Guidelines for ETL Project

This document contains guidelines, requirements, and suggestions for Project 2.

Team Effort

Due to the short timeline, teamwork will be crucial to the success of this project! Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week.

Working in a group enables you to tackle more difficult problems than you'd be able to working alone. In other words, working in a group allows you to **work smart** and **dream big**. Take advantage of it!

Project Proposal

Before you start writing any code, remember that you only have one week to complete this project. View this project as a typical assignment from work. Imagine a bunch of data came in and you and your team are tasked with migrating it to a production database.

Take advantage of your Instructor and TA support during office hours and class project work time. They are a valuable resource and can help you stay on track.

Finding Data

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

- data.world
- Kaggle

You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

Data Cleanup & Analysis

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

- The sources of data that you will extract from.
- The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).
- The type of final production database to load the data into (relational or non-relational).
- The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

Project Report

At the end of the week, your team will submit a Final Report that describes the following:

- **E**xtract: your original data sources and how the data was formatted (CSV, JSON, MySQL, etc) The source of data used were the following websites: https://www.fire.ca.gov/incidents/, https://www.ncdc.noaa.gov/cdo-web/search
    - PyMongo, CSV, BeautifulSoup, created lists, dictionaries, APIs

  Transform: what data cleaning or transformation was required -

    - used splinter to move through webpages by scraping data through loops,
    - Remove auto-inserted _id key for unique object ID on each iteration of for loop (fix a py mongo code) active_fire_dict.pop('_id', None) - to get around a loop that will try to stop a for loop thinking that it repeats a step unnecessarily that is needed
    - Limitations of number of weather stations that can be downloaded from the website - manually downloaded stations names and created API request for historical data for the weather stations.
    - Store latitude, longitude elements from web scraping to use in API call for weather forecasts of fire areas.

- **L**oad: the final database, tables/collections, and why this was chosen.
  Mongodb for ease of use with storing various data sources as collections without

primary/foreign key relationships, JSON format for easy interpretation when working with Javascript during the front-end portions of project 2.

Please upload the report to Github and submit a link to Bootcampspot.

ETL Project Topic: **California Wildfires**

Data sources-

1. https://www.fire.ca.gov/incidents/
    a. Web scrape data for 2019-2013 as historic_fire_data collection
    b. Web scrape data for active as current_fire_data collection
2. https://www.ncdc.noaa.gov/cdo-web/search
    a. Pull 2 year historical weather data for all weather stations in California
    b. Will be helpful in identifying trends in rising temperatures, drought conditions, etc for areas commonly affected by
3. https://visual-crossing-weather.p.rapidapi.com/forecast
    a. Weather API for 5 day forecasts
4. https://www.kaggle.com/annieichen/top-20-largest-california-wildfires/data
    a. Top 20 California Wildfires CSV

MongoDB Schema-

```
> show dbs
admin     0.000GB
config    0.000GB
local     0.000GB
wildfire  0.000GB
> use wildfire
switched to db wildfire
> show collections
Active_wildfires
ca_wildfires_2019
largest_fires
weather_forecast
```