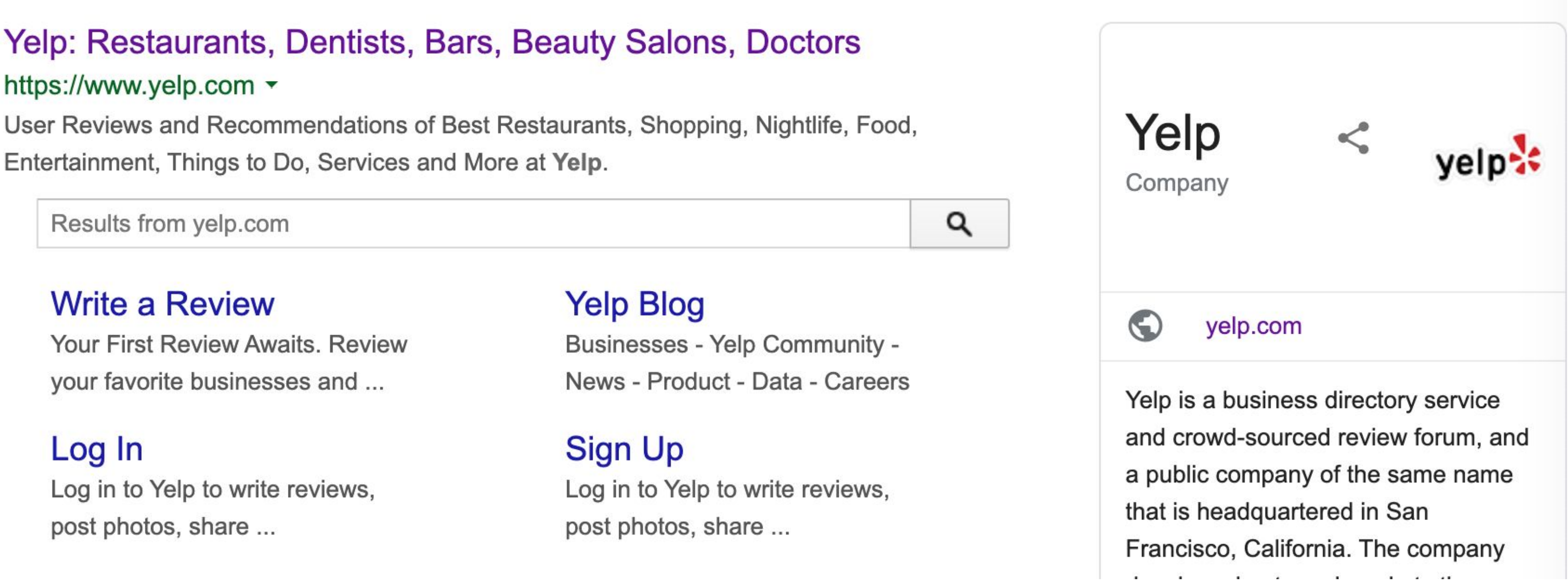
YelpReview

Yen

OUTLINE

1. What’s Yelp?
2. Yelp dataset
3. Data modeling
4. Data process
5. APP demo & Data analysis

What’s Yelp : A crowd-sourced review forum



Yelp: Restaurants, Dentists, Bars, Beauty Salons, Doctors

https://www.yelp.com

User Reviews and Recommendations of Best Restaurants, Shopping, Nightlife, Food, Entertainment, Things to Do, Services and More at **Yelp**.

Results from yelp.com

Write a Review

Your First Review Awaits. Review your favorite businesses and ...

Yelp Blog

Businesses - Yelp Community - News - Product - Data - Careers

Log In

Log in to Yelp to write reviews, post photos, share ...

Sign Up

Log in to Yelp to write reviews, post photos, share ...

