ReAlpha Case Study

By Spencer Vilicic
https://github.com/quence-dev/realpha-case-study

Chosen Data Sources

- AllRecipes webscraping - full demonstration
- Quandl API - partial demonstration (no db interaction)
- Homesnap webscraping - partial demonstration (limited results)

Quandl API

- Most straightforward to interact with, while the data itself requires some knowledge (i.e. must know what you're looking for)
- Can use standard API requests or Quandl's custom python module

Homesnap (Webscraping)

- Most challenging data source with active measures to prevent web scraping
- Randomly generated IDs and classes, hidden elements, and no discernable request URLs in their network when loading search results

AllRecipes (Webscraping)

- More straightforward approach for webscraping with consistent IDs, classes, etc.
- Bottleneck comes from loading each page individually and scraping results (could be sped up with multithreading)
- Still unable to load more than initial page of 24 results

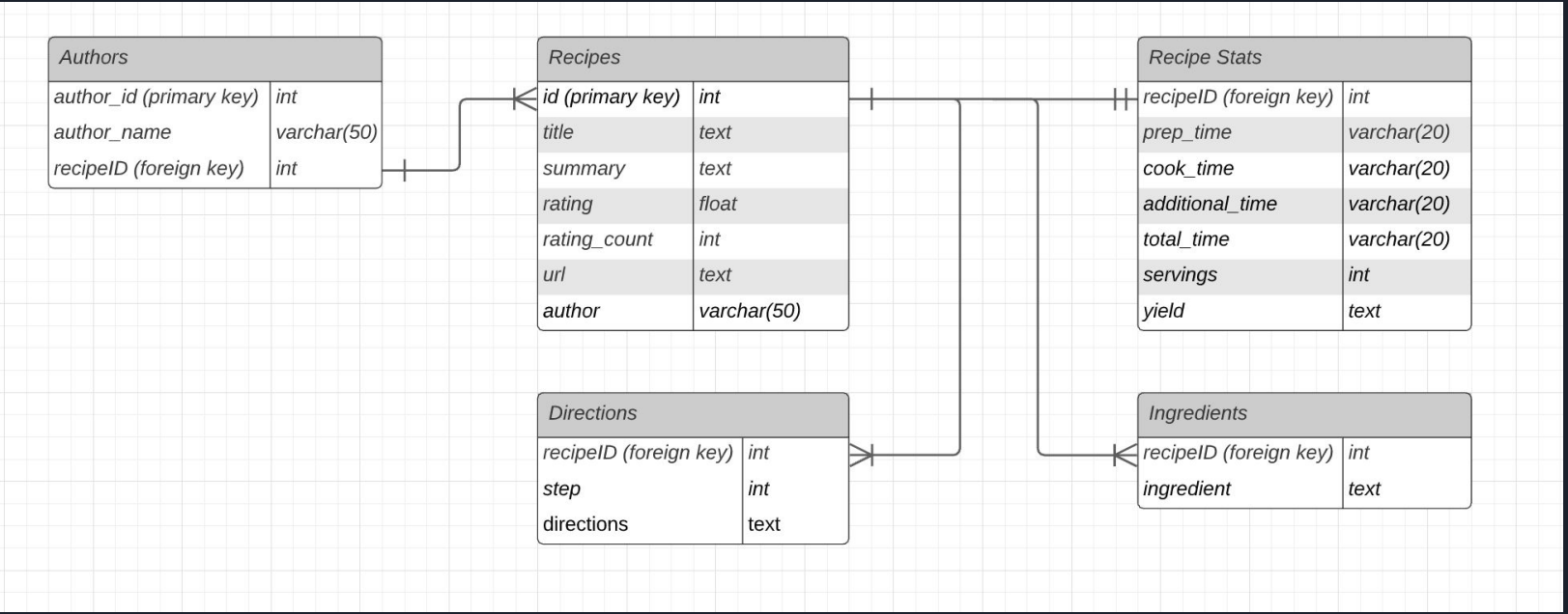
Raw data.json

[{"title": "Easy Mac and Cheese Pizza", "summary": "We loved Cici's pizza so much that we made our own mac and cheese pizza! It is quick and easy!", "rating": 4.83, "rating_count": 3, "url": "https://www.allrecipes.com/recipe/237150/easy-mac-and-cheese-pizza/", "author": "JMU Jen", "metadata": {"prep": "10 mins", "cook": "20 mins", "additional": "5 mins", "total": "35 mins", "servings": "4", "yield": "4 servings", "ingredients": ["1 (12 inch) pre-baked pizza crust", "\u00be cup cavatappi (corkscrew macaroni)", "\u2154 (16 ounce) jar cheese sauce (such as Ragu\u00ae Double Cheddar), divided, or as needed", "1 tablespoon butter", "salt and ground