Project 2: ETL Mini Project

Fundamentals in Extraction, Transforming, and Loading Data

Authors : Geetha Rajendram, Elisangela Jardim, Kurt Goesch, Justin DeMaagd

Background:

In today’s world, data analysis has become integral component in many facets of business operation, social awareness, and technology innovation. The more we can interpret data to better provide information to stakeholders, and other people of interest, the more productive we become, and the drive for improving daily processes can be streamlined quicker, more efficiently. In efforts to better streamline data, it is accustomed to compile raw data (the initial collection of information) into a format that is usable by others. This data needs to be relatively easy to understand, and conclusive in its results. In many cases, data may have different sources (i.e. html, API, SQL, Excel, etc.). The skills to extract data from its source, then run it through a data pipeline are greatly valued by others who need this information transformed into more comprehensible material. Data that will be better accustomed by the end user.

In this exercise we showcased our ability to extract, transform, and load data from a crowdfunding dataset. We successfully developed a methodology to present the flat data coherently into a relational database.

Methodology:

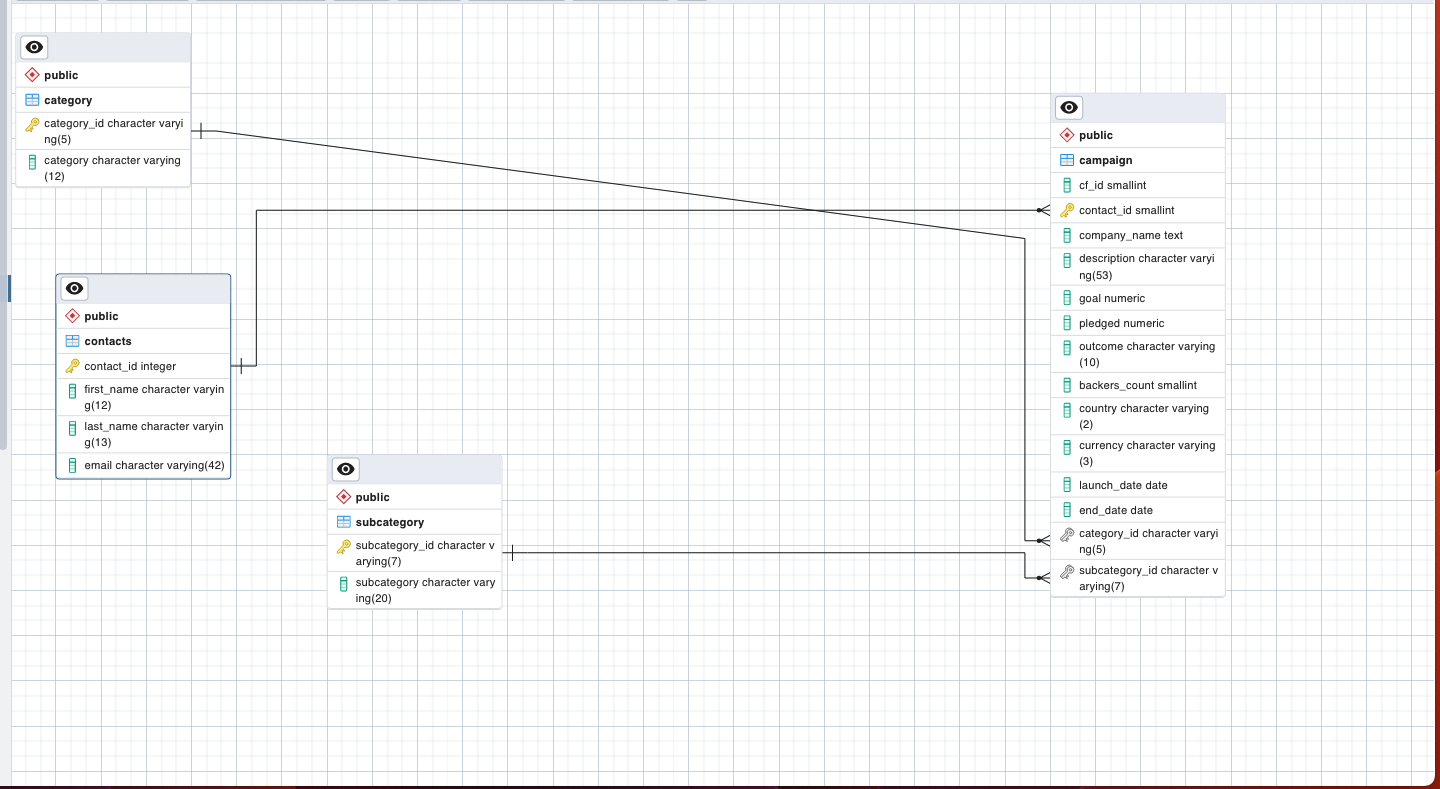
**Extract and transform crowdfunding.xls and category.xls into \*.csv**

**Utilizing the panda’s dependency, we created data frames for all:**

* Category and Subcategory
* Campaign
* Contacts
  + The Contacts table was created by Option 1 (Use of Python Dictionaries)

**Import \*csv files into SQL Server to create a relational database.**

* All \*.csv files were imported into a SQL Database scheme called Crowdfunding\_db\_schema.sql
* An entity relational diagram was created in postgres to illustrate and describe the relationships between the tables.
  + Primary keys and foreign keys were assigned appropriately.



* Database integrity was validated by running a SELECT Query for each table.

