**ETL Project Report**

**Project Scope:**

The U.S. has almost 500 students for every guidance counselor. Underserved youth lack the network to find their career role models, making CareerVillage.org the only option for millions of young people in America and around the globe with nowhere else to turn. To date, 25,000 volunteers have created profiles and opted in to receive emails when a career question is a good fit for them. To help students get the advice they need, the team at CareerVillage.org needs to be able to send the right questions to the right volunteers. CareerVillage.org has provided several years of anonymized data and each file comes from a table in their database. Currently data is being saved in several different CSV files so using files from various locations we cannot answer complex business questions that can be answered by ETL.

Goal of the project is to define ETL process to provide clean data for analyze questions like How well does the solution match professionals to the questions they would be motivated to answer? Final dataset will help us to **develop a method to recommend relevant questions to the professionals who are most likely to answer them.**

**Data source Link:** <https://www.kaggle.com/c/data-science-for-good-careervillage/data>

**Source API:** **kaggle competitions download -c data-science-for-good-careervillage**

Examples:

kaggle datasets download data-science-for-good-careervillage/data

kaggle datasets download data-science-for-good-careervillage/data -f answers.csv

Please note that BigQuery datasets cannot be downloaded.

**Initialize metadata file for dataset creation**

usage: kaggle datasets init [-h] [-p FOLDER]

optional arguments:

-h, --help show this help message and exit

-p FOLDER, --path FOLDER

Folder for upload, containing data files and a special dataset-metadata.json file (https://github.com/Kaggle/kaggle-api/wiki/Dataset-Metadata). Defaults to current working directory

**CSV files from data source:**

answer\_scores.csv

answers.csv

comments.csv

emails.csv

group\_memberships.csv

groups.csv

matches.csv

professionals.csv

question\_scores.csv

questions.csv

school\_memberships.csv

students.csv

tag\_questions.csv

tag\_users.csv

tags.csv

**CareerVillage.org has provided several years of anonymized data and each file comes from a table in their database.**

* **answers.csv:** Answers are what this is all about! Answers get posted in response to questions. Answers can only be posted by users who are registered as Professionals. However, if someone has changed their registration type after joining, they may show up as the author of an Answer even if they are no longer a Professional.
* **comments.csv:** Comments can be made on Answers or Questions. We refer to whichever the comment is posted to as the "parent" of that comment. Comments can be posted by any type of user. Our favorite comments tend to have "Thank you" in them :)
* **emails.csv:** Each email corresponds to one specific email to one specific recipient. The frequency\_level refers to the type of email template which includes immediate emails sent right after a question is asked, daily digests, and weekly digests.
* **group\_memberships.csv:** Any type of user can join any group. There are only a handful of groups so far.
* **groups.csv:** Each group has a "type". For privacy reasons we have to leave the group names off.
* **matches.csv:** Each row tells you which questions were included in emails. If an email contains only one question, that email's ID will show up here only once. If an email contains 10 questions, that email's ID would show up here 10 times.
* **professionals.csv:** We call our volunteers "Professionals", but we might as well call them Superheroes. They're the grown ups who volunteer their time to answer questions on the site.
* **questions.csv:** Questions get posted by students. Sometimes they're very advanced. Sometimes they're just getting started. It's all fair game, as long as it's relevant to the student's future professional success.
* **school\_memberships.csv:** Just like group\_memberships, but for schools instead.
* **students.csv:** Students are the most important people on CareerVillage.org. They tend to range in age from about 14 to 24. They're all over the world, and they're the reason we exist!
* **tag\_questions.csv:** Every question can be hashtagged. We track the hashtag-to-question pairings, and put them into this file.
* **tag\_users.csv:** Users of any type can follow a hashtag. This shows you which hashtags each user follows.
* **tags.csv:** Each tag gets a name.
* **question\_scores.csv:** "Hearts" scores for each question.
* **answer\_scores.csv:** "Hearts" scores for each answer.

ETL Process step by step process:

* 1. Read CSV file from the source (API)
  2. Create ERD diagram to build primary key and foreign key relationship.
  3. Create script to build postgres database.
  4. Import CSV files into postgres database tables.
  5. Write queries to perform analysis.
  6. Optional: write python code to read csv from local folder and add index and change date format and save file to postgres table.