Continental
The Future in Motion

BACKGROUND STORY

Cameras in cars are recording the roads over the globe. Images are generated from these videos by taking frames in proper intervals. After Data Managers selecting the proper images, those are sent to labelling teams for annotating the visible objects.

Many labelling teams use the old labeling tool called Papaya, meanwhile some labelling teams started to use the new, cutting edge labelling tool called Kumquat. The structure of the generated output of the two software are different. Our current database was designed for storing the output of the Papaya software, but Data Managers would like to handle the result of the labelling with the two tools together.

YOUR TASKS

You can use any RDBMS system for your implementation. You can create the database tables using the attached SQL files (labels.sql and images.sql). An image is identified by its record file name and timestamp column. The labels related to that image can be identified using the same columns. For your solution please use Python and SQL. Try to use the ETL approach for the data loaders. Add unit tests where it makes sense.

- Create a data loader to load the output of the old Papaya software (papaya_output_images.csv and papaya_output_labels.csv) into the database tables.
- Create a data loader to load the output of the new Kumquat software (kumkvat_output.csv) into the database tables.

The size of the output files are getting bigger and bigger due to the ongoing labelling activities. As a result of it loading everything again and again is not on option anymore.

- Optimize your data loaders to load only the data that is not yet presented in the database (delta loading).

Sometimes images are re-labelled when new use cases are on the horizon. In that case new labels are introduced for that record file name and timestamp combination with a higher version number.

- Create a module that deletes the old version of the re-labelled images (deduplication).

The Data Operations team turns to you with a request. They would like to have the country codes in the country column, instead of country names.

- Integrate an attribute map function into the already existing data loaders to translate the value for this column.

A team member of the AI Developers team contacts you and asks information about the data in our database. He is interested in the following:

- List of images that do not have labels on them.
- For each record file list the 2 label classes with the biggest size.
- Find the record file which has the most labels with at least 0.6 size.

Please create SQL scripts answering the above questions.