black pepper to taste", "\u00bd cup shredded Cheddar cheese"], "directions": ["Preheat oven to 450 degrees F (230 degrees C). Place pizza crust on a baking sheet.", "Bring a small pot of lightly salted water to a boil. Cook cavatappi in the boiling water, stirring occasionally, until cooked through but firm to the bite, 10 to 11 minutes. Drain and return pasta to the pot.", "Stir 1/2 the cheese sauce and butter into the cavatappi pot over medium heat until butter is melted and pasta-cheese mixture is combined. Season with salt and pepper.", "Spread a thin layer of cheese sauce over the pizza crust. Sprinkle shredded Cheddar cheese over the cheese sauce onto the crust. Pour pasta-cheese mixture over the Cheddar cheese, and top with another 1/2 the cheese sauce.", "Bake pizza in the preheated oven until golden and bubbling, 8 to 10 minutes. Let rest

Processed example.json


```
[
  {
    "recipe_title": "Macaroni and Cheese Pizza Bake",
    "total_time": "55 mins"
  },
  {
    "recipe_title": "Honey-Roasted Carrot and Goat Cheese Pizza",
    "total_time": "40 mins"
  },
  {
    "recipe_title": "Fig and Goat Cheese Pizza",
    "total_time": "1 hr 18 mins"
  },
  {
    "recipe_title": "Pear and Gorgonzola Cheese Pizza",
    "total_time": "25 mins"
  },
  {
    "recipe_title": "Mac-N-Cheese Pizza",
    "total_time": "55 mins"
  },
  {
    "recipe_title": "1-Dish Pepperoni Cheese Pizza Bake",
    "total_time": "50 mins"
  },
  {
    "recipe_title": "Apple Cheese Pizza",
    "total_time": "40 mins"
  },
```

Recipe ERD



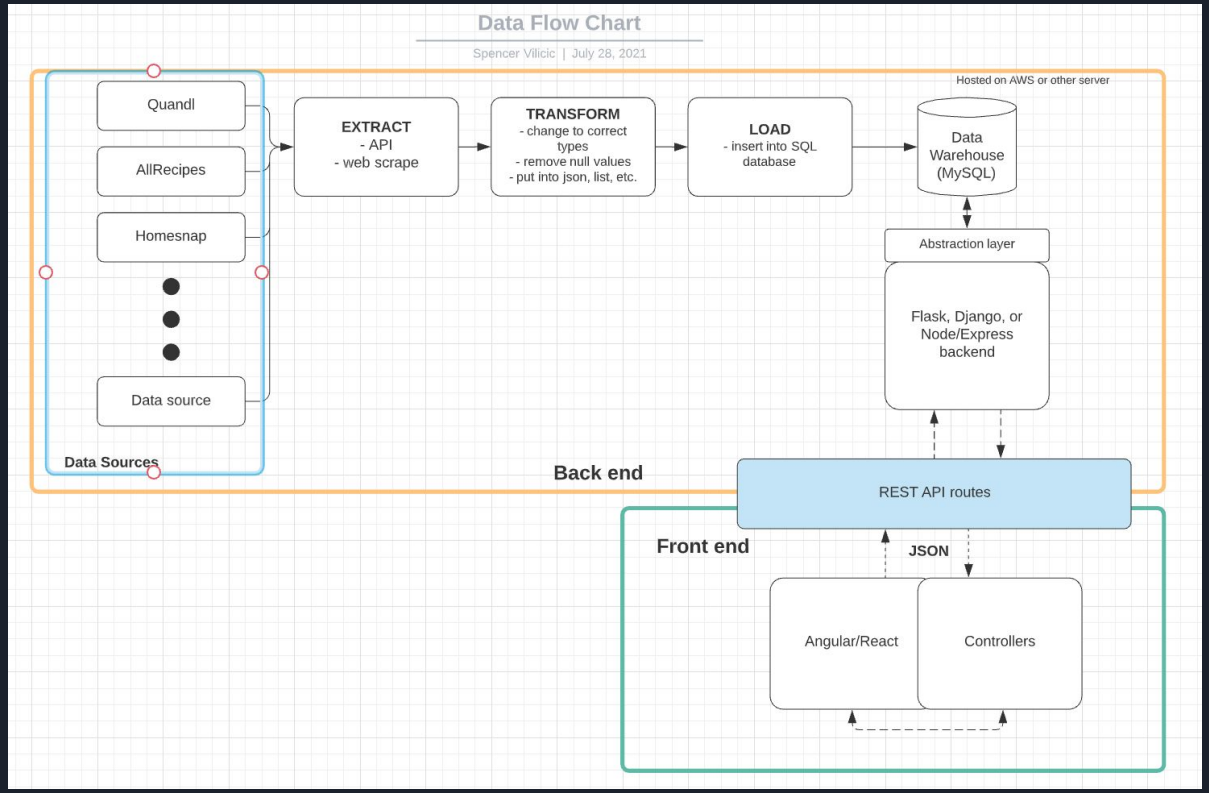
https://lucid.app/lucidchart/invitations/accept/inv_9f241866-2a42-4a68-82c2-229fa8ca026a?viewport_loc=169%2C-33%2C1714%2C827%2C0_0

