BIZTRENDER

Search & Visualization on Yelp Dataset

INSTRUCTOR : Prof James Abello

TA : Mr Fangda Han

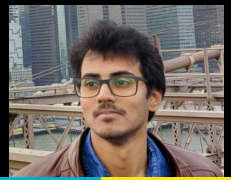
GRADER : Mr Nishant Kumar

GROUP 9:

ABHISHEK
ab2083

CHAITANYA
cd817

PRANEETH
pt357



THE DATASET

Public data from Yelp Dataset Challenge
https://www.yelp.com/dataset

Raw data uncompressed - 8.6 GB

Business.json - 192, 609 records

Checkin.json - 161, 950 records

Review.json - 6, 685, 900 records

Tip.json - 1, 223, 094 records

User.json - 1, 637, 138 records

Data at rest

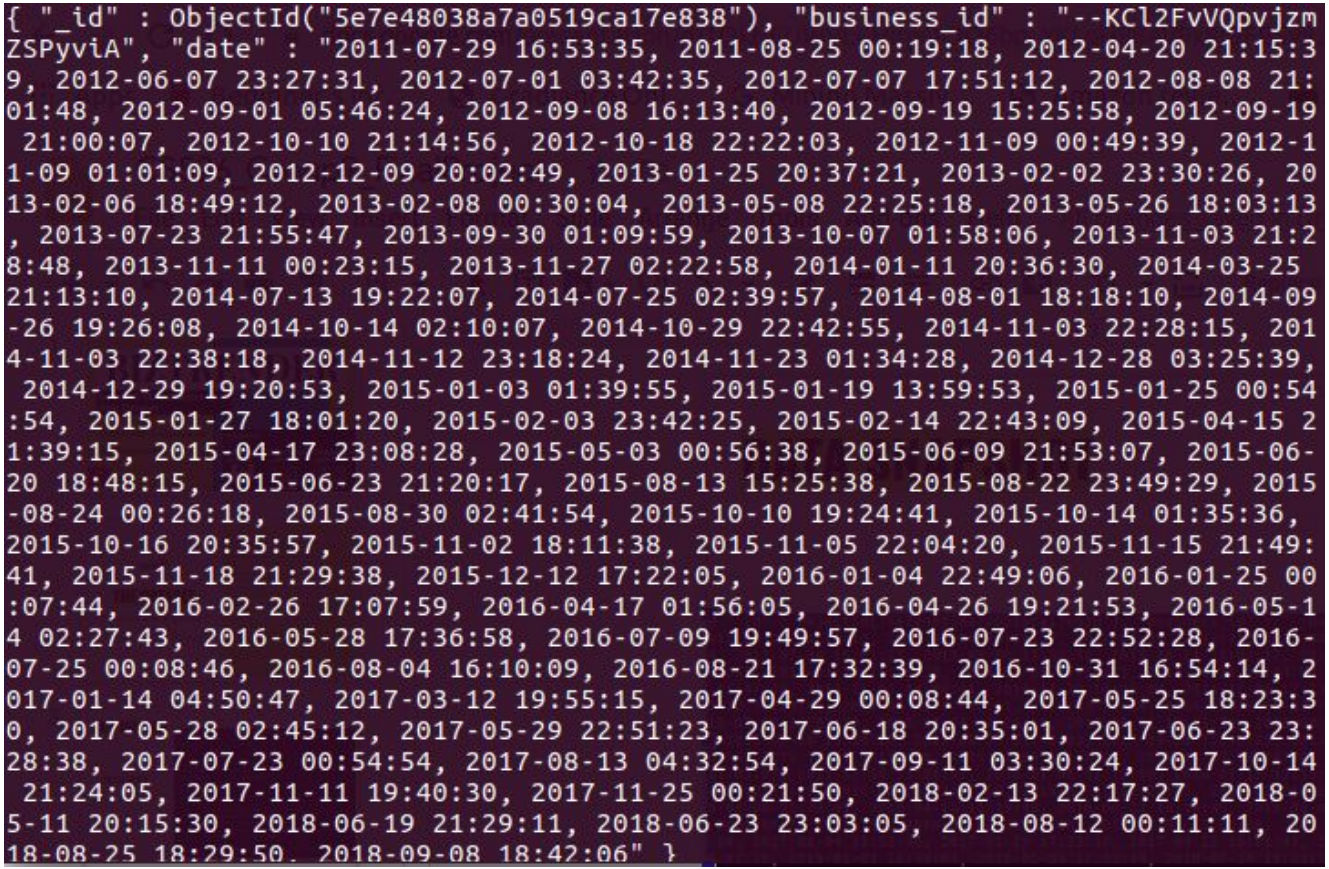
DATA SNAPSHOT

BUSINESS


```
"business_id" : "6KgGE8B1RsR7jc9R5nuH0Q",
"name" : "Ruby Tuesday",
"address" : "4 E University Dr",
"city" : "Tempe",
"state" : "AZ",
"postal_code" : "85281",
"latitude" : 33.422192,
"longitude" : -111.939615,
"stars" : 2.5,
"review_count" : 9,
"is_open" : 0,
"attributes" : {
        "RestaurantsTakeOut" : "True",
        "RestaurantsPriceRange2" : "2",
        "OutdoorSeating" : "False",
        "BusinessAcceptsCreditCards" : "True",
        "Alcohol" : "'full_bar'"
},
"categories" : "American (Traditional), Restaurants",
"hours" : null
```