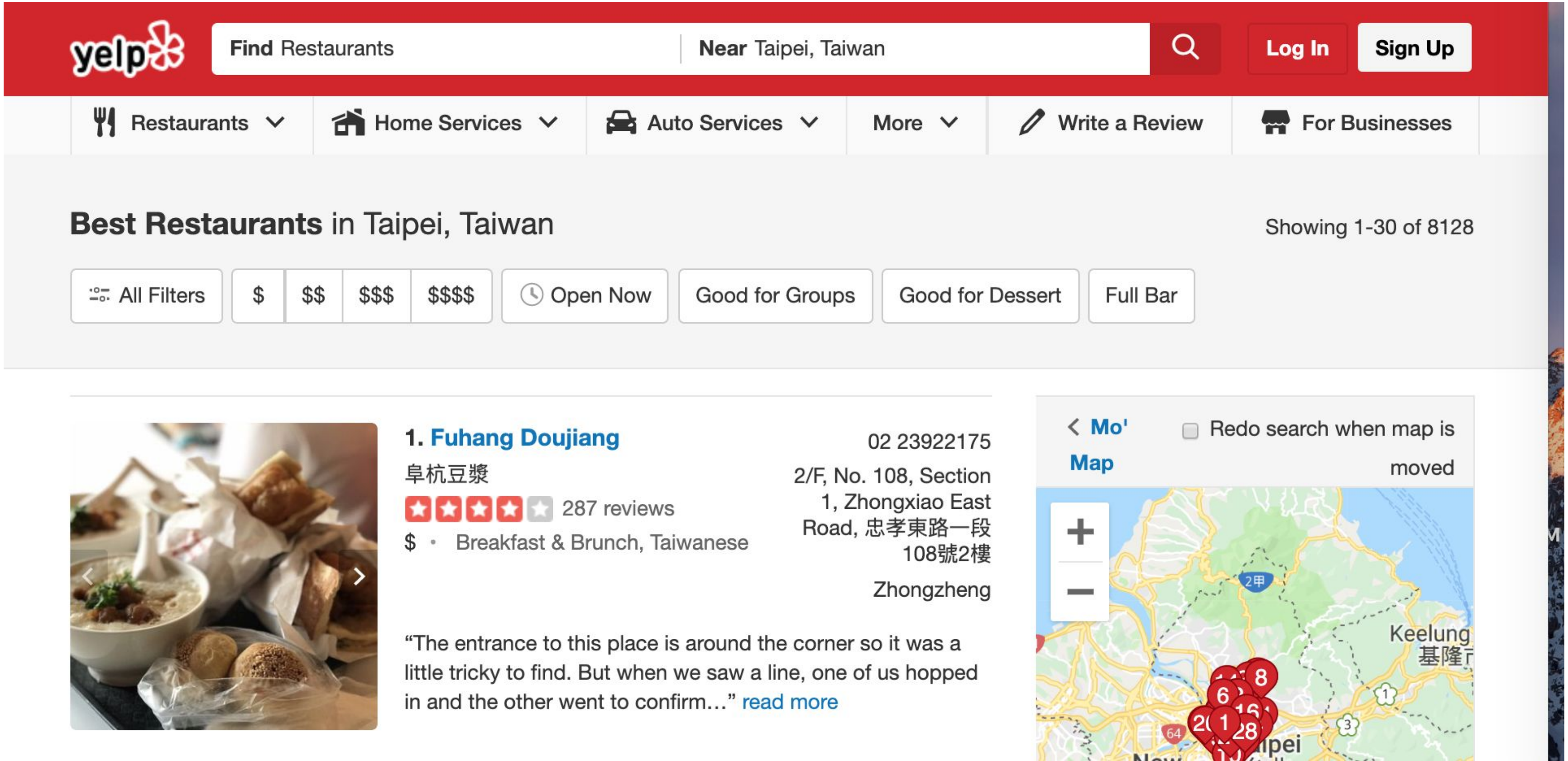
Yelp

Company

yelp.com

Yelp is a business directory service and crowd-sourced review forum, and a public company of the same name that is headquartered in San Francisco, California. The company

What’s Yelp : Business (e.g. restaurants, bar..)



