Search Engine For Restaurants

Group 8: Eric Chen, Nan Chen, Rong Huang, Wenyu Chen, Zhipeng Zhang

Background and Definition of the Data Analytics Problem

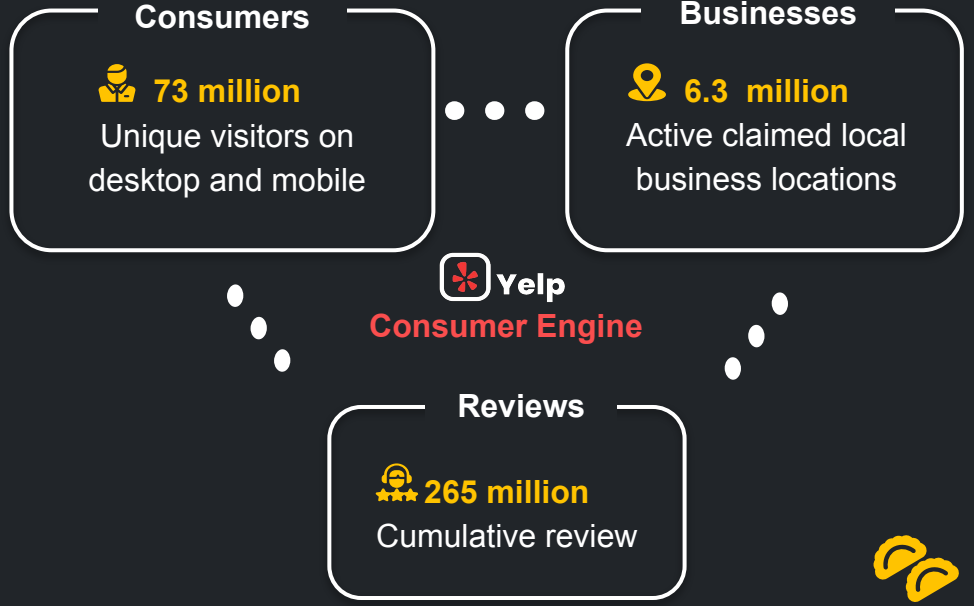
Background

Yelp, the user-driven reviews app, helps customers make decisions about where to spend their money and helps businesses improve from customer feedbacks. “**About one-third of customers rely on online reviews when choosing a restaurant** and over half of 18 to 34-year-olds factor reviews into their dining decisions” (Nakayama and Wan, 2018).

Data Analytics Problem

We are planning to develop **a search engine** for restaurant information, the average rating score, and top 3 reviews based on characteristics (such as funny, cool and useful). Users can search for reviews based on restaurant information (name, zip code, state, city, stars and categories).

Mission: Connecting people with great local business



Sources: https://www-sciencedirect-com.ezproxy.cul.columbia.edu/science/article/pii/S0261517717302388
https://www.yelp-ir.com/

Data Source Specification: Yelp Open Dataset

Procurement way: directly downloading from *https://www.yelp.com/dataset*

- Yelp offers two alternatives: connect to API or download the dataset directly. Through connecting to API, there is limit for the number of results we can one-time download, 50. And a region has to be specified, like a state or longitude and latitude. Since our goal is to cover the restaurants across US, so we choose to download the dataset directly to mitigate the risk of leaving omittance.

Details of Yelp Open Dataset

- Includes: *business.json, review.json, user.json, checkin.json, tip.json, photo.json*
- These 6 json file contain over 6,990,280 reviews for 150,346 business in 11 metropolitan areas, and 200,100 corresponding photos

Data Source Specification: Yelp Open Dataset

We will use *business.json* and *review.json*

➢ From the business file, we will access the information of restaurants, and mapping by the business_id, we will access the average rating and the top 3 reviews based on the reviews' usefulness (we define it as the sum of number of useful/funny/cool votes received).

review.json		
column name	**data type**	**description**
review_id	string	22 character unique review id
user_id	string	22 character unique user id, maps to the user in user.json
business_id	string	22 character business id, maps to business in business.json
stars	integer	star rating
date	string	date formatted YYYY-MM-DD
text	string	the review itself
useful	integer	number of useful votes received
funny	integer	number of funny votes received
cool	integer	number of cool votes received

business.json		
column name	**data type**	**description**
business_id	string	22 character unique string business id
name	string	the business's name
address	string	the full address of the business
city	string	the city
state	string	2 character state code, if applicable
postal code	string	the postal code
latitude	float	latitude
longitude	float	longitude
stars	float	star rating, rounded to half-stars
review_count	integer	number of reviews
is_open	integer	0 or 1 for closed or open, respectively
attributes	object	business attributes to values
categories	array	an array of strings of business categories
hours	object	an object of key day to value hours, hours are using a 24hr clock

