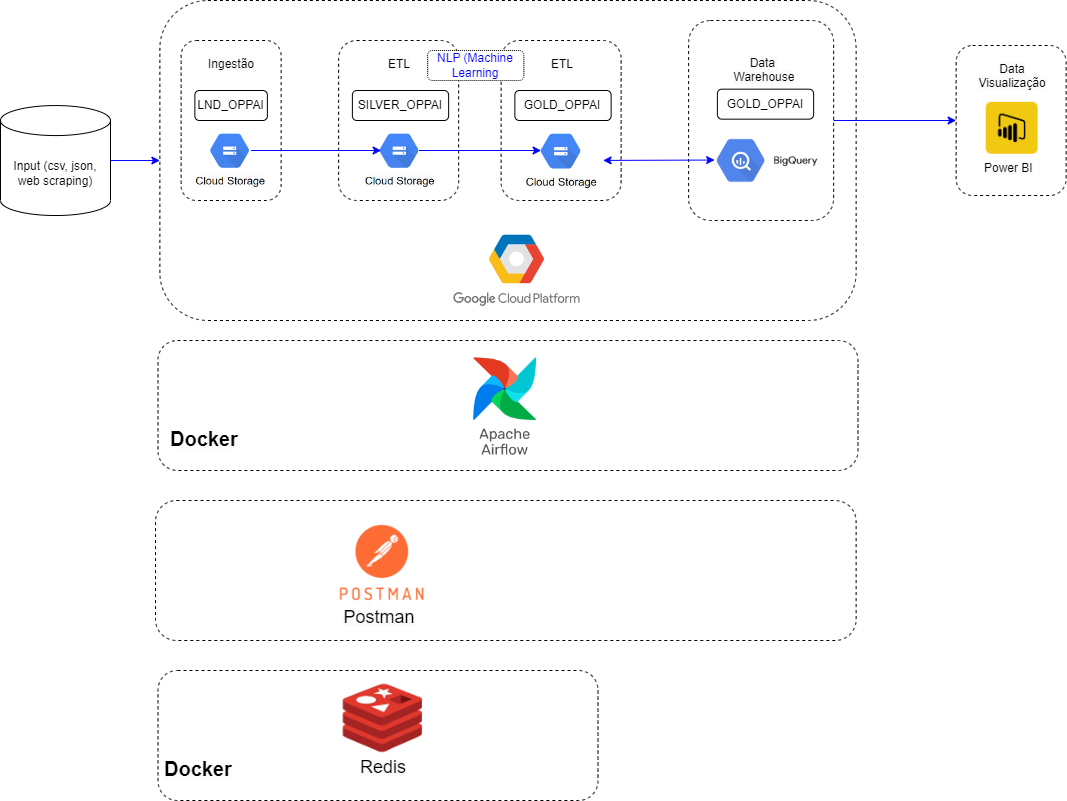
**Link do video sobre o projeto (**[**https://youtu.be/NtuZG4KOhPk**](https://youtu.be/NtuZG4KOhPk)**)**

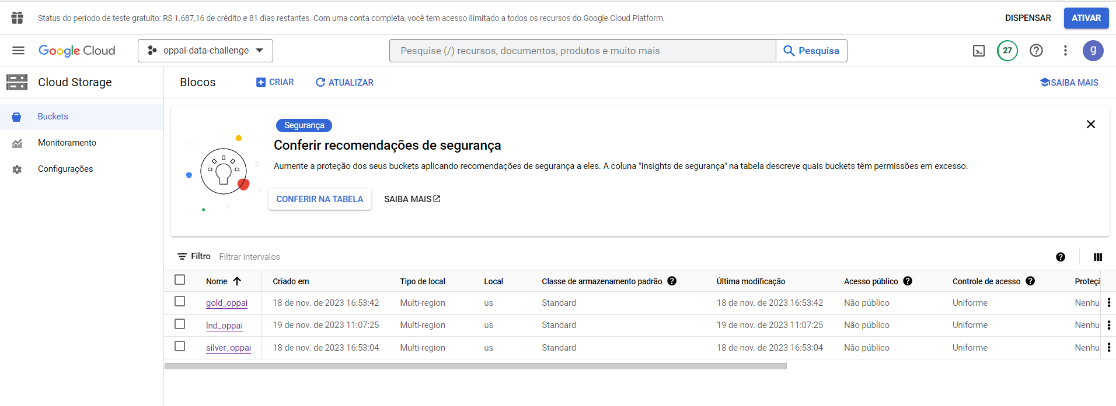
Explicação e entendimento das implementações feitas na arquitetura do DataLake:

* Criação da arquitetura dentro do ambiente Google Cloud.
* Utilizou-se Cloud Storage como os buckets de ingestão e de ETL.
* Utilização do Big Query como DW (Data Warehouse) - No ambiente Big Query usa-se SQL, assim caso seja necessário poderá criar as Views ou fazer joins entre tabelas Fatos e Dimensionais.
* Utilização do Python como linguagem principal, no entanto uma boa alternativa seria o PySpark, Dask ou Polars.
* Arquivos salvos em parquet e csv.
* Apliacação do NLP (Processamento de Linguagem Natural) - Rodou-se um algoritimo da qual retorno tipos de sentimentos baseado nos tipos de comentários (Primeiramente foi feito uma transformação para a língua inglesa, devido a biblioteca utilizada).
* Utilização do Airflow para orquestrar os Dags dos pipelines.
* Utilização do Postman para caso seja necessário ativar os pipelines por meio de API do AirFlow.
* Utilização do Redis - A utilização do Redis é uma alternativa viável para a incrementação dos dados futuramente, pois visando o a gravação das 'Datas' na memórias, tornaria o processo mais rápido.
* Utilização do Power BI para a visualização (Foi feito a conexão do Power BI no Big Query - as conexões pode ser to tipo direct ou import).

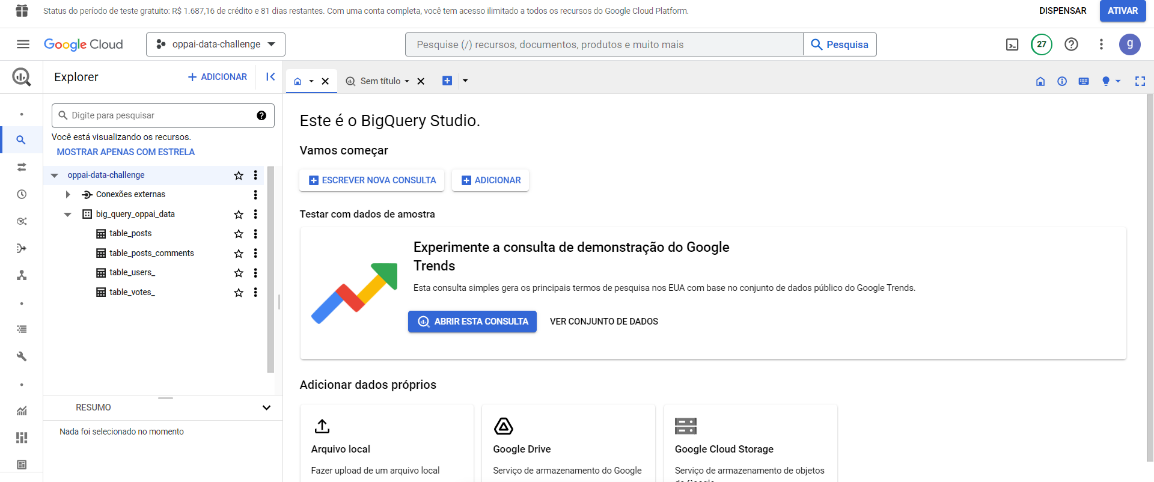
Arquitetura DataLake



Camadas dentro do DataLake



Camadas no DW

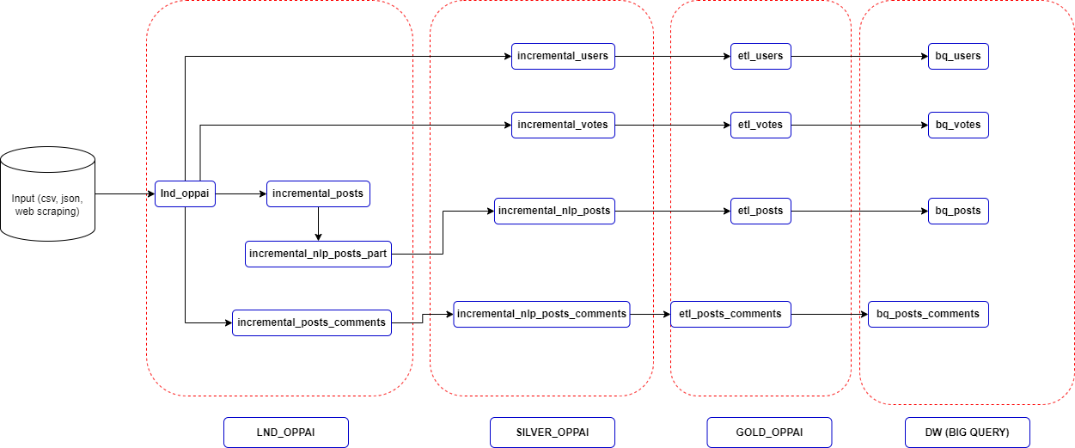


Machine Learning (NLP)

- Foi utilizado o modelo textBlob juntamento com RandomForest.