DATA SNAPSHOT

CHECKIN



DATA SNAPSHOT

REVIEW


```
{
        "_id" : ObjectId("5e7e48408a7a0519ca1a60c9"),
        "review_id" : "2TzJjDVDEuAW6MR5Vuc1ug",
        "user_id" : "n6-Gk65cPZL6Uz8qRm3NYw",
        "business_id" : "WTqjgwHlXbSFevF32_DJVw",
        "stars" : 5,
        "useful" : 3,
        "funny" : 0,
        "cool" : 0,
        "text" : "I have to say that this office really has it together, they ar
e so organized and friendly!  Dr. J. Phillipp is a great dentist, very friendly
and professional.  The dental assistants that helped in my procedure were amazin
g, Jewel and Bailey helped me to feel comfortable!  I don't have dental insuranc
e, but they have this insurance through their office you can purchase for $80 so
mething a year and this gave me 25% off all of my dental work, plus they helped
me get signed up for care credit which I knew nothing about before this visit!
I highly recommend this office for the nice synergy the whole office has!",
        "date" : "2016-11-09 20:09:03"
}
```


TIP

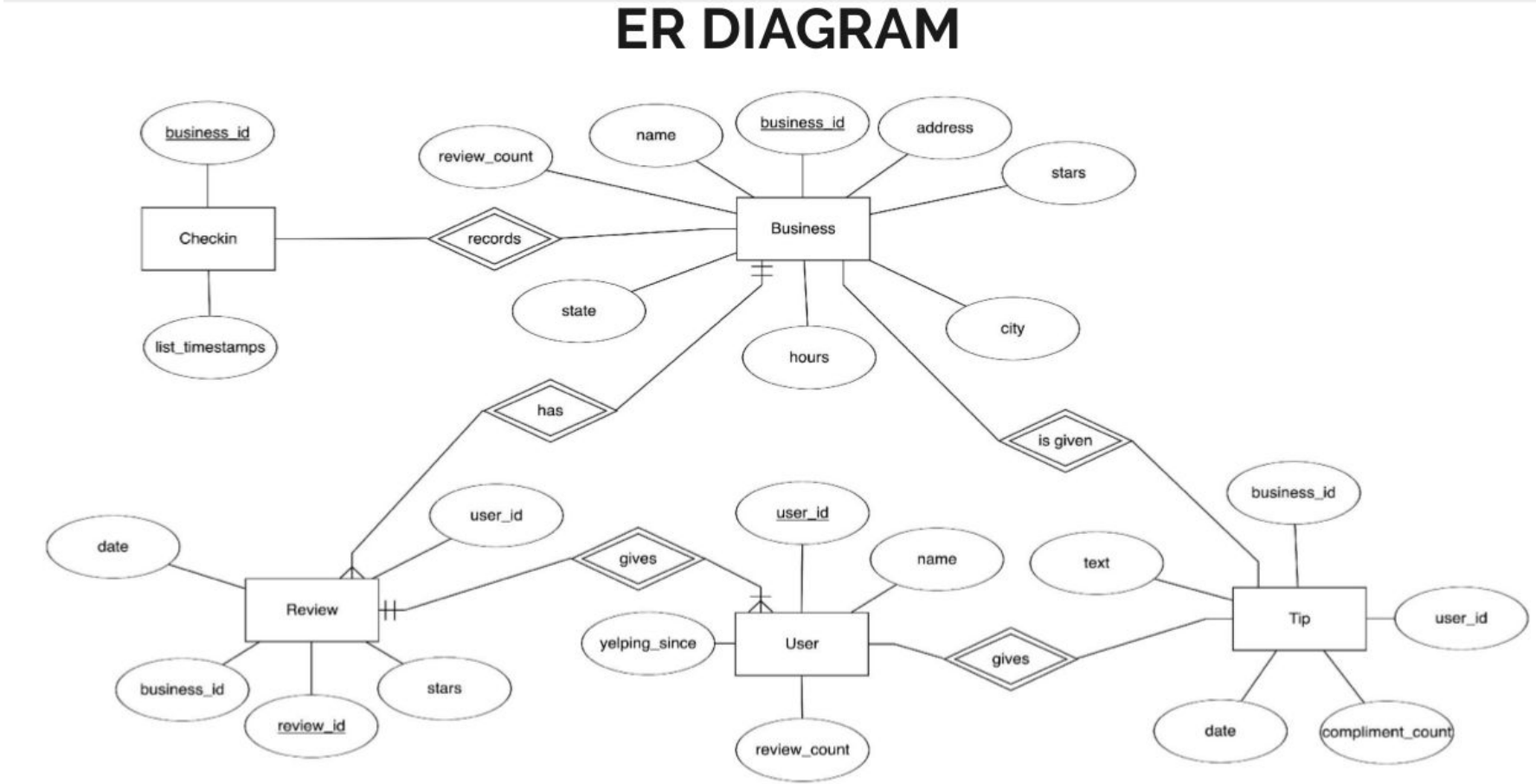

```
{
        "_id" : ObjectId("5e7e4ab88a7a0519ca806599"),
        "user_id" : "8zXgNCkusOHMwh5Tj_8yCQ",
        "business_id" : "qrSsS0pk7SL67MP5nN8tlg",
        "text" : "More parking in the back of the restaurant",
        "date" : "2013-12-22 05:10:12",
        "compliment_count" : 0
}
```