Best Restaurants in Taipei, Taiwan

Showing 1-30 of 8128

1. **Fuhang Doujiang**

阜杭豆漿

★★★★☆ 287 reviews

$ • Breakfast & Brunch, Taiwanese

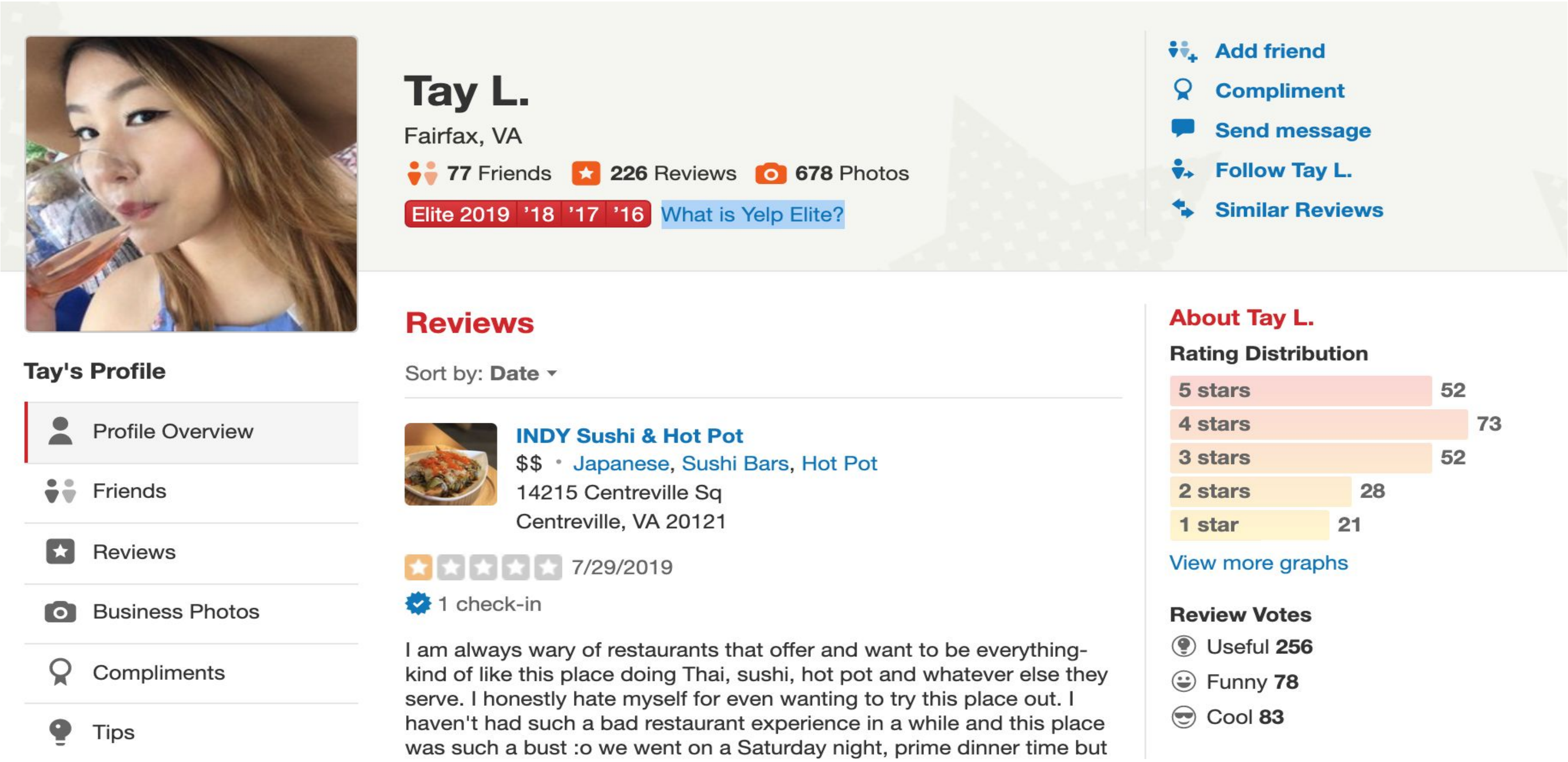
02 23922175

2/F, No. 108, Section 1, Zhongxiao East Road, 忠孝東路一段108號2樓

Zhongzheng

“The entrance to this place is around the corner so it was a little tricky to find. But when we saw a line, one of us hopped in and the other went to confirm…” read more