Technical Scheme

Back-end technologies

- **Data Processing Tool:** Spark – Batch Processing
  - Overall data size is not small (over 5GB).
  - Distributed computing framework.
- **Database Type:** MongoDB
  - Store JSON format data which are relatively complex
  - Horizontal scaling
  - Fast Search and Indexing via Map-Reduce support
- **CAP Analysis** - We choose CP
  - The frequency of updates to restaurant information is low, the location of restaurants does not change frequently.
  - The use of distributed computing framework Spark and horizontally scalable database MongoDB makes partition tolerance critical for the system's fault tolerance.
  - The compromise of availability is acceptable, given that the consistency of data is more important for the search engine's accuracy and relevance.

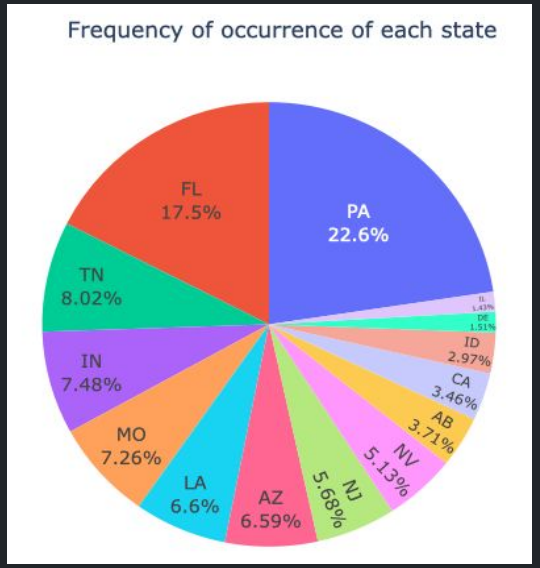
Front-end technologies

- **HTML/CSS/JavaScript:** Build the user interface of the web application
- **Flask:** Build back-end APIs for communication with the database and the front-end.

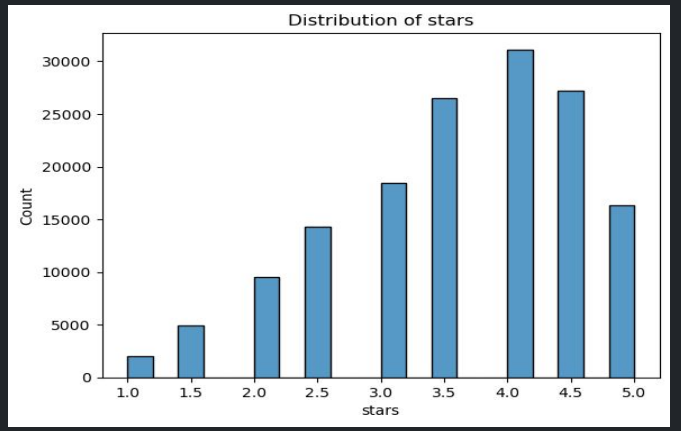
Explore Data

- **Number of Observations:**
  business: 150,346
  review: 6,990,280
  new_df: 150,346

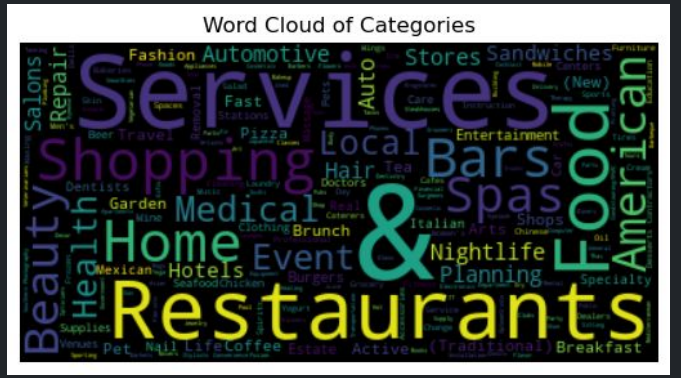
- Majority of business entities comes from Philadelphia and Florida