DATA SNAPSHOT

USER


```
{
        "_id" : ObjectId("5e7e4b038a7a0519ca930f4a"),
        "user_id" : "l6BmjZMeQD3rDxWUbiAiow",
        "name" : "Rashmi",
        "review_count" : 95,
        "yelping_since" : "2013-10-08 23:11:33",
        "useful" : 84,
        "funny" : 17,
        "cool" : 25,
        "elite" : "2015,2016,2017",
        "friends" : "c78V-rj8NQcQjOI8KP3UEA, alRMgPcngYSCJ5naFRBz5g, ajcnq75Z5xxkvUSmmJ1bCg, BSMAmp2-wMzCkhTfq9ToNg, jka10dk9ygX76hJG0gfPZQ, du
t0e4xvme7QSlesOycHQA, l4l5lBnK356zBua7B-UJ6Q, 0HicMOOs-M_gl2eO-zES4Q, _uI57wL2fLyftrcSFpfSGQ, T4_Qd0YWbC3co6WSMw4vxg, iBRoLWPtWmsI1kdbE9ORSA, x
jrUcid6Ymq0DoTJELkYyw, GqadWVzJ6At-vgLzK_SKgA, DvB13VJBmSnbFXBVBsKmDA, vRP9nQkYTeNioDjtxZlVhg, gT0A1iN3eeQ8EMAjJhwQtw, 6yCWjFPtp_AD4x93WAwmnw,
1dKzpNnib-JlViKv8_Gt5g, 3Bv4_JxHXq-gVLOxYMQX0Q, ikQyfu1iViYh8T0us7wiFQ, f1GGltNaB7K5DR1jf3dOmg, tgeFUChlh7v8bZFVl2-hjQ, -9-9oyXlqsMG2he5xIWdLQ,
 Adj9fBPVJad8vSs-mIP7gw, Ce49RY8CKXVsTifxRYFTsw, M1_7TLi8CbdA89nFLlH4iw, wFsNv-hqbW_F5-IRqfBN6g, 0Q1L7zXHocaUZ2gsG2XJeg, cBFgmOCBdhYa0xoFEAzp_g
, VrD_AgiFvzqtlR15vir3SQ, cpE-7HK514Sr5vpSen9CEQ, F1UYelhPFB-zIKlt0ygIZg, CQAL1hvsLMCzuJf9AglsXw, 1KnY1wr15WfEWIRLB9IS6g, QWFQ-kXBiLbid-lm5Jr3d
Q, nymT8liFugCrM16lTy0ZfQ, qj69bdd885heDvUPCyHd2Q, DySCZZcgbdrlHgEovk5y9w, lZMJIDuvhT9Dy4KyquLXyA, b_9Gn7wS93AoPZPR0dIJqQ, N07g1IaLh0_6sUjtiSRe
4w, YdfPX_7DxSnKvvdCJ57iOw, 8GYryZPD22W7WgQ8kvMkEQ, cpQmAgOWatghp14h1pn1dQ, EnchhymLYMqftCRjqvVWmw, -JdfKhFktE7Zs9BMDFcPeQ, uWhC9eof98zPkvsalga
qJw, eyTlNDDaiPatfe6mheIZ0g, VfHq0o73aKsODvfAhwAQtg, kvD5tICngLAaQDujSFWupA, dXacwEhqi9-3_XT6JeH0Og, NfU0zDaTMEQ4-X9dbQWd9A, cTHWBdjSKbctSUIvWs
gFxw, 3IEtCbSDF5t7RkZ20T6s9A, HJJXTrp6UybYyPdQ9DA0JA, JaXogQFVjzGRAeBvzamBHg, NUonfKkjS1iVqnNITtgXZg, D5vaJAYp0sOrGfsj9qvsMA, H27Ecbwwu4FGAlLgI
Courw, S8HrLmMiE4u8FWYWkNEoTw, Io36Y3xWQcIX9rYvPcYfXQ, J5mcqh8KxYpqjaLBNlwcig, -nTB3_08g06fD0GT8AtDBQ, wMpFA46lihK8oFns_5p65A, RZGFJHeomGJCWp3x
cL3ejA, ZoQSzzXoSP1RxOD4Amv9Bg, qzM0EB0SkuuGIFv0adjQAQ, HuM6vvuveken-fPZ7d4olA, H3oukHpGpn9n_mJwSDSQyQ, PkmsJsQ8FIZe8eh8c_u96g, wSByVbwME4MzgkJ
aFyfvNg, YEVqknoDmrHAoUbHX0nPnA, li3vsK1XAPmeJYAUTYflHQ, MKc8yXi0glbPYt0Qb4PECw, fQPH6W9fksi27gkuUPnFaA, amrCMrDsoRetYFg2kwwdFA, 84dVQ6n6r2ezNa
Tuc7RkKA, yW9QjWY0olv5-uRKv3t_Kw, 5XJDj7c3eoidfQ3jW18Zgw, txSc6a6pIDctvwyBeu7Aqg, HFbbDCyyqP9xPkUlcxeIdg, hTUv5oh2do6Z3OppPuuiJA, gSqonG9J4fNM-
fl_fE71AA, pd9mgTFpBTg5F9x-MsczNg, j3VE22V2GcHiH8UZxfFLfw, NYXlMW-T-3V4Jqr4r-i0Wg, btxgAZedxX8IWhMifA7Xkg, -Hp5mPLiRJNFnyeX5Ygzag, P6-DwVg6-t2J
uQwIUEk0iQ, OI2TvxYvZrAodBG_RF53Xw, bHxf_VPKmZur1Bier-6A2A, Et_Sb39cVm81_Xe9HDM8ZQ, 5HwGl2UyYbaRq8aD6YC-fA, ZK228WMcCKLo5thcjD7rdw, iTf8wojwfm0
NOi7dOiz3Nw, btYRxQYNJjpecflNHtFH0A, Kgo42FzpW_dXFgDKoewbtg, MNk_1Q_dqOY3xxHZKeO8VQ, AlwD504T9k0m5lkg3k5g6Q",
        "fans" : 5,
        "average_stars" : 4.03,
        "compliment_hot" : 2,
        "compliment_more" : 0,
        "compliment_profile" : 0,
        "compliment_cute" : 0,
        "compliment_list" : 0,
        "compliment_note" : 1,
        "compliment_plain" : 1,
        "compliment_cool" : 1,
        "compliment_funny" : 1,
        "compliment_writer" : 2,
        "compliment_photos" : 0
}
```

