**Dataset1:**

Yelp users give rating and provide reviews about businesses on Yelp. The dataset is a subset of businesses , reviews and user data. These files are available in json format.

High Level pipeline steps -

Get the static raw data files - yelp\_academic\_dataset\_business, yelp\_academic\_dataset\_checkin.json, yelp\_academic\_dataset\_review.json, yelp\_academic\_dataset\_tip.json, yelp\_academic\_dataset\_user.json

I want to create a data lake which will have both processed and unprocessed data. The data can reside in local db , later can be moved to cloud.

Write a python script to collect the raw data from JSON files and load into PostgreSQL staging database(later moved to the Cloud)

Data Exploration after data is loaded into a database. Data cleansing may be required after loading the raw data in tables. Will come up with validation, cleansing rules if needed.

Load incremental data from APIs for testing upsert functionality (depending on structure of API and the static dataset) . At this point Business Search API can fetch about 1000 records at a time. Useful link-

<https://www.yelp.com/developers/documentation/v3/get_started>

/businesses/matches – can also be used to pull businesses based on restaurant name, address form another dataset

Upsert logic – update a flag is a business is open or permanently closed. Using the is\_open flag.

Write spark jobs for ELT for reading data from staging zone, apply transformation, and load into processed zone.

Load the data in distributed environment.(Not sure at this point how it works). Also, may have to fake data for volume.

Load the data in Cloud(AWS)

Data pipeline Orchestration – To be updated

**Datset2:**

Use census data to get demographic information of a location(details to be provided after data exploration)

Another idea is to use Facebook graph API to pull list of restaurants in NYC and then find the reviews from Yelp.