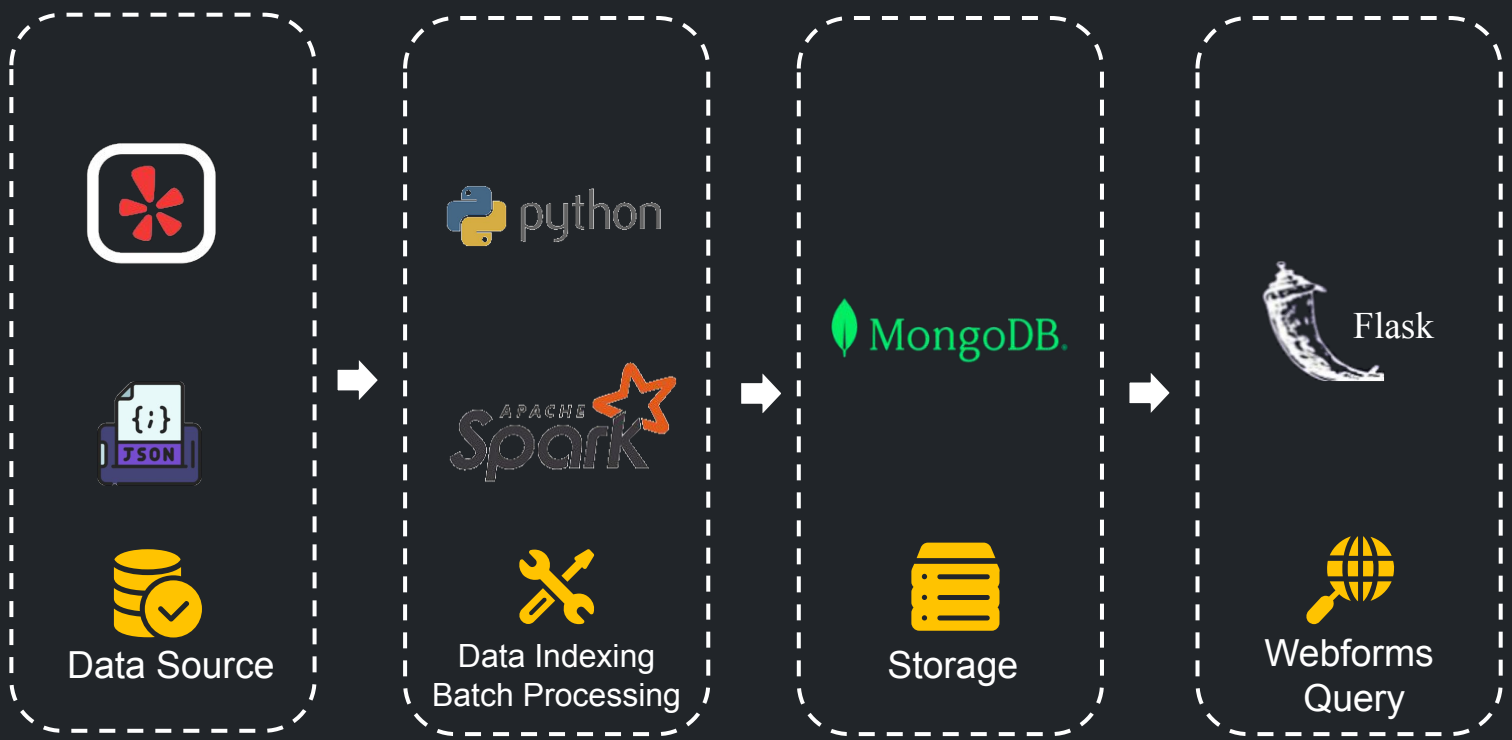
- Distribution of business' stars



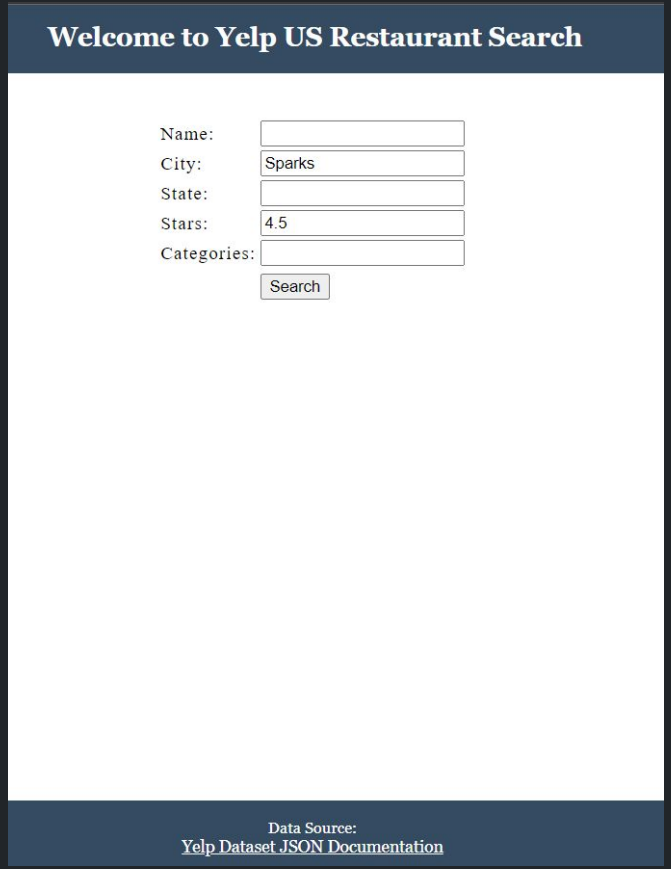
- Most of business are restaurants and services