RAW DATA - ENTITIES



THE QUESTIONS

USE CASE

WHAT?

Discover trends in the dataset (business, ratings, reviews, visits, users). Support data-driven “human” decision making.

WHY?

Customer vs Analyst. Platform for analysts.

HOW?

Centralized datalake. Web application for search and visualization.

PROPOSED SOLUTION

Target Users

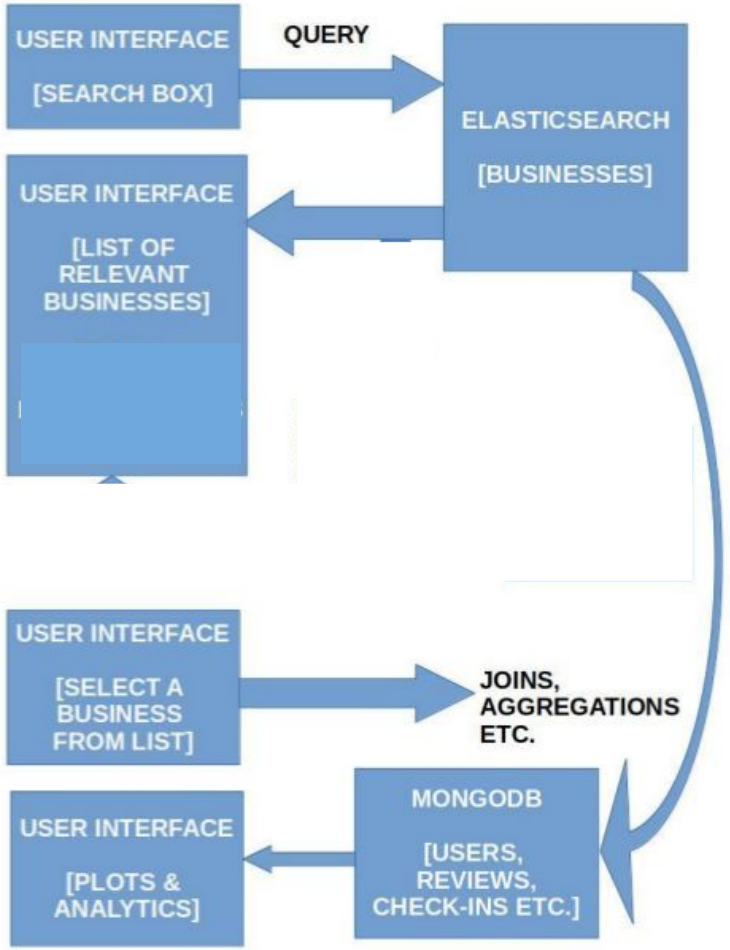
- BUSINESS ANALYSTS
- BUSINESS

OWNERS

- REAL ESTATE

CONSULTANTS

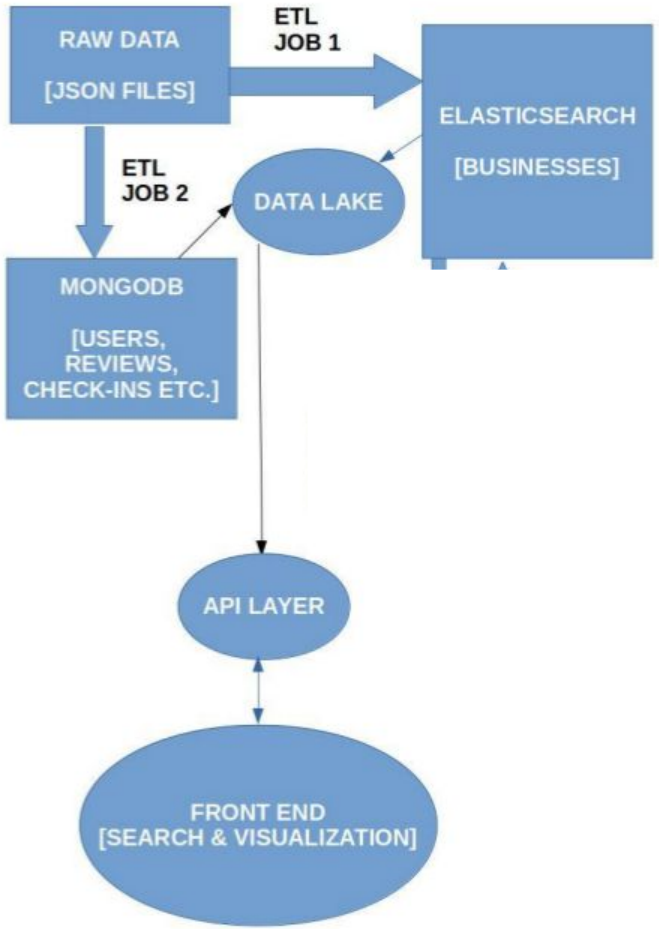
Conceptual Flowchart



MODE OF PROCESSING

- No pre-processing
- Data Lake : Elasticsearch + MongoDB
- DB Indexing
- Request Caching
- Streaming HTTP Response Body
- Filtering and Aggregation

Data Lineage



SOFTWARE STACK

BACKEND

JAVA Spring Boot, MVC Architecture

FRONTEND

Javascript, AngularJs, Highcharts

SCRIPTING / COMMANDS

Bash, Mongo shell, Python

DEPLOYMENT

Localhost, AWS EC2 & ES (experimental)

VISUAL REPRESENTATIONS

- Use of standard UI fonts, sizes, color and texture palettes.

- Elements : Search box, click buttons, dropdown menus, unordered lists, navigation buttons, canvas divisions for plots.

- Interactive plots: 3D column chart with stacking and grouping, map (US and Canada) with latitudes/longitudes, packed bubble chart, dynamic column charts, dynamic time-series line chart, histogram, fixed placement column chart, and word cloud.

- Multi-colored labels and legends.

INTERACTIVITY

- Bidirectional vertical sliders.
- Both textual and graphic representations for answered user queries.
- Mouse clicks, mouse hovering, mouse selection and their combinations.
- Menu driven plots and tabs.
- Mouse controlled zooming and panning.
- Hierarchical abstraction : Generic (trend summary) and specific (search) views.
- Built-in options for full-screen view and exporting plots.
- Different views linked through navigation using mouse clicks.

KEY INSIGHTS

- Largest number of businesses in the dataset are either restaurants or food-based.
- Religious organizations is the only category that has no business with rating 1.
- For any category, most number of businesses are rated between 2 & 4.
- Smallest number of businesses in each category are the ones that are rated the worst at 1.
- All businesses lie in one of the clusters within these states/provinces - Alberta, Quebec or Ontario in Canada, & Nevada, Arizona, Illinois, Wisconsin, Ohio, Pennsylvania, North Carolina or South Carolina in the USA.

KEY INSIGHTS

- Higher rating does not imply more visits for any business.
- Religious organizations receive the highest number of visits on Sunday, which is significantly higher than the visits received on any other days.
- Businesses receive more visits during late evening or early morning as compared to daylight hours.
- For poorly rated restaurants (1 to 2), the review sentiment is mostly negative with few positive scores in between.

KEY INSIGHTS

- For middle-rated restaurants (2 to 3, the review sentiment is mostly centered (roughly symmetrical) about zero (neutral).