Raw homedata.json


```
{"datatable": {"data": [["ZSFH", "10001", "2021-05-31", 202951.0], ["ZSFH", "10001",
"2021-04-30", 200827.0], ["ZSFH", "10001", "2021-03-31", 198750.0], ["ZSFH", "10001",
"2021-02-28", 195494.0], ["ZSFH", "10001", "2021-01-31", 195764.0], ["ZSFH", "10001",
"2020-12-31", 197303.0], ["ZSFH", "10001", "2020-11-30", 197721.0], ["ZSFH", "10001",
"2020-10-31", 196277.0], ["ZSFH", "10001", "2020-09-30", 186197.0], ["ZSFH", "10001",
"2020-08-31", 187253.0], ["ZSFH", "10001", "2020-07-31", 186914.0], ["ZSFH", "10001",
"2020-06-30", 97918.0], ["ZSFH", "10001", "2020-05-31", 97436.0], ["ZSFH", "10001", "2020-04-30",
96703.0], ["ZSFH", "10001", "2020-03-31", 96307.0], ["ZSFH", "10001", "2020-02-29", 95999.0],
["ZSFH", "10001", "2020-01-31", 95921.0], ["ZSFH", "10001", "2019-12-31", 96008.0], ["ZSFH",
"10001", "2019-11-30", 96028.0], ["ZSFH", "10001", "2019-10-31", 96199.0], ["ZSFH", "10001",
"2019-09-30", 96065.0], ["ZSFH", "10001", "2019-08-31", 96075.0], ["ZSFH", "10001", "2019-07-31",
95772.0], ["ZSFH", "10001", "2019-06-30", 95512.0], ["ZSFH", "10001", "2019-05-31", 95038.0],
["ZSFH", "10001", "2019-04-30", 95160.0], ["ZSFH", "10001", "2019-03-31", 95414.0], ["ZSFH",
"10001", "2019-02-28", 95347.0], ["ZSFH", "10001", "2019-01-31", 94898.0], ["ZSFH", "10001",
"2018-12-31", 93923.0], ["ZSFH", "10001", "2018-11-30", 93171.0], ["ZSFH", "10001", "2018-10-31",
92330.0], ["ZSFH", "10001", "2018-09-30", 91774.0], ["ZSFH", "10001", "2018-08-31", 91071.0],
["ZSFH", "10001", "2018-07-31", 90453.0], ["ZSFH", "10001", "2018-06-30", 89773.0], ["ZSFH",
"10001", "2018-05-31", 89321.0], ["ZSFH", "10001", "2018-04-30", 88814.0], ["ZSFH", "10001",
"2018-03-31", 88167.0], ["ZSFH", "10001", "2018-02-28", 87736.0], ["ZSFH", "10001", "2018-01-31",
87261.0], ["ZSFH", "10001", "2017-12-31", 87131.0], ["ZSFH", "10001", "2017-11-30", 86878.0],
```

ETL Flow Chart

https://lucid.app/lucidchart/invitations/accept/inv_d5bf3314-392e-43da-af9a-7f5e86b75f08?viewport_loc=-221%2C-7%2C2325%2C1121%2C0_0



Next Steps

- Front end and REST API for making calls
- Optimization (multithreading, caching database hits)
- More purposeful data processing to prep for machine learning / AI

Reflection

- If I hit a roadblock, I would move to a different datasource rather than sticking with one. This might’ve slowed me down in the long run
- Webscraping gets easier the more I do it
- Still not sure when I’m “finished” without clear end in mind
- Barebones data processing, much to learn there

Thank you!