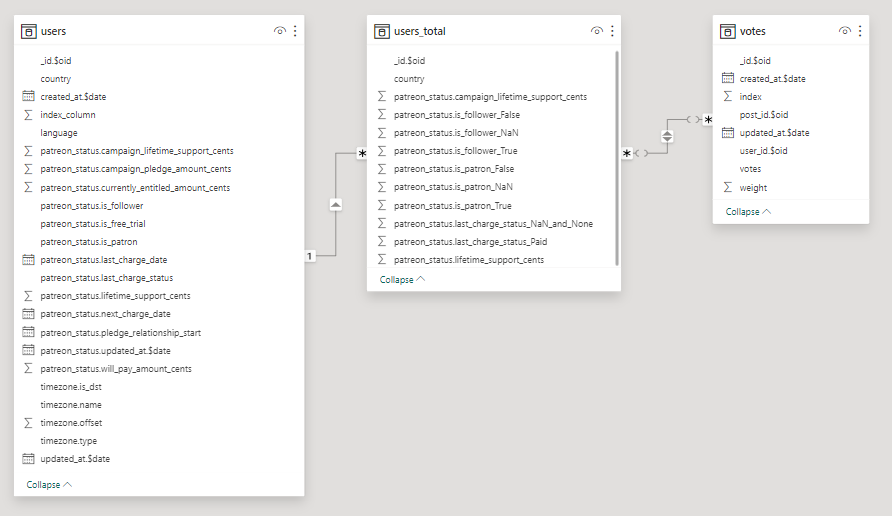
- Primeiramente todos os comentários foram traduzidos para a língua inglesa, devido ao fato da biblioteca funcionar unicamente somente na língue inglesa e também é um método padrão a ser aplicado, pois fazer estudo de NLP (Processamento de linguagem natural) com diversos idiomas tornaria o processo inviavel e demorado.

Workflow os pipelines estruturados pelo Airflow

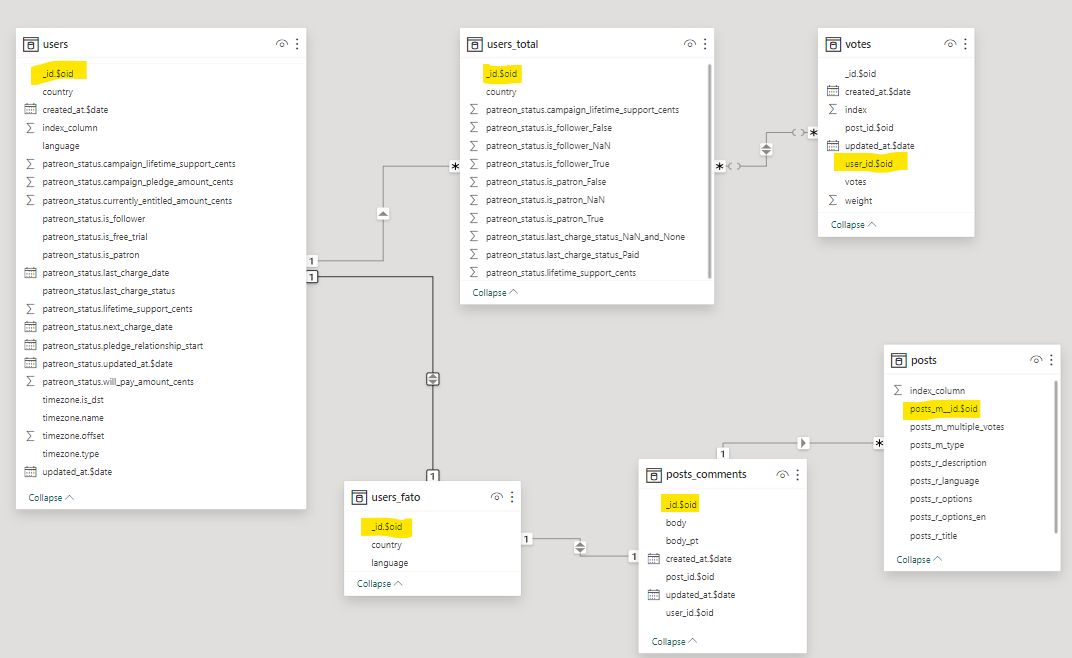


* A implementação das incrementais ocorre de acordo com a difereça de ‘datatime’, ou seja, durante a ingestão sempre será acrescentado apenas os dados com datas novos.
* Outro método a ser utilizado seria a gravação de data mais recente dentro do Redis, da qual serviria para o mesmo propósito.
* Foi criado duas pastas no Dag do Airflow Anexos\_simulando\_novos\_Files e Anexos para simular esse processo de incrementação.

