

## CptS - 451 Introduction to Database Systems Spring 2020 Project Description

This project is for the CptS451 students majoring in **Data Analytics**

In your semester long CptS 451 course project you will develop a data search application for Yelp.com's business review data. The emphasis will be on the database infrastructure of the application.

### Learner Objectives:

At the conclusion of this assignment you will gain experience in:

- ✓ Database modeling and design
- ✓ Populating the database with large datasets
- ✓ Querying large databases
- ✓ Optimizing query performance through indexes
- ✓ JSON parsing
- ✓ Database Application Development

### Overview:

In 2013, Yelp.com has announced the “Yelp Dataset Challenge” and invited students to use this data in an innovative way and break ground in research. In your project you would query this dataset to extract useful information for local businesses and individual users.

The Yelp data is available in JSON format. The original Yelp dataset includes and **1.2M** tips by **1.6M** users for **192.6K** businesses from United States, Canada, UK, and Germany. (<https://www.yelp.com/dataset>) In your project you will use a smaller dataset that your instructor created. This simplified dataset includes only **19,983** businesses, **189,298** users, **287,288** tips written for those businesses, and **3,786,310** check-ins to those businesses.

You will be given sample code (Python) to parse some of the Yelp JSON files (available on Blackboard). The Yelp JSON files that you will use in this project are available at the instructor's website at:

[https://eecs.wsu.edu/~arslanay/CptS451/project/yelp\\_dataset/yelp\\_CptS451\\_2020.zip](https://eecs.wsu.edu/~arslanay/CptS451/project/yelp_dataset/yelp_CptS451_2020.zip)

*(Note: Please make sure to use the dataset available on the above link, not the one from the Yelp.com website)*  
See Appendix-B for an overview of the Yelp Academic Dataset.

### Requirements:

You will develop a target application which runs queries on the Yelp data and extracts useful information. The primary users for this application will be customers/investors that would like to get information about businesses in a given location.

Using this application, the users can gather information about:

- the businesses in a given state, city, and/or zipcode,
- the businesses that belong to a certain category,
- the popular businesses, businesses with convenient service.
- etc.

You will implement this project as a standalone Python application.

A detailed description of the application and example screenshots are available in Appendix-A. In evaluating your work instructor's primary focus will be primarily on how you design your database and how efficiently you

can search the database. However your GUI should provide the basic functionality for easy search of the business. Creativity is encouraged! Additional functionality will be considered for extra credit.

You will be given more detailed milestone descriptions when they are assigned.

## Submission Instructions:

You will submit the deliverables for milestones on **Blackboard** ([learn.wsu.edu](https://learn.wsu.edu)). For each milestone you will create a .zip files that contains all deliverables for that milestone, name the .zip files as *<yourteamname>\_milestoneX.zip*, and submit it to the corresponding milestone dropbox on Blackboard. Specific submission details for each milestone will be provided under milestone descriptions.

## Project Milestones:

### I. Milestone-0: (no submission required)

Download and install PostgreSQL Database Server. You may download the latest version from the link <https://www.postgresql.org/>

### II. Milestone-1:

#### 1) Parse JSON Data:

Download the Yelp dataset from:

[https://eecs.wsu.edu/~arslanay/CptS451/project/yelp\\_dataset/yelp\\_CptS451\\_2020.zip](https://eecs.wsu.edu/~arslanay/CptS451/project/yelp_dataset/yelp_CptS451_2020.zip). Look at each JSON file and understand what information the JSON objects provide. Pay attention to the data items in JSON objects that you will need for your application. The milestone-1 description will specify which data items you shouldn't parse in the *business*, *tip*, *check-in*, *user* JSON objects.

Download the sample program from Blackboard (*Project/ Sample JSON Parsing Code*). The sample code:

- reads JSON objects form a file and extracts certain key and value pairs from JSON objects, and
- writes the extracted data into a text file.

