CIS 450/550 : Database and Information Systems - Spring 2015
(Release date: Feb 10, 2015)

Course Project: Yelp Dataset

The goal of this project is to create an application of your own choosing over a Yelp dataset which will be provided. Developing the project will exercise: schema design and view-based access control; cloud hosting; JSON data interchange; SQL queries; potential use of a key-value database; and performance considerations. The project is designed to offer lot of flexibility in choosing features to implement, and to think about how to make these features efficient.

You are allowed to think of any application/website/game which uses the Yelp dataset, so be creative! It is your choice as to which details you want to emphasize: businesses, users, reviews, or all of them. The intent of the project is to help you understand the importance of good database design in implementing and extending your features, in the context of the Amazon Web Services platform.

You should build your application using a relational technology, and include in your design user login and credential checking. You may augment your basic design with non-relational (noSQL) technologies.

The Yelp Dataset

Get the dataset here: Yelp Dataset

The dataset consists of a single zip-compressed file, composed of one json-object per line. Every object contains a 'type' field, which tells you whether it is a business, a user, or a review.

More information on the Dataset can be found at : Yelp Dataset Challenge!

business

```
{
   'type': 'business',
   'business_id': (encrypted business id),
   'name': (business name),
   'neighborhoods': [(hood names)],
   'full_address': (localized address),
   'city': (city),
   'state': (state),
   'latitude': latitude,
   'longitude': longitude,
   'stars': (star rating, rounded to half-stars),
   'review_count': review count,
   'categories': [(localized category names)]
```

```
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

review

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

user

```
{
    'type': 'user',
    'user_id': (encrypted user id),
    'name': (first name),
    'review_count': (review count),
    'average_stars': (floating point average, like 4.31),
    'votes': {(vote type): (count)},
    'friends': [(friend user_ids)],
    'elite': [(years_elite)],
    'yelping_since': (date, formatted like '2012-03'),
    'compliments': {
        (compliment_type): (num_compliments_of_this_type),
        ...
    },
    'fans': (num_fans),
}
```

check-in

```
{
    'type': 'checkin',
    'business_id': (encrypted business id),
    'checkin_info': {
        '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
        '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
        ...
        '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
        ...
        '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)
    }, # if there was no checkin for a hour-day block it will not be in the dict
}
```

tip

