Projeto: Extração de dados da API do youtube

**Introdução:**

Este projeto tem como a proposta de elaborar uma arquitetura de extração da API do youtube até a elaboração do dashboard.

Conceitos Rápidos:

DAG: É uma representação do fluxo de trabalho, sem ciclos

Operators: São classes no airflow que representam uma tarefa específica.

Hooks: São usados para fazer iterações entre os sistemas, seja API, banco de dados

Endpoints utilizados:

1 – Pesquisar assunto

youtube/v3/search?: url para pegar os vídeos por assunto -> salva todo o histórico de busca, independente do idioma do canal

2 – Url para filtrar canais brasileiro:

Url : /youtube/v3/channels

Após salvar os registros, é usada a url /youtube/v3/channels para consultar o idioma do canal e gerar uma lista de vídeos dos canais

3 – Url para obter dados dos vídeos:

Após obter a lista de vídeos brasileiros: é usada a URL: /youtube/v3/videos para obter os dados completos, onde nesses dados serão usados os campos viewCount, likeCount, favoriteCount, commentCount, sempre gravando os dados históricos.

4 – Url para extrair os comentários:

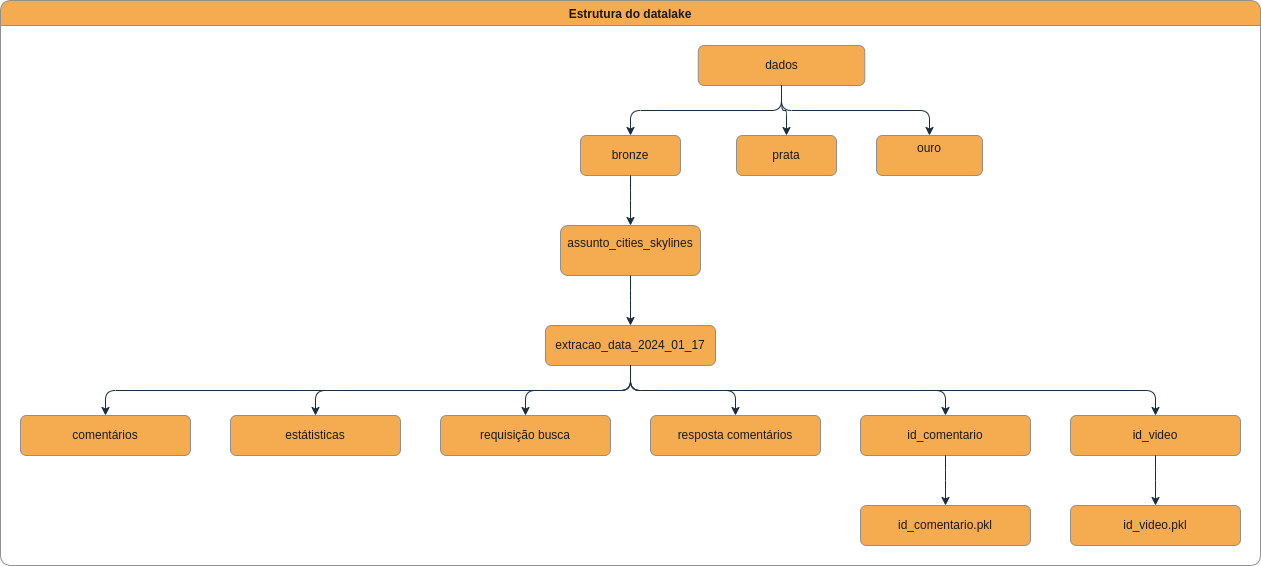
Depois de extrair a lista de vídeos, será executada a url de comentários (youtube/v3/commentThreads), gravando a requisição original em um datalakes

5 – Url para extrair as respostas dos comentários:

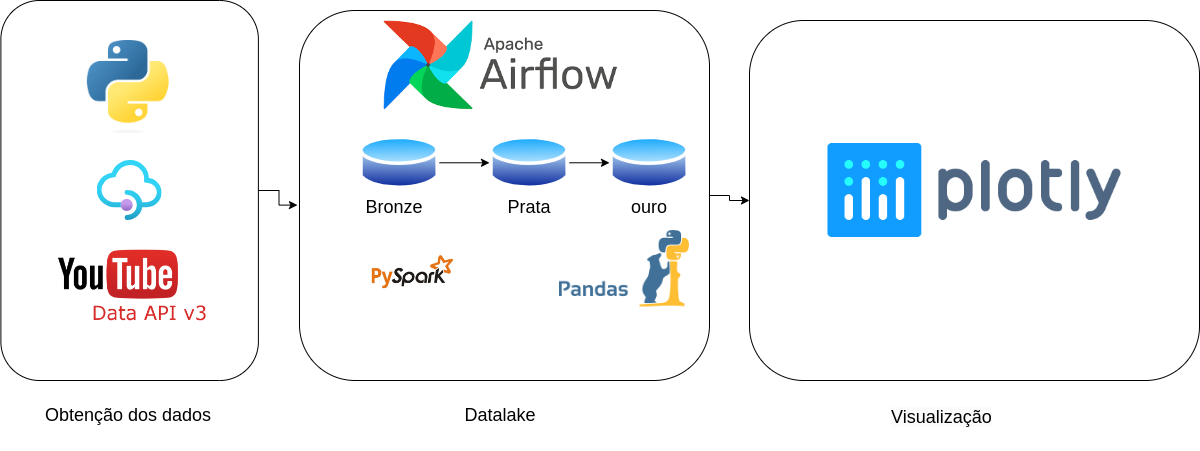
Depois de extrair os comentários, cada comentário tem um id que mostra se possui respostas para aquele comentário: a URL para a extração das resposta dos comentários é essa: (youtube/v3/comments).

