ETL Project Proposal

Step 1: Extract

We will pull data pertaining to the retail industry and merge it together by industry code. We plan to pull the following data:

* Monthly retain trade data from Census downloaded in in csv format
  + <https://www.census.gov/retail/index.html>
* NACIS industry code and descriptions from the NAICS association.
  + <https://www.naics.com/six-digit-naics/?code=44-45>.
  + We will scrape this information using Pandas
* <https://www.bls.gov/iag/tgs/iag441.htm>
* <https://www.bls.gov/iag/tgs/iag44-45.htm>

Step 2: Transform

Each of the data files we pull have NAICS codes. However, they are not all at the same level or in the same format. We expect to perform the following transformations:

* Collapsing and aggregating detailed industry codes to summary levels
* Joining data by NAICS code
* Dropping or otherwise labeling NAICS codes that do not appear in all data files.

Step 3: Load

* Our initial plan is to load the data into a Postgres SQL database because two of our three team members have professional experience with SQL.
* We will also investigate options for uploading data to AWS, time permitting.