What’s Yelp : User (profile, comment, friends)



Tay L.

Fairfax, VA

77 Friends 226 Reviews 678 Photos

Elite 2019 ’18 ’17 ’16 What is Yelp Elite?

Tay's Profile

Reviews

Sort by: Date

INDY Sushi & Hot Pot

$$ • Japanese, Sushi Bars, Hot Pot

14215 Centreville Sq
Centreville, VA 20121

★☆☆☆☆ 7/29/2019

1 check-in

I am always wary of restaurants that offer and want to be everything-kind of like this place doing Thai, sushi, hot pot and whatever else they serve. I honestly hate myself for even wanting to try this place out. I haven't had such a bad restaurant experience in a while and this place was such a bust :o we went on a Saturday night, prime dinner time but

About Tay L.

Rating Distribution

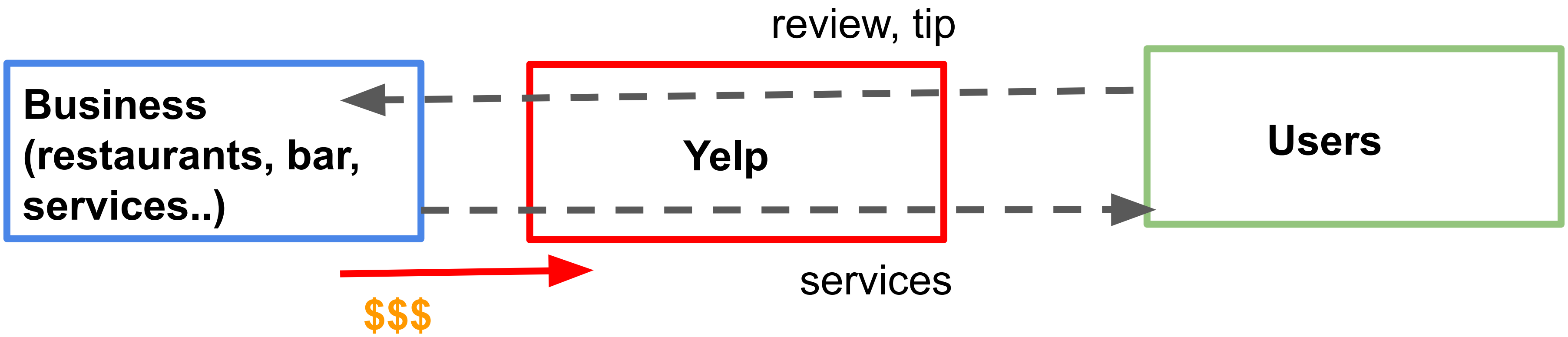
5 stars	52
4 stars	73
3 stars	52
2 stars	28
1 star	21