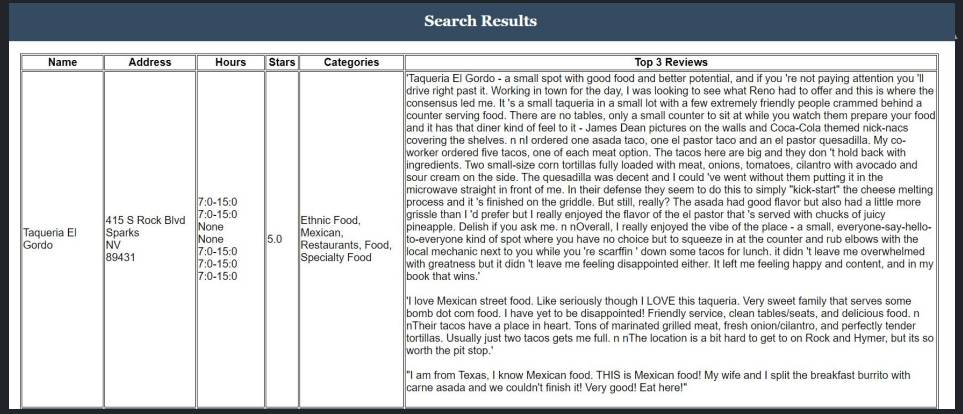


ETL Pipeline: Spark & Data Warehouse: MongoDB



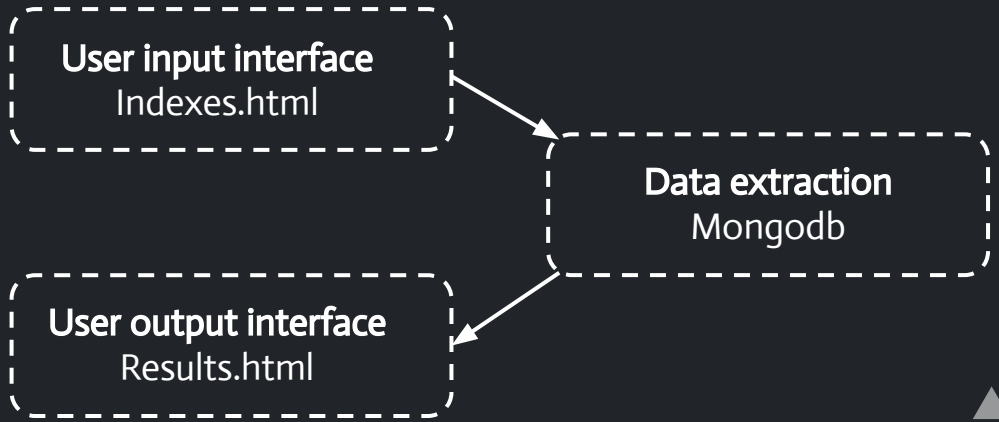
Flask API





Search Results

Name	Address	Hours	Stars	Categories	Top 3 Reviews
Taqueria El Gordo	415 S Rock Blvd Sparks NV 89431	7:0-15:0 7:0-15:0 None None 7:0-15:0 7:0-15:0 7:0-15:0	5.0	Ethnic Food, Mexican, Restaurants, Food, Specialty Food	'Taqueria El Gordo - a small spot with good food and better potential, and if you 're not paying attention you 'll drive right past it. Working in town for the day, I was looking to see what Reno had to offer and this is where the consensus led me. It 's a small taqueria in a small lot with a few extremely friendly people crammed behind a counter serving food. There are no tables, only a small counter to sit at while you watch them prepare your food and it has that diner kind of feel to it - James Dean pictures on the walls and Coca-Cola themed nick-nacs covering the shelves. n nI ordered one asada taco, one el pastor taco and an el pastor quesadilla. My co-worker ordered five tacos, one of each meat option. The tacos here are big and they don 't hold back with ingredients. Two small-size corn tortillas fully loaded with meat, onions, tomatoes, cilantro with avocado and sour cream on the side. The quesadilla was decent and I could 've went without them putting it in the microwave straight in front of me. In their defense they seem to do this to simply "kick-start" the cheese melting process and it 's finished on the griddle. But still, really? The asada had good flavor but also had a little more grissle than I 'd prefer but I really enjoyed the flavor of the el pastor that 's served with chucks of juicy pineapple. Delish if you ask me. n nOverall, I really enjoyed the vibe of the place - a small, everyone-say-hello-to-everyone kind of spot where you have no choice but to squeeze in at the counter and rub elbows with the local mechanic next to you while you 're scarffin ' down some tacos for lunch. it didn 't leave me overwhelmed with greatness but it didn 't leave me feeling disappointed either. It left me feeling happy and content, and in my book that wins.' 'I love Mexican street food. Like seriously though I LOVE this taqueria. Very sweet family that serves some bomb dot com food. I have yet to be disappointed! Friendly service, clean tables/seats, and delicious food. n nTheir tacos have a place in heart. Tons of marinated grilled meat, fresh onion/cilantro, and perfectly tender tortillas. Usually just two tacos gets me full. n nThe location is a bit hard to get to on Rock and Hymer, but its so worth the pit stop.' "I am from Texas, I know Mexican food. THIS is Mexican food! My wife and I split the breakfast burrito with carne asada and we couldn't finish it! Very good! Eat here!"



Scalability and Cost Implications