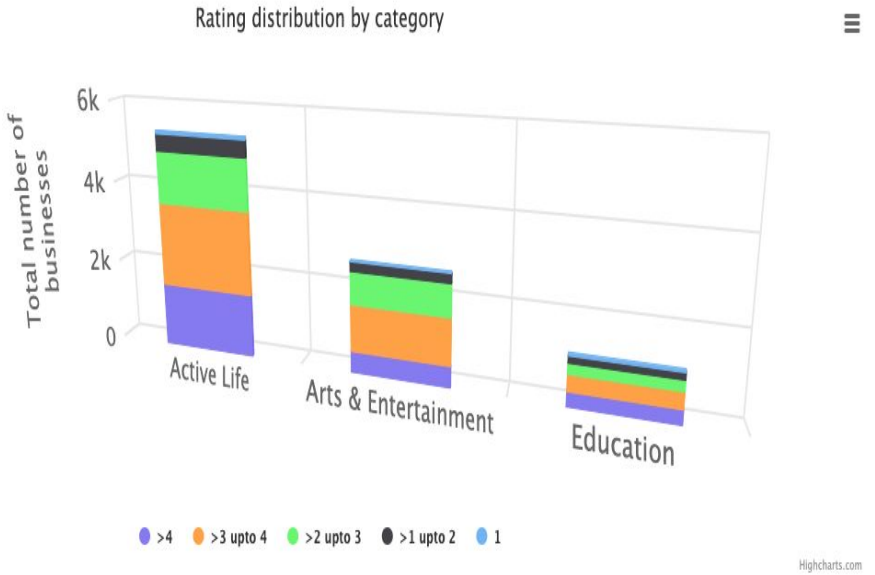
- For restaurants with good ratings (above 3), the review sentiment is mostly positive with few negative scores in between.

- For a specific business, the number negative reviews corresponds to the number of users rating the business below their average user rating.