View more graphs

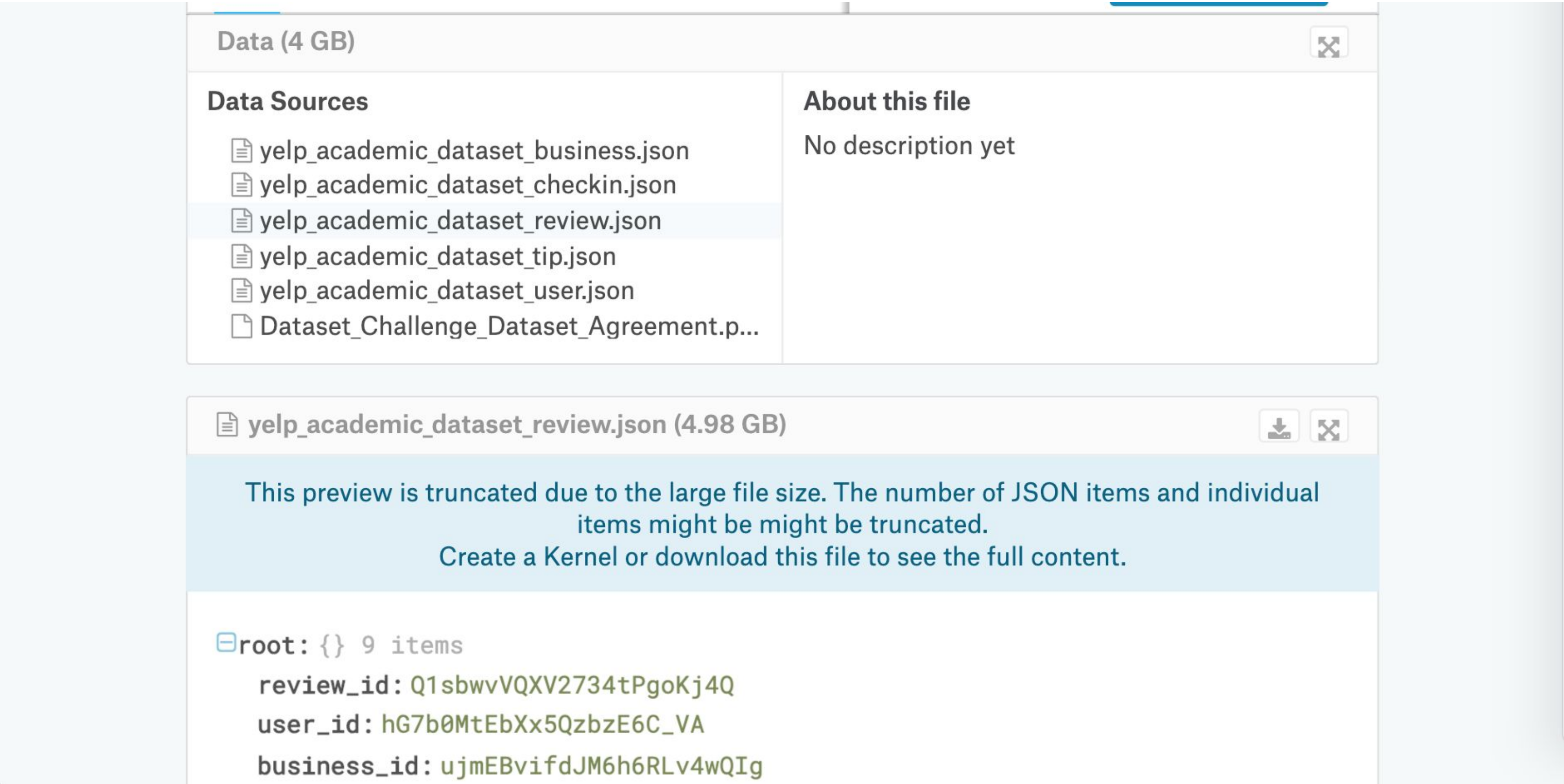
Review Votes

Useful	256
Funny	78
Cool	83

What’s Yelp : Yelp model



Yelp Dataset



Data (4 GB)

Data Sources

- yelp_academic_dataset_business.json
- yelp_academic_dataset_checkin.json
- yelp_academic_dataset_review.json
- yelp_academic_dataset_tip.json
- yelp_academic_dataset_user.json
- Dataset_Challenge_Dataset_Agreement.p...