Modelagem



* Tanto dataset ‘**users**’ quanto ‘**votes**’ apresenta lista de dicionarios nos valores das suas colunas, logo foi feito um ‘**explode**’ para obter os valores contidos neles.
* Join entre o dataset ‘**users\_total**’ com dataset ‘**votes**’, por meio das colunas **\_id.$oid** e **user\_id.$oid**.
* Foi feito essa relação para descobrir-se os paises provenientes de cada ‘**users**’ em **votes**, assim teremo um dataset enriquecido para posteriormente fazer a visualização das quantidades de ‘**users**’ com relação os ‘**country**’ paises.
* Realização de um ‘**groupby**’ e um ETL para a criação do dataset ‘**users\_total**’.



* Para enriquecer a tabela **post\_comments** foi feito um join entre os **\_id.$oid**.
* Para enriquecer a tabela posts realizou-se o cruzamento da tabela **post\_comment** já enriquecido, pois não houve possíveis keys para fazer um cruzamento direto entre **users** e **posts**.

Após a modelagem, responda as seguintes perguntas:

Porque você modelou dessa forma?

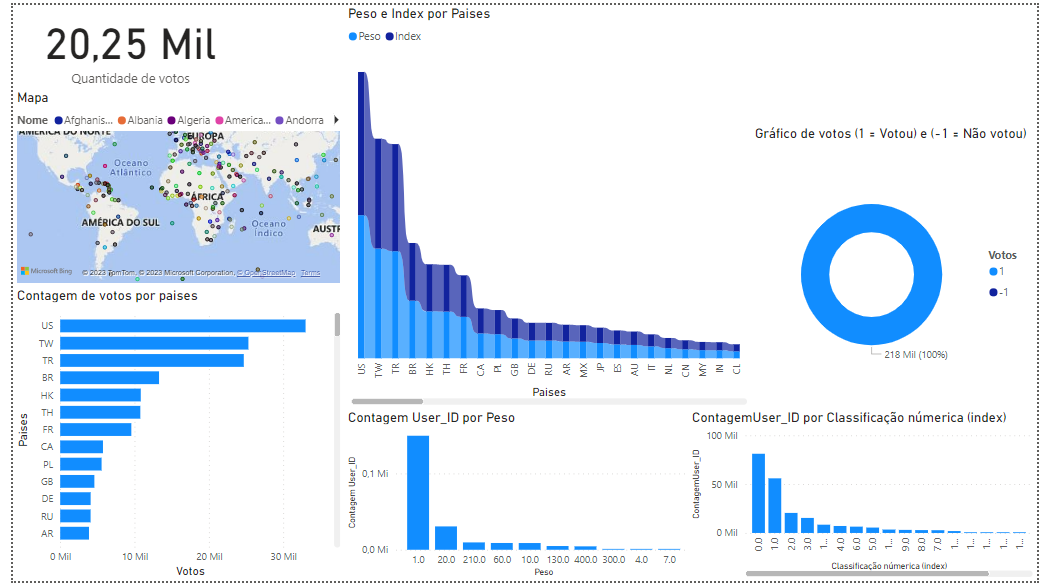
*- Foi modelado desta forma para o enriquecimento de alguns dataset e assim conseguir plotar ela gráficamente com mais detalhes representativos.*

Essa modelagem serviria pra um treinamento de um modelo de machine learning?