DEMO

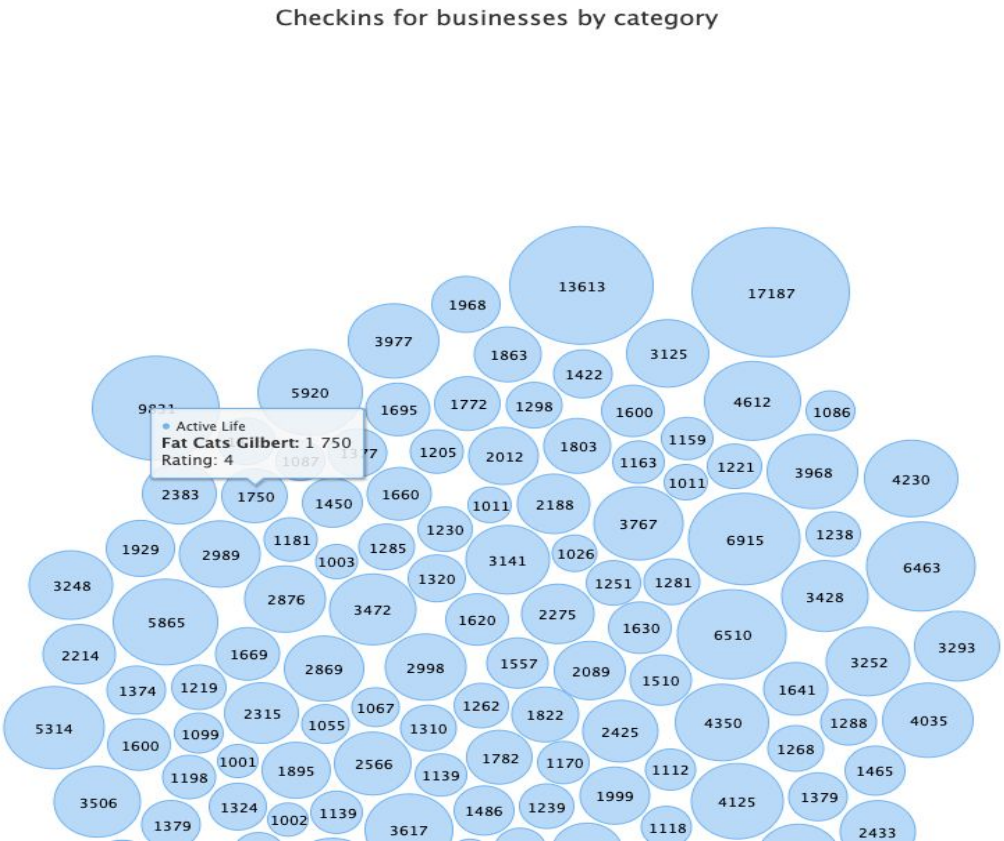
VISUAL REPRESENTATIONS, INTERACTIVITY, INSIGHTS

APPLICATION VIEWS

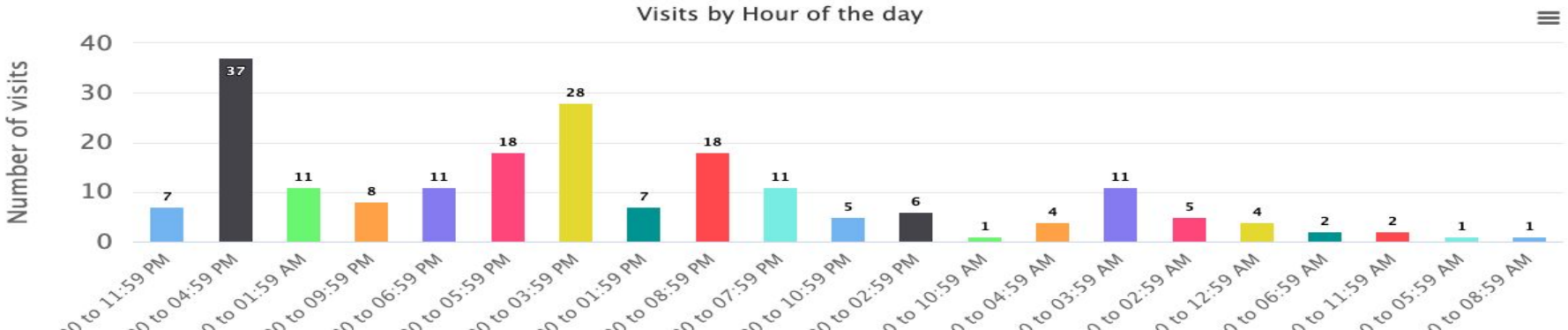


APPLICATION VIEWS

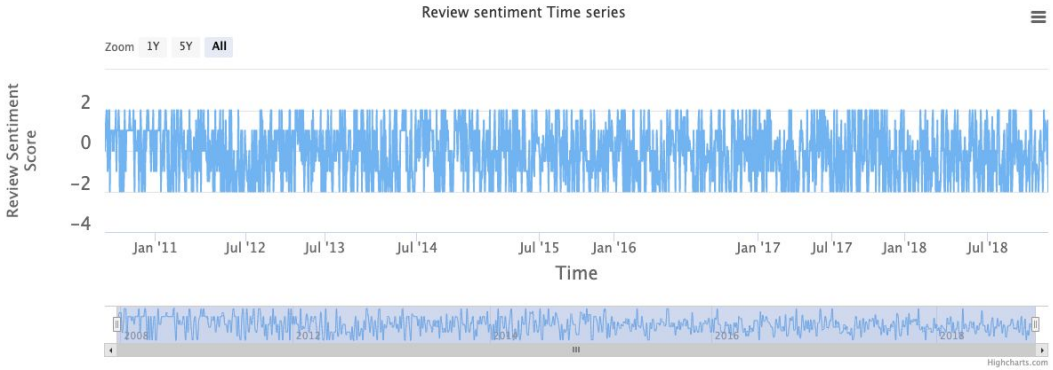
Rating Distribution | Geographic Distribution | Ratings vs Visits | Visits Distribution | Review Sentiment vs Time