About this file

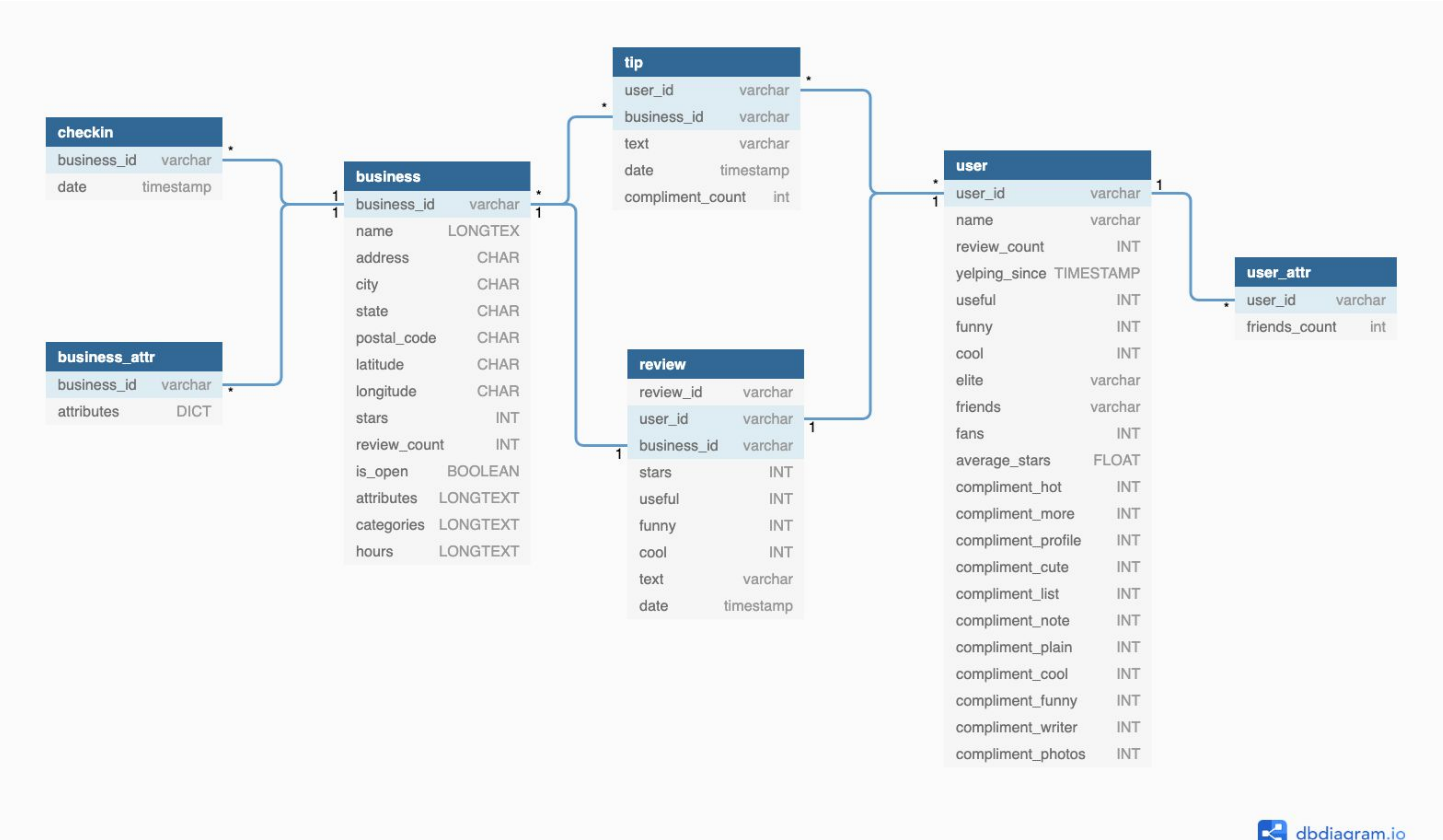
No description yet

yelp_academic_dataset_review.json (4.98 GB)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated.
Create a Kernel or download this file to see the full content.

```
⊟root: {} 9 items
  review_id: Q1sbwvVQXV2734tPgoKj4Q
  user_id: hG7b0MtEbXx5QzbzE6C_VA
  business_id: ujmEBvifdJM6h6RLv4wQIg
```

