Lab 5: ETL - Extract-Transfer-Load

Due date: Thursday, May 19.

This is a team lab. Each group submits one set of deliverables.

Lab Overview

This lab starts your software development in earnest. Your goal is to create the database for your project (we refer to it as the CED Database) and populate it with some information. The information will come from the following sources:

- **Data files shared with you.** The data files shared with you (Excel and CSV formats) contain information about several aspects of the database:
  - A set of questions for the undergraduate research experience survey
  - A set of questions for the job experience survey
  - A list of possible answers to each question
  - A list of users
  - A list of survey responses for each of the two surveys
  - A list of profile characteristics with descriptions
  - For each survey - a mapping from survey questions to profile characteristics

  Note: some data contained in the files shared with you requires careful parsing. However, all formats should be fairly straightforward.

- **O*NET Database.** This database ([https://www.onetcenter.org/database.html#individual-files](https://www.onetcenter.org/database.html#individual-files)) contains information about a variety of jobs. Using the web site facilities, web scraping (if necessary) and data download (O*NET data is distributed under Creative Commons license).
Assignment

Each team shall ingest data into its version of the CED Database from both sources of information.

Start by inspecting the data and mapping the information available to you from each data file/data source to the tables and specific attributes in those tables.

If you find information in the data files that is not captured in your data model (typically - an extra attribute here and there) - adjust your data model and your relational model to incorporate this information.

Conversely, if your data model calls for some additional data (attributes) associated with different entity and/or relationship sets, which you are not seeing in the data files provided to you, you have the following options:

- Consult the instructor. Reach out via Slack, identify extra information missing. Some of this information may be provided by the instructor.
- Adjust the data files accordingly. While some of the data: survey questions, survey response scales, profile characteristics are provided by the customers and are not subject to modification, other data: user accounts, surveys taken, and so on – are "fake data" created for the purpose of testing your software. You are allowed to modify any "fake data" to adjust it to fit your data model. (If in doubt - talk to the course instructor).

Build whatever scripts necessary to ingest the data into the database.

Ingesting O*NET Data

The O*NET database is bigger than the data that the CED Database needs to ingest. In general, your goal is to

- Determine the list of STEM-related jobs in the O*NET database.
- Extract the descriptions of the STEM-related jobs from the O*NET database.
- Ingest the appropriate parts of the O*NET job descriptions into the CED Database.

When deciding what information to ingest, err on the side of "more is better". Some of the data in the O*NET Job descriptions is needed to create O*NET job profiles in the CED Database (the detailed information on this is coming early next week in the next assignment), and you want to make sure that all data necessary to compute specific values of different profile characteristics is available in the CED Database.
What You Can Omit For Now

There are three sets of data in the CED Database that are built "on the fly" from other data in the CED database using a variety of algorithms. While some of these algorithms have already been discussed, for this particular assignment, placement of these data items into the CED Database is optional. You will be working on these tasks in the next part of the project.

The three sets of data are:

- **Experience profiles** associated with survey responses. The dataset actually contains these profiles for each survey response, and placing them into the CED database is fairly straightforward. However, at this point, there is no need to build software that automatically generates experience profiles from a survey response.

- **O*NET Job profiles.** We will be disclosing the algorithm for mapping O*NET job descriptions into job profiles at the end of the week/early next week, and implementation of this algorithm is going to be included as a task into your next project assignment.

- **Recommendations/Matches.** Some versions of CED databases may choose to store matches/recommendation directly in the CED database. The matching is constructed based on comparing a user experience profile/desired profile to an O*NET job profile. You will be implementing the algorithms for doing this in the next assignment.

In addition, the current data does not contain any desired profiles. You can choose to build your own, or leave creation of desired profiles off until the next assignment.

**Deliverables**

You will have some hard-copy deliverables and the software deliverable.

**Code.** Share your github repository with the instructor ("dekhtyar").

**Revisions.** Because there is a delay in providing you feedback on your data model, submit the revised E-R model and relational model with your Lab 5 deliverables (create a googledocs Lab 5 directory, and share it with the instructor).

**Database Demo.** In the first lab after the due date, demonstrate your database to the instructor.