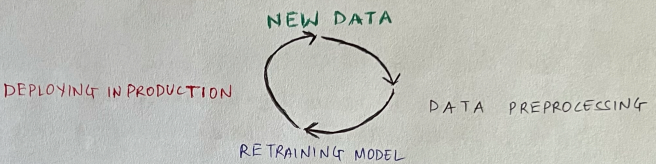
TAIYO.AI ASSIGNMENT

TASK : Our current GPT-3 fine-tuned model is only trained on static data. Our task is to ensure that the model acknowledges the new data coming in. The dataset referred is 'world_bank-preprocessed.csv'. For that we have to create a pipeline that does the job of addressing new data, Pre-processing the data, Re-training the current model such that it trains on the new data and finally deploying the model in the production environment.



STEPS : The most efficient way to create a machine learning pipeline is to use a cloud computing service but for the scope of the task we'll approach the problem differently.

1. To address the new data, we have to create a script that detects when the new data is added to the dataset. We could use Watchdog, pyinotify etc...
2. Create a fundion that triggers the preprocessing.
3. Creating a function that executes the retraining part.
4. Calling these functions in case any modifications are detected in the csv file in any 'on_modify' methods.
5. The real-time analysis can be

ASSUMPTIONS MADE:

1. The dataset 'world_bank_preprocessed.csv' is constantly being fed with raw data.
2. The batch size is set to 1000 entries per update.

INTERACTION:

1. The user has to keep the 'observer.ipynb' file (server) running in the background while using the model.
2. The user must provide the 'PROJECT NAME' and 'COUNTRY'. If the details are available in the dataset then the fine tuned model will give a brief summary of that project.

POSSIBLE FUTURE SCOPE:

1. A cloud computing service could be used to make the machine learning pipeline more efficient
2. A UI can be built where the model interacts with the user to answer the specific questions.
   Libraries that could be used:
   (i) Dash
   (ii) Flask
   (iii) tkinter