Data Modeling : snowflake pattern



checkin	
business_id	varchar
date	timestamp

business_attr	
business_id	varchar
attributes	DICT

business	
business_id	varchar
name	LONGTEX
address	CHAR
city	CHAR
state	CHAR
postal_code	CHAR
latitude	CHAR
longitude	CHAR
stars	INT
review_count	INT
is_open	BOOLEAN
attributes	LONGTEXT
categories	LONGTEXT
hours	LONGTEXT

tip	
user_id	varchar
business_id	varchar
text	varchar
date	timestamp
compliment_count	int

review	
review_id	varchar
user_id	varchar
business_id	varchar
stars	INT
useful	INT
funny	INT
cool	INT
text	varchar
date	timestamp

user	
user_id	varchar
name	varchar
review_count	INT
yelping_since	TIMESTAMP
useful	INT
funny	INT
cool	INT
elite	varchar
friends	varchar
fans	INT
average_stars	FLOAT
compliment_hot	INT
compliment_more	INT
compliment_profile	INT
compliment_cute	INT
compliment_list	INT
compliment_note	INT
compliment_plain	INT
compliment_cool	INT
compliment_funny	INT
compliment_writer	INT
compliment_photos	INT

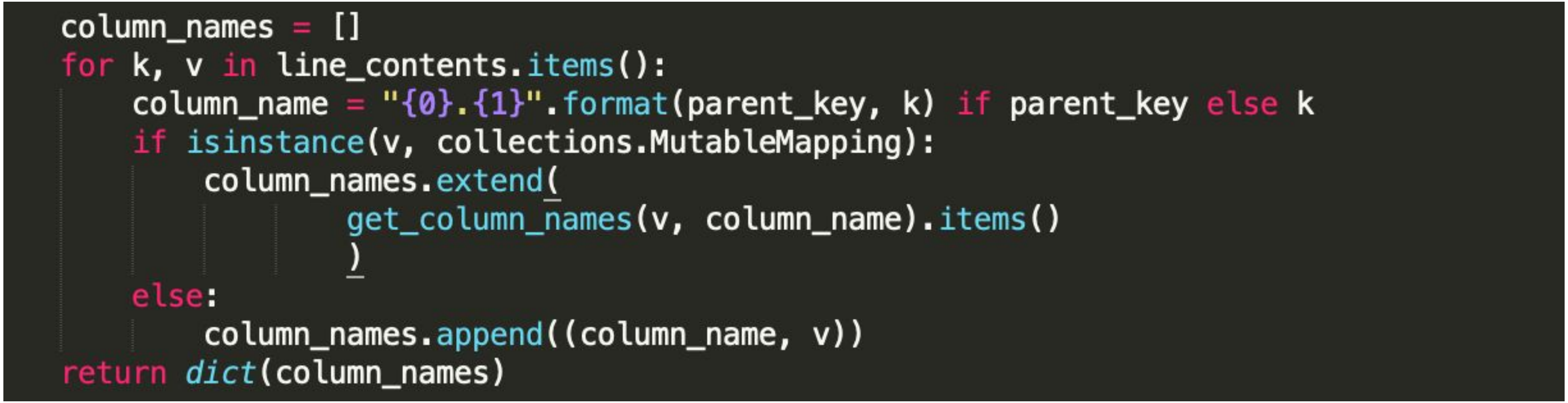
user_attr	
user_id	varchar
friends_count	int

Data Modeling

1. “SNOWFLAKE” pattern
2. Review table as “fact” table at center, connected with other tables (as “dimension” table)
    a. review
    b. user, business, tip, checkin
3. Attribution table (via ETL) connected to dimension table
    a. business_attr
    b. user_attr

Data Process

1. JSON -> csv -> fixed csv -> mysql
    a. flatten json to csv
    b. fix potential outlier data
    c. insert fixed csv data to mysql