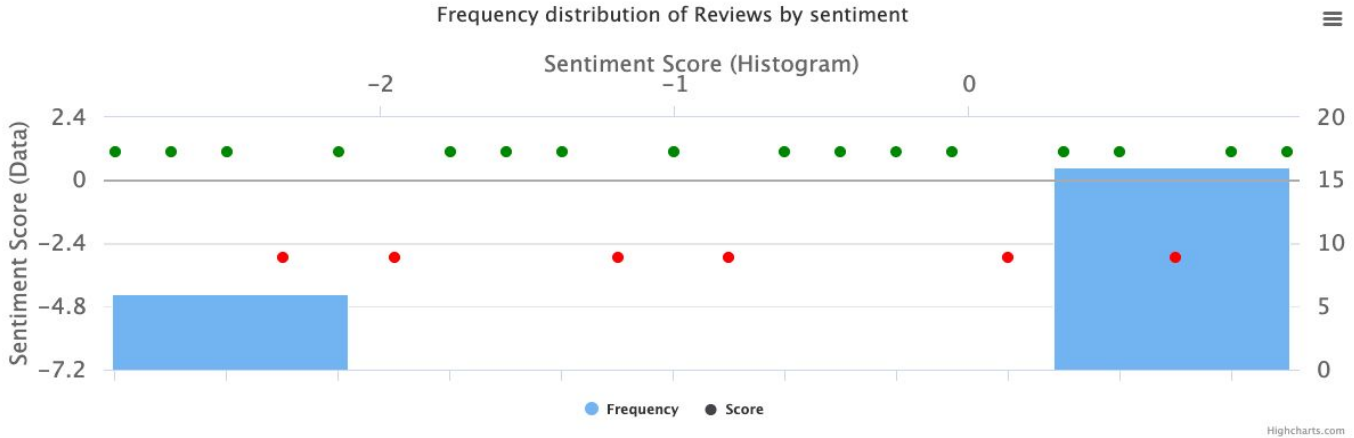


APPLICATION VIEWS

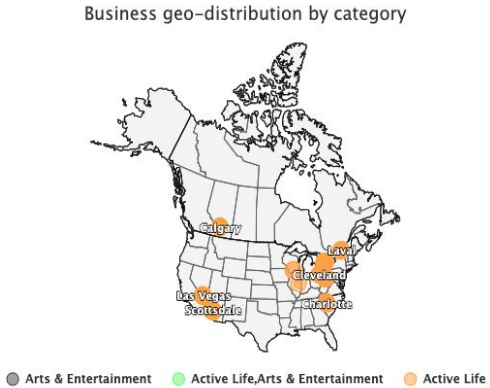


APPLICATION VIEWS

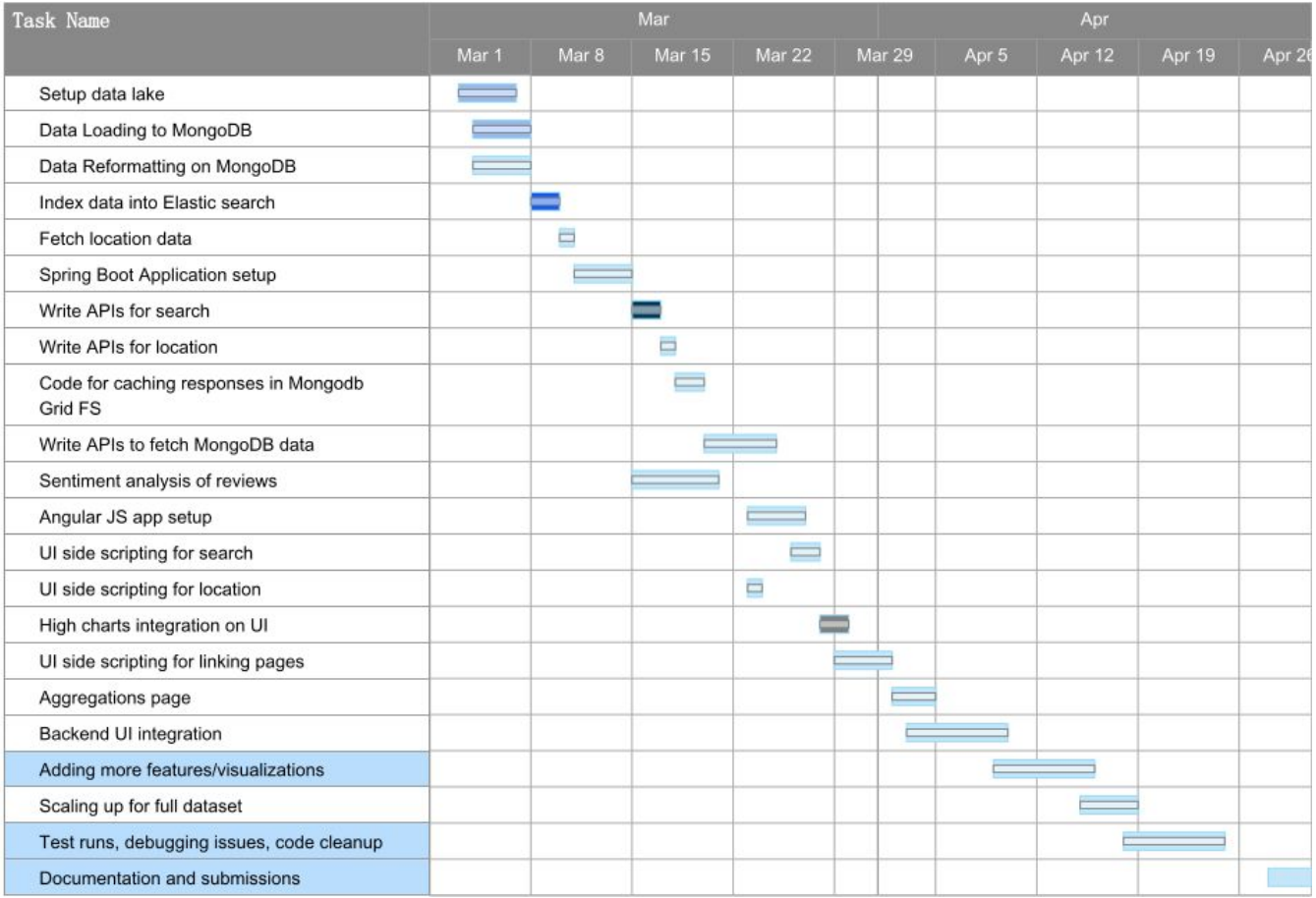




APPLICATION VIEWS



PROGRESS - GANTT CHART



FUTURE WORK

- Analyzing user-user relationships to discover insights into how one user's ratings affect other user's ratings.

- Adding user profiling and security to the application, with ability to assign different levels of visibility to each user.

- Adding location based data like population, demographics, income etc., and analyzing them against existing dimensions to discover more interesting trends.

- Make the solution generic by giving the user the ability to add data from multiple sources, so that the centralized datalake is further enriched and complex insights can be discovered.

REFERENCES

- Inspiration : http://people.ischool.berkeley.edu/~sayantan.satpati/yelp/
- Yelp Dataset Challenge : https://www.yelp.com/dataset/challenge
- Yelp Data dictionary : https://www.yelp.com/dataset/documentation/main
- Highcharts : https://www.highcharts.com/demo
- Highmaps : https://www.highcharts.com/maps/demo

THANK YOU

Q&A