual a zero;
- dislikes: assemelha-se ao campo denominado "likes". A partir de dezembro de 2021, passou a necessitar de autorização do canal detentor do vídeo para poder ser coletado pela *API*. Dessa maneira, a consideração de inclusão desse dado em nosso sistema, atualmente, é inviável;
- comment_count: novamente, é um campo numérico que retrata a quantidade de comentários que o vídeo recebeu até a captura dos dados. Pertence ao domínio de inteiros longos sem sinal;
- thumbnail_link: é a *URL* da *thumbnail* (imagem de capa) do vídeo. Por ser um *link*, que é composto por vários caracteres, alfanuméricos e símbolos especiais, em sequência, escolhemos o tipo *string* para armazená-los;
- comments_disabled: indica se o canal que publicou o vídeo optou por desativar a postagem de comentários. Nos arquivos *CSV*, esse campo é assinalado por "True" ou "False", devendo, evidentemente, ser tratado como um valor de domínio *booleano*;
- ratings_disabled: informa se o canal desativou a opção dos espectadores curtirem (ou descurtirem) o vídeo. Assim como "comments_disabled", somente assume valores "True" e "False" e, por isso, é traduzido para valores do espectro *booleano*;
- description: é a descrição do vídeo fornecida pelo seu publicador na língua original. É um texto de tamanho completamente variável (no máximo 5000 caracteres) e pode conter símbolos da codificação *UTF-8*. Logo, é do tipo *string*.

2. Social Blade: Esta é uma plataforma online (*site*) que incluímos como apoio para as nossas observações. Podemos caracterizá-la como um sistema de consultoria para produtores de conteúdo de uma multiplicidade de redes sociais (*YouTube, Twitch, Facebook, TikTok,*

Twitter (novo *X*), *Instagram*, só para citar algumas) que pretendem entender, dentre outras coisas, como suas postagens e atividades nas redes estão engajando (através da análise, em variados intervalos de tempo, do número de visualizações de seus perfis ou da variação da quantidade de seguidores, por exemplo) e qual a sua relevância em comparação a outros criadores segundo vários critérios. O sistema ainda é capaz de realizar projeções de quantidades de acessos e seguidores de 2 meses a 5 anos no futuro. Certamente, o *Social Blade* possui um banco de dados extremamente complexo e é capaz de performar inúmeras operações complicadas. Nosso papel com esse conjunto de dados não é extraí-lo e agregá-lo ao nosso esquema (até porque isso significaria um projeto muito além dos nossos atuais propósitos), mas usá-lo como referência para potenciais análises feitas no escopo de nosso modelo, servindo como uma forma confirmação.

