## **Data Engineer Assignment**

1. In [this link](https://drive.google.com/drive/folders/1u49DTGX-gIGnESgZFFE2z5h63r9hels9?usp=sharing) you have data of json files that contains the annotated **ground truth** information of our videos. The dataset is aimed to train a deep learning model for the task of object detection. For that purpose, different objects were recorded in different sites.
   1. The ground truth folder contains subfolders - each subfolder is the name of the site the movies were recorded in.
   2. Each subfolder contains json files with the ground truth - one json per movie. The movies are not included as they are not required for the purpose of this assignment.
   3. Each json follows the structure described in [this file](https://drive.google.com/file/d/1fJXC4eS6IlJSdunbDCSzKQc_9UhD_zzP/view?usp=sharing).   
      Notes:

* The zero frame includes a summary of the json content (all the frames in a specific movie).
* “nm” is a unique ID number, for tracking an object within a single video.
* Ignore\_frame means that for some reason, the frame wasn’t annotated.
* Attr\_key\_frame means there is a significant novelty (bug change) in this specific frame (in comparison to the previous frame).
  1. Please index this data into a database of your choosing.
  2. Explain why you chose this specific platform/engine for such a dataset
  3. Please analyze the data and find if there are any biases or deficiencies in the data. Please provide visualization of your statistical analysis that you think includes interesting information regarding the data, as well as visualization of biases in the data, if it exists.
  4. Please provide your suggestions, based on your analysis, regarding what should the next data collection events be focused on (sites, objects, objects’ attributes, etc.) in order for us to have a “richer” and more holistic dataset.

1. Create a demo data pipeline Python code that monitors an empty folder and performs the following processing:
   1. Once movies files are uploaded to this folder, the files should be copied to a predefined folder named “Raw”.
   2. Each movie placed in the “Raw” folder should be separated into frames and all frames should be copied to a predefined folder named “Train”. The frames should be saved in a \*.png format and placed under a subfolder with the name of the movie.
   3. In addition, all frames that were placed in the “Train” folder, should have a compressed copy placed in a predefined folder named “Annotation”. Here the frames should be saved in a \*.jpg format and placed under a subfolder with the name of the movie. The quality value of the jpeg compression should be a parameter of the pipeline.
2. We Would like to build a data pipeline, which needs to support numerous data collection events. We need you to design and explain the architecture and data flow for such a data pipeline:
   1. Each data collection event includes recording movies from cameras on laptops
   2. The collected data needs to be annotated by an external data annotation.
   3. The annotated data needs to be stored in the company data lake.
   4. The annotated data needs to be registered into a platform that gives us the ability to query it and monitor the data status
   5. For preview and annotation purposes a low resolution variant of the movies is sufficient. For model training we need the high resolution movies
   6. Some of the data needs to be copied to training machines for model training and some need to be transferred to the outsource annotation company.
   7. Where (infrastructure) do you plan to build the data lake?
   8. How do you upload the movies from the laptops into the data lake?
   9. How Would you protect the data in each step?

Note: please supply a location were we can get the code you have created for review and test