Please note that the sample code includes examples of extracting simple key values only. In a JSON object the key value can be a composite string, and array, or another JSON object (for example: business categories, attributes, and hours), therefore you need to recursively parse those objects until you extract all data stored in JSON objects. You will write the code for parsing business, user, tip, and checkin JSON objects.

- 2) i) Design a database schema that models the database for the described application scenarios in Appendix-A and provide the ER diagram for your database design. Your database schema doesn't necessarily need to include all the data items provided in the JSON files. Your schema should be precise but yet complete. It should be designed in such a way that all queries/data retrievals on/from the database run efficiently and effectively. In Milestone2 you will revise your ER model.

ii) Translate your ER model into relations and produce DDL SQL statements for creating the corresponding tables in a relational DBMS. Note the constraints, including key constraints, referential integrity constraints, not NULL constraints, etc. needed for the relational schema to capture and enforce the semantics of your ER design.

- 3) Build a very simple database application:

Download the "Milestone1DB.csv" file from the link

[http://www.eecs.wsu.edu/~arslanay/CptS451/project/yelp\\_dataset/milestone1DB.csv](http://www.eecs.wsu.edu/~arslanay/CptS451/project/yelp_dataset/milestone1DB.csv). Create a database on

PostgreSQL with name “milestone1DB” and create a table named “business”. You will import the CSV file into this table. Detailed instructions are available in Milestone-1 specification. (Note that the schema of this table should comply with the columns of the CSV file.)

Write a simple application (either web or standalone) which connects to the milestone1DB database and runs simple queries on the business table. The goal of this exercise is to get you started in database programming early on. In Milestones 2 and 3, you will develop a larger application with all required features.

The instructor will provide a video which explains how to establish connectivity with PostgreSQL in Python using `psycopg2` driver. Instructor will provide the queries you need to run on your table (see Milestone 1 specification).

#### Milestone-1 Deliverables:

1. (25%) Source code for parsing all JSON data. Only submit your source code, not the data files.
2. (50%) The E-R diagram for your database design. To create your ER diagram, I suggest you to use draw.io (<https://www.draw.io/>) . You may also use your favorite drawing tool (e.g., Visio, Word, PowerPoint). Should be submitted in .pdf format. Name this file “<your-team-name>\_ER\_v1.pdf”
3. (25%) Source code for your application. Only submit your source code, not the data files.

Create a zip archive “<your-team-name>\_milestone1.zip” that includes your source code for JSON parsing and your sample application. Upload your milestone-1 submission on Blackboard until the deadline.

**You will demonstrate your Milestone1 to the instructor and the TA.**

### III. Milestone-2: (Deadline TBA)

- 1) Revise your database schema (ER model and relations).
- 2) Populate your database with the Yelp data. Generate INSERT statements for your tables and run those to insert data into your DB. You will also write and additional scripts to update the information stored in your database.

Details will be available in Milestone2 specification.

- 3) Build the alpha-prototype of your application.

#### Milestone-2 Deliverables:

(Weights of the deliverables are TBA)

1. The revised E-R diagram. **Should be submitted in .pdf format.** Name this file “<your-team-name>\_ER\_v2.pdf”
2. SQL script file containing all SQL statements (i.e., CREATE TABLE statements, UPDATE statements) . Name this file “<your-team-name>\_SQL.sql”
3. Source code for parsing/inserting Yelp data into the database.
4. The paper explaining your proposed metrics for classifying businesses.
5. Alpha version of your Yelp application.

Check the Milestone2 specification for additional deliverables.

Create a zip archive “<your-team-name>\_milestone2.zip” that includes your ER diagram and SQL script files. Upload your milestone-2 submission on Blackboard until the deadline.

**You will demonstrate your Milestone2 to the instructor and the TA.**

#### **IV. Milestone-3: (Deadline: TBA)**

In this milestone you will complete the implementation of your application. A detailed description of the application requirements is provided in Appendix-A.