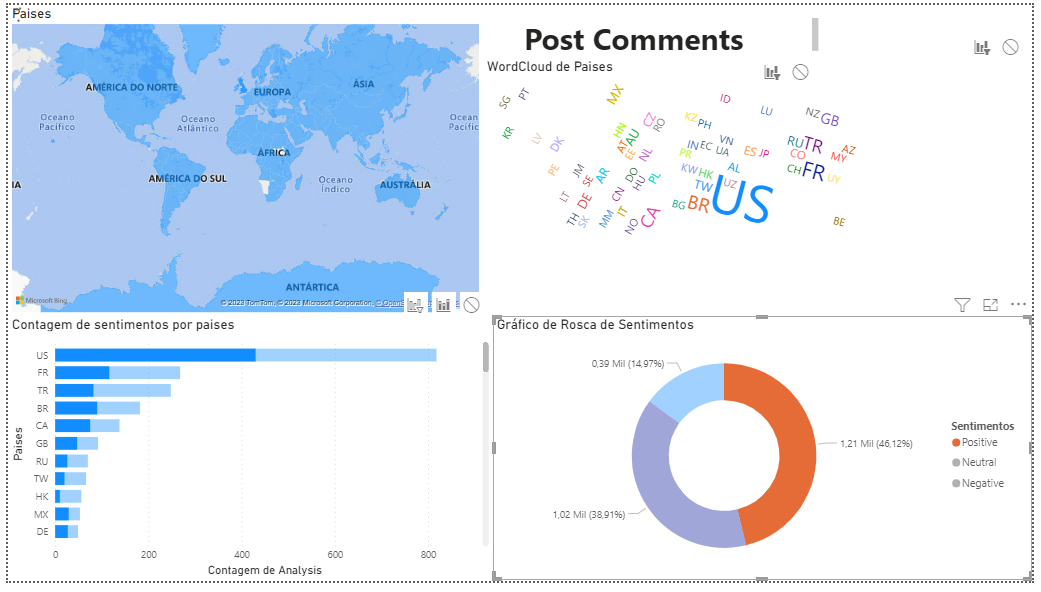
*- Serviu para fazer o treinamento de NLP (sentimentos), pois gráficamento conseguimos ver melhor paises com positividades , negatividades e neutralidades de acordo com os cometários relacionados aos games.*

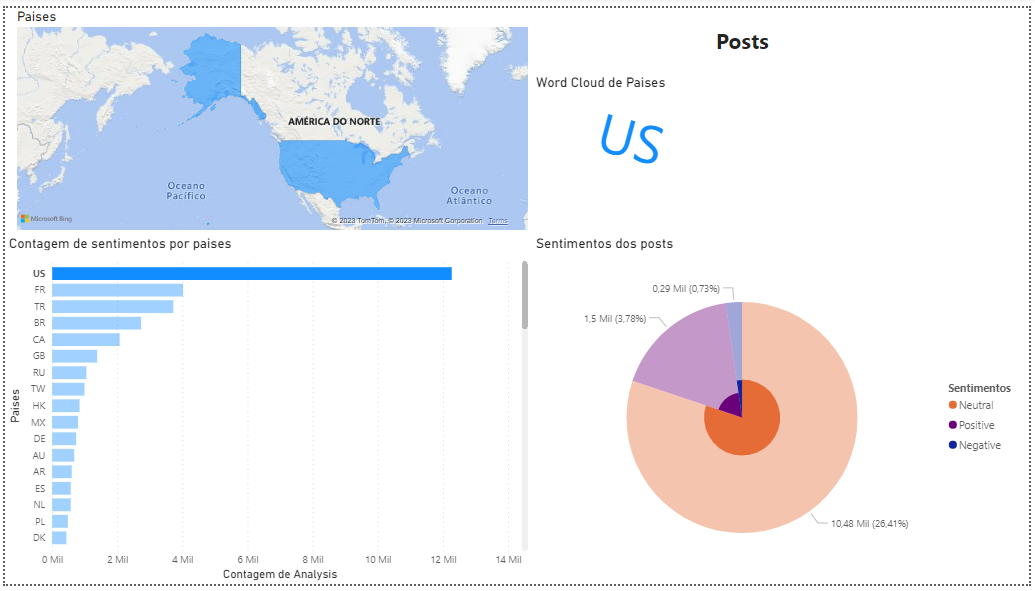
As estruturas das collections fornecidas podem ser melhoradas para uma melhor análise dos dados? Se sim, sugira melhorias para as estruturas

*- Não há muito o que fazer, pois são dados que provém diretamento dos bancos de dados e dos bancos de cometários. Talvez uma limpeza primária ajudaria, pois tornaria alguns data insight mais visiveis e assim facilitaria o processo da criação de KPI’s e também para os negócios.*



Esse painel representa o total de votos por cada país, onde 1 representa aqueles que votaram e -1 representa aqueles que não votaram. Além disso, cada classificação (índice) foi dividida de acordo com um peso de validação. Para cada país, há uma representação específica; para obtê-la, basta clicar nos países.





Tanto os comentários de posts quanto os próprios posts são painéis que representam a análise de sentimento dos comentários, podendo ser neutros, positivos ou negativos. Concomitantemente, o quadro 'nuvem de palavras' (Word Cloud) representa o país com o maior voto majoritário.