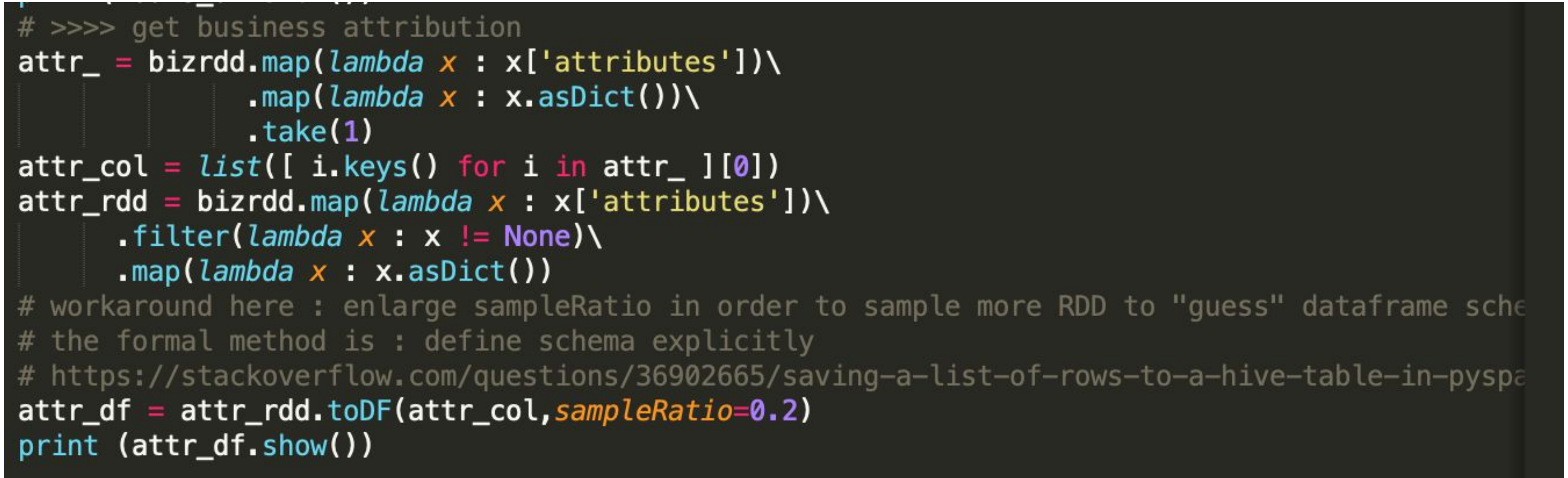


```
column_names = []
for k, v in line_contents.items():
    column_name = "{0}.{1}".format(parent_key, k) if parent_key else k
    if isinstance(v, collections.MutableMapping):
        column_names.extend(
                get_column_names(v, column_name).items()
                )
    else:
        column_names.append((column_name, v))
return dict(column_names)
```

Data Process

2. JSON -> attr csv -> mysql

   a. Transform data form make it easy to access via Spark SQL
   b. Doing aggregation/statistics on complex nest json via Spark
   c. Insert attribution data to mysql
   d. Run spark task via Docker



```
# >>>> get business attribution
attr_ = bizrdd.map(lambda x : x['attributes'])\
              .map(lambda x : x.asDict())\
              .take(1)
attr_col = list([ i.keys() for i in attr_ ][0])
attr_rdd = bizrdd.map(lambda x : x['attributes'])\
        .filter(lambda x : x != None)\
        .map(lambda x : x.asDict())
# workaround here : enlarge sampleRatio in order to sample more RDD to "guess" dataframe sche
# the formal method is : define schema explicitly
# https://stackoverflow.com/questions/36902665/saving-a-list-of-rows-to-a-hive-table-in-pyspa
attr_df = attr_rdd.toDF(attr_col,sampleRatio=0.2)
print (attr_df.show())
```

App Demo & Ananlysis