Estrutura do datalake:

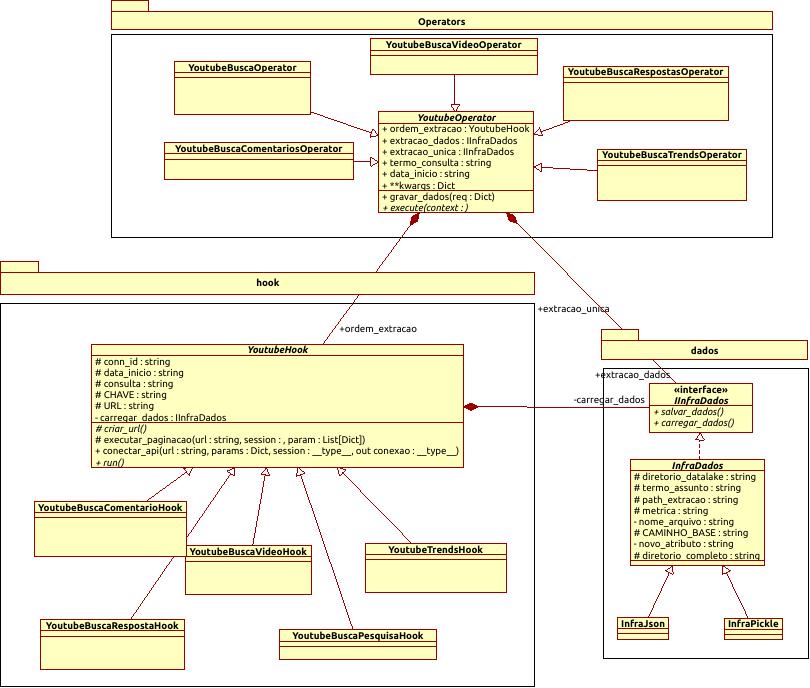
A figura abaixo mostra a estrutura do datalake que foi construido.



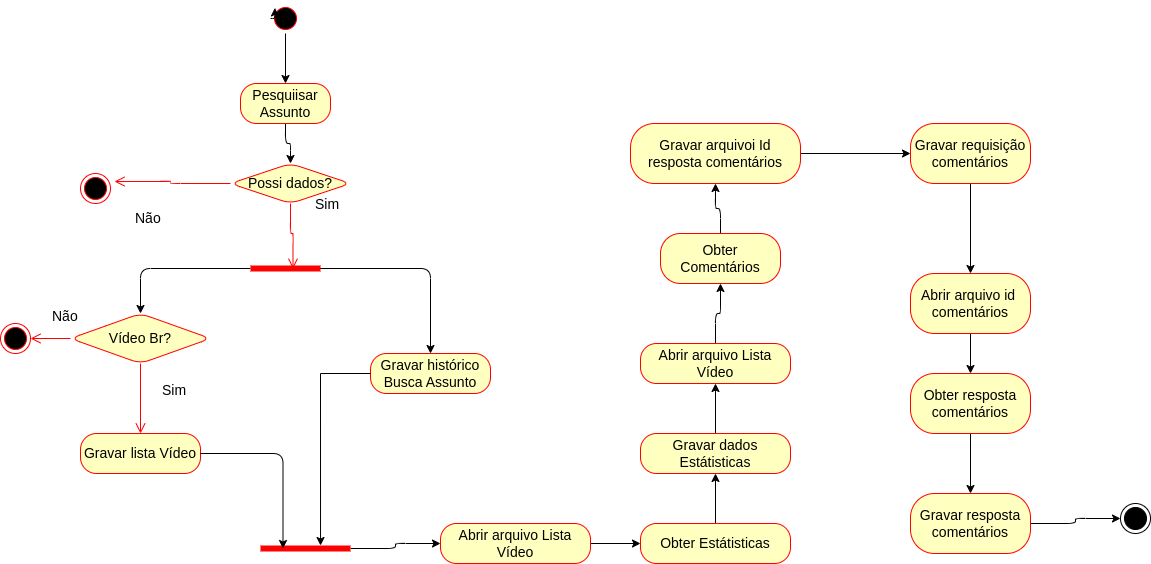
A Figura abaixo mostra o fluxo de dados até a geração do dashboard



A figura abaixo mostra o diagrama de classe da composição da dag. A lógica ficaria assim: para cada endpoints da api do youtube, iria herdar da classe base com seus requisitos específicos.



A figura abaixo mostra uma proposta de diagrama de atividade para a DAG



Observações:

Mesmo que o diagrama de atividade apresente, no final do processo, uma execução completa, uma proposta seria de implementar uma verificação de vídeos que contenham comentários antes de fazer todo o fluxo de comentários.