Cost Implication: We are using publicly available data with signed agreement “YELP DATASET TERMS OF USE”.

- Amazon DocumentDB cost
  - Server usage (180 hours/month)
  - Instance Type (db.r6g.large)
  - Storage (10 GB) , I/O (1 million)
  - Price: $48.57 / month
- Metrics tracking cost to measure the success of product: **$20** / month
- Total Estimated Cost: **$68.57** / month

Scalability:

- We estimate our platform user base will reach approximately 0.01% of the amount of Yelp monthly user, around **17,800** as our initial user base.
- To handle the traffic will incur, we will take horizontal scaling to cope with new demands and query optimization to return results faster.

Evaluation criteria including quantitative and qualitative success metrics

Qualitative metrics

- Pop-up page & Chat window to collect customer feedback
- Survey emails sent to subscribers

Quantitative metrics

- Click through rate
- Time spent on each search
- Monthly active users amount
- User rating in the scale from 1 to 10
- New users per month
- Customer subscription rate

Sources: https://techjury.net/blog/yelp-statistics/#gref

Data Quality Dimensions & Licensing

- **Completeness**: All records are unique but the location distribution of businesses is uneven. Businesses in some states may not be found because of data missing.
- **Consistency**: We stored our data into MongoDB after joining into one table, thus the consistency can be guaranteed.
- **Accuracy**: We obtained data from Yelp. The reviews are written by customers based on dining experience and the business information is provided by restaurants and verified by Yelp.
- **Relevancy**: Comments made by customers and business information could be utilized in building our search engine.
- **Quantity**: There are 150,346 business records and 6,990,280 review records. We only consider top 3 helpful reviews so there are sufficient reviews to support the filter.
- **Accessibility**: We will set data access rights for users.

Conclusion & Recommendation

Conclusion:

- Utilizing the ETL process of extracting, transforming, and loading data to build a search engine on top of the data warehouse.
- Facilitating users in rapidly retrieving information about restaurants and gaining insights into them from a customer perspective using other customers' reviews.
- Ensuring consistency and partition tolerance.

Recommendation:

- Employing Yelp Fusion API to obtain real-time data.
- Fuzzy searching could be helpful when searching for information that may have multiple spellings or alternative phrasings.
- Inclusion of images and a restaurant reservation website would enhance the system's usability and convenience.
- Improving the system based on feedback received from customers during its operation.
- Storing data on cloud based platform such as AWS could be a practical option to explore in case of limited computational capabilities.




Thanks for listening!

Q & A