Brainstorm  
  
Se va a montar un entorno utilizando **Google Cloud**, para poder realizar el el análisis, procesado de datos y entrenamiento de los mensajes de odio que nos rodean en el día a día en las RRSS.

Esto previamente teniendo datasets, que nos servirán para entrenar al sistema y detectar dichos comportamientos en los comentarios de las redes.

En esta PoC se ha decidido optar por la red social YouTube.

Diseño del DAaaS

Definición la estrategia del DAaaS

Para este proyecto, vamos a realizar la siguiente estrategia:

* Crear una web y un software que, mediante **Google Cloud Functions**, que conectará a la API de YouTube (marco Data Source), y los descargará en “raw”, sin tratar, todos los ficheros siempre vendrán en formato CSV y se almacenarán en un bucket de **Google Cloud Storage**, y dicho **Google Cloud Storage** estará dentro de un marco SIT (Storage, Ingest, Transform).

Para ello se hará uso de las API de YouTube

* + API de YouTube (https://www.googleapis.com/youtube/v3)
* Desde la web, el usuario tendrá que introducir una URL válida de YouTube y solicitará analizar los comentarios del vídeo, esta petición lanzará el software de la **Google Cloud Functions** hacia la API de YouTube. Esta consulta a la API de YouTube se tendrá que hacer con credenciales, por lo que se tendrá que recuperar la API key de la cuenta de YouTube dentro del **Google Cloud Secret Manager**.

Una vez se ejecute dicho software y ya los datos estén en el bucket raw del **Google Cloud Storage**, mediante herramientas como **Google Colab,** con Notebooks en lenguaje Python, se podrán realizar todas las funciones del proceso ETL para tratar los datos “raw”.

* Cuando ya se han tratado los datos, se volverán a llevar a un nuevo bucket processed de **Google Cloud Storage**
* Por otro lado, y aun dentro del mismo marco SIT, se procederá a la carga de dichos comentarios en la web para poder visualizarlos, a la espera de los resultados.
* Finalmente, y ya en un nuevo marco Visualize, mediante la herramienta **Tableau**, se construirán dashboards “Ad-Hoc” para que los puedan utilizar y visualizar usuarios, viendo datos, gráficas, e incluso los conocidos emojis, muy usados para comentar.

Arquitectura DAaaS

* Fuentes de datos:
  + Datasets de terceros
  + API de YouTube
* Componentes:
  + Google Cloud Functions
  + Google Cloud Storage
  + Notebooks Python en Google Colab
  + Tableau

DAaaS Operating Model Design and Rollout

1. Delimitar que se van a tener 3 marcos
   1. El primero, con las fuentes de origen, marco Data Source
   2. A continuación, donde estará la ingesta, tratamiento y almacenamiento de los datos, marco SIT (Storage, Ingest, Transform)
   3. Finalmente, para poder visualizar todos esos datos de forma, limpia, ordenada y personalizada, tendremos el marco Visualize, que dará uno de los resultados a visualizar
2. Crear y preparar el entorno cloud con todas las instancias necesarias, como:
   1. Google Cloud Functions
   2. Google Cloud Storage
   3. Notebooks Python en Google Colab
   4. Tableau
3. El software desarrollado, consultará las API en busca de los datos solicitados.
4. Se descargarán los ficheros CSV generados en la consulta a la API y se compararán con los ya previamente entrenados de los dataset.

Link a Diagrama:

Adjunto fichero draw.io en el repositorio de la práctica:

https://github.com/PFBKCXII/petTeam/tree/main/Arquitectura