```
{
    'type': 'tip',
    'text': (tip text),
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'date': (date, formatted like '2012-03-14'),
    'likes': (count),
}
```

NOTE: You can use the Yelp API to fetch more information on ‘businesses’.

Milestone 1:
Form a team (size 3 or 4), develop initial idea, and set up infrastructure [February 17th]

The initial step is to select your teammates and do the following:

1. Determine the technologies you wish to standardize on as a group. Amazon Web Services should be used for hosting your database and deploying your application.
2. Setup Subversion/Git to share source code and starter data files. See http://www.seas.upenn.edu/cets/answers/subversion.html for details, and be sure that whoever sets it up grants access to everyone in the group.
3. Upload a PDF document via Canvas, stating who is in your group, what technologies you plan to use. The document should also include a timeline for the different milestones of your project, and a preliminary division of responsibilities.

Based on this description, we will assign each group an overseeing TA who will follow your progress throughout the remainder of the semester. You should consult them early and often, and get their input as you develop an idea of the features to be implemented in your project (Milestone 2).

Each group member will receive an Amazon Web Services token, which grants $75 in usage credits. You likely won't be able to apply all of the credits to the same user account, so each of your team members should register with Amazon at http://aws.amazon.com/. Redeem your credits as you proceed with your Project, rather than redeeming everything initially.

With a total group amount of about $325, you should have enough to complete the project. However, if you exceed this amount through carelessness you will be responsible for overages. By this, we mean that you should turn off instances whenever they are not being used and NEVER publicly share your id and password or put them somewhere that they can be compromised. **We had an incident last year where hackers used the AWS Keys and deployed several EC2 instances and a bill of more than 1000$ was generated.**

Milestone 2:
Project outline and schema design [March 17th]

In this phase, you will explain your project idea in more detail, and discuss how it uses the Yelp database. Your project idea should contain the following:

- Motivation for the idea
- Features that will definitely be implemented in the application
- Features that might implemented in the application, given enough time
- Technology and tools to be used
- Member responsibility for project components

It is important to establish early on specific project component responsibilities – each group member should have aspects of the project that they "own" and are responsible for. "Own" does not necessarily mean they will be doing all of the coding / development, but rather that they are responsible for making sure the feature is complete.

You should also design a relational schema for your application, as well as a description of the noSQL component, if you choose to add one. Your schema should be based on the application rather than a straightforward copy of the dataset provided.

For this milestone you should submit a file with the information above, along with the relational schema and noSQL description.

Milestone 3:
Populate the database [March 24]

Now that you have a baseline schema to work on, the next part is to populate the database. You must extract from the Yelp dataset provided to you the data of interest, and clean it if necessary. You are allowed to add more data from other sources, but it should definitely use a good portion of the dataset provided.

You should use the AWS Getting Started handout to create your own Oracle database on Amazon RDS. For the milestone you should submit a text file with a full JDBC/SQLPLUS connect string, including guest user ID and password and database schema name, to us via Canvas. (From this we should be able to dump your SQL tables.)

Milestone 4:
Demo basic functionality [April 9th]

In this phase, you should have a running application with some basic features. Submit the source code and a brief document of the list of features through Canvas; you should also set up a time to demo what you did to your overseeing TA to get their feedback.

Final Milestone:
Project Demonstrations [April 23rd-28th]

Your final demo should contain all the basic and/or advanced features mentioned in your report along with any extra credit implemented.   You should also give the instructors an updated copy of your project description (Milestone 2) prior to starting the demo.

Extra Credit

1. Use Google Maps API to route the user from his/her current location to the closest location of business of choice(which he/she selects based on the review summary provided by the app). And also make sure that the business is open by the users predicted arrival time.
2. Trigger Bing Search, see http://datamarket.azure.com/dataset/bing/search, to return additional information.
3. Import login and user information from Facebook.
4. NoSql component (via S3, MongoDb, DynamoDB, Berkeley DB e.t.c)
5. Include friends and friend groups.
6. Anything you think is intuitive and adds interest to the application.

Experimental Validation and Report

A modern software infrastructure project isn't done until you understand how it performs, and where the bottlenecks are. Instrument your application to collect timings on various aspects. You should at least be able to determine what the latency in handling each request is, and extra credit will be awarded if you can also see what happens under multiple concurrent requests.

Your final report should include a write-up of:

1. Introduction and project goals
2. Basic architecture (not a dump of the classes)
3. Key technical challenges and how they were overcome
4. Performance evaluation
5. Potential future extensions

FINAL DELIVERABLES [May 1, 11:59 pm]

The entire project code along with the final report should be zipped and submitted on Canvas.

Yelp Dataset Challenge Awards

We encourage you to come up with an appealing project and submit to the Yelp Dataset Challenge awards. Show them how you use their data in insightful, unique, and compelling ways. You can make your submissions here before June 30, 2015: Submit to Challenge

Sample Ideas to trigger your creativity!

Visualization APP:
Create a web application from the perspective of a business owner and a user.

Business owner perspective could depict visualizations of:

- A cluster of positive and negative words from user reviews
- How the business stands when compared to it's competitors(use ratings)
- If the business has several branches, compare the relative prosperity of the business

Option to visualize any of the above based on a range of time(this week, this month, past year e.t.c.) and a radius of area.

User Perspective could depict visualizations of:

- Given a favorite business, similar businesses he/she could be interested along with a prediction of estimated interest.
- Predictions using trends of another user with similar taste, reviews and ratings

Option to visualize any of the above base of a range of time(this week, this month, past year e.t.c.) and a radius of area.

Shopping Buddy APP:
Given a shopping list, build a web application that uses the Yelp data to narrow down to which stores the user has to go to and the locations to the nearest stores.

Team Outing APP:
A web app which lets a host add his/her friends to a circle, based on their history and location predict a list of possible restaurants they might be interested it. Or if the host is planning to cook, predict a list of cuisines the guests might enjoy for the dinner party!

Plagiarism Policy

There are a lot of applications that have been developed over the Yelp Dataset. You can refer to them for ideas, but you are **STRICTLY NOT ALLOWED** to use the Codes directly.

In case you would like to use some code or snippets, please consult your mentoring TA before you do so. Please make sure that you **cite** the original author/source if you are approved to use it.

If you are caught under Plagiarism, academic measures will be taken as directed by:
http://gethelp.library.upenn.edu/PORT/documentation/plagiarism_policy.html