Capítulo 3

Projeto do Banco de Dados

3.1 Modelo Conceitual

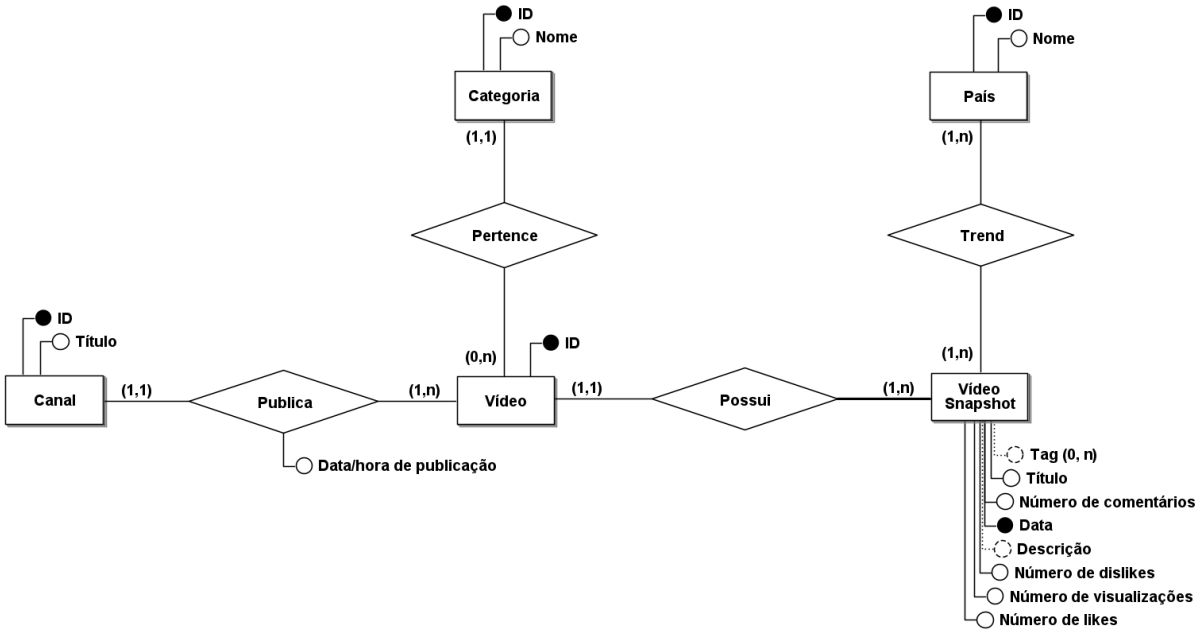

Figura 3.1: Modelagem conceitual do *Dataset* utilizado.

Como a base de dados tem os vídeos em alta como foco, as entidades do modelo correspondem aos conceitos relevantes para os vídeos do *YouTube* e seus dados. A entidade principal é "Vídeo', sendo cada um identificado por um *ID* e publicado por um único "Canal", nome dado às contas através das quais os produtores de conteúdo publicam suas criações. Cada canal possui um título, pelo qual é conhecido por seu público na plataforma, e um *ID*, que o identifica para os

sistemas do *YouTube*. Além disso, cada vídeo é, no momento da publicação, associado a uma categoria por seu criador, que assume um dentre 15 valores pré-determinados pela plataforma, como "Gaming" ou "Comedy", por exemplo.

As informações de categoria e canal de um determinado vídeo são fixas, e não mudam com o tempo. No entanto, a maioria dos dados sobre um vídeo, como quantidade de *likes* e visualizações, são constantemente atualizadas, sendo cruciais para o registro da sua popularização. Para representar os dados de um vídeo como estavam no momento em que entrou em alta em um certo país, é incluída a entidade "Vídeo Snapshot", que representa um "fotografia" do estado do vídeo em uma data. A *snapshot* possui os dados do vídeo que se referem a um momento do tempo, como número de *likes*, *dislikes* e comentários. E, como um canal tem o direito de editar título e descrição de seus vídeos após a publicação, esses dados também são considerados dependentes do tempo, e, portanto, são modelados como atributos da *snapshot*, e não da entidade vídeo. É importante ressaltar que, como um mesmo vídeo pode ficar em alta durante vários dias, consecutivos ou não, cada um pode se associar a várias *snapshots*, cada uma referente a um data diferente. Finalmente, para representar que um vídeo esteve em alta em um dado país, é incluída a relação "Trend" entre "País" e "Snapshot".

TODO: Apresentar e descrever as consultas realizadas. Cada consulta deve ser definida através de uma explicação textual acompanhada do respectivo comando SQL.

3.2 Modelo Lógico

3.3 Modelo Físico

Capítulo 4

Aplicação

TODO: Descrever a aplicação web desenvolvida, explicando como foi implementada e detalhando, com exemplos, as funcionalidades que ela disponibiliza.

Capítulo 5

Distribuição do trabalho

TODO: Indicar explicitamente as atividades realizadas por cada membro do grupo.

Capítulo 6

Considerações finais

TODO: Avaliar os resultados alcançados de acordo com os objetivos propostos.

Bibliografia

[1] Alphabet Inc. *Youtube*. Acessado em 11 de Outubro de 2024. URL: https://www.youtube.com.

[2] Google LLC. *Kaggle*. Acessado em 12 de Outubro de 2024. URL: https://www.kaggle.com/datasets/datasnaek/youtube-new.

[3] mitchelljy. *Trending YouTube Scraper*. Acessado em 28 de Setembro de 2024. URL: https://github.com/mitchelljy/Trending-YouTube-Scraper/tree/master.

[4] Jason Urgo. *Social Blade*. Acessado em 11 de Outubro de 2024. URL: https://socialblade.com/.