##### Milestone-3 Deliverables:

*(Weights of the deliverables are TBA.)*

1. The source code of your application. **Please only upload your source code, not your DB files.s**

Create a zip archive “<your-team-name>\_milestone3.zip” that includes your source code. Upload your milestone-3 submission on Blackboard until the deadline.

**You will demonstrate your final project to the instructor and the TA. The demonstration schedule will be announced in mid-April.**

#### **References:**

1. Yelp Dataset Challenge, <https://www.yelp.com/dataset>
2. Samples for users of the Yelp Academic Database, <https://github.com/Yelp/dataset-examples>
3. Yelp Challenge, University of Washington Student Paper 1  
<http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p08-fants.pdf>
4. Yelp Challenge, University of Washington Student Paper 2,  
<http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p10-michelmj.pdf>

## Appendix-A

### Application Specification

The primary users for this application will be potential customers seeking for businesses. Using this application, the users can gather information about:

- the businesses in a given state, city, and/or zipcode,
- the businesses that belong to a certain category,
- the popular businesses in a zipcode, etc.
- the businesses that provide convenient service.

You should implement this project as a standalone Python application. Below you will find screenshots to help you visualize the required functionality.

### The application will have two main windows:

#### A. User Information:

User can view his/her own information, the list of his/her own friends, the latest tips that each friend has provided, and friend suggestions.

#### Use Case:

1. The user enters his name and chooses his/her own user id (among the users who has the same name). The system displays the following for the selected user:
  - user’s profile information (including, his/her *name*, *average stars*, *the date he/she joined yelp*, *number of fans*, *average stars*, and *count of votes*)).
  - user’s friends (name and star rating of each friend and the date he/she yelps since)
  - the latest review that each of those friends posted (Note that this is different than the list of most recent reviews by friends. You need to return the latest review that each friend has posted.).
  - friend suggestions. These are the users who are not in your friend list but have similar interests and behaviors. For example, they may have provided tips for same businesses or they may be the friends of your friends. In milestone-2, you will propose a metric to measure “friend similarity” and use it to come up with friend suggestions.

**User Page:**

**Login**

Username: Kelsee

Password: 6zN34Ao92JJRkogkDh-4lIra4RzqjC9shMZTJEk8ZACM-DicbEME-ga9f9VQDJXS

**My Information**

Name: Kelsee

Stars: 3.85

Yelping Since: 2009-05-06 22:00:28

Votes:

Funny: 59 Cool: 68 Useful: 221

**My Friends**

	Friend Name	Avg Stars	Fans	Tip Count	Yelping Since
1	Vanessa	4.7	4	5	2012-02-22 06:12:24
2	Jenny	4.7	2	0	2011-08-11 19:38:14
3	Michael	4.69	2	2	2011-02-15 22:25:39
4	Kristy	4.63	0	0	2011-09-11 18:30:55
5	Lindsey	4.33	551	397	2010-01-11 04:24:19
6	Matt	4.27	1	0	2011-09-02 22:52:53
7	Angelique	4.26	40	0	2010-01-10 03:35:13
8	N	4.2	33	150	2009-07-18 03:10:01

**Friend Suggestions:**

	Friend Name	Avg Stars	Fans	Tip Count	Yelping Since
1	hillery	5.0	1	2	2011-02-18 17:38:51
2	Tyler	5.0	0	1	2013-04-26 06:10:32
3	Lauren	5.0	1	0	2016-09-10 00:03:47
4	Tarah	5.0	0	2	2010-05-24 16:17:01
5	Sheila	5.0	1	1	2011-08-06 17:16:01
6	trankie	5.0	7	4	2009-11-30 06:07:32
7	Craig	5.0	0	0	2016-09-30 02:43:30

**Latest Tips of My Friends**

	Friend Name	Business	City	Date	Review
1	Elle	Sprinkles ...	Scotts...	2012-08...	How do I love thee? Let me count t...
2	Kristen	La Grande ...	Phoenix	2013-02...	\$9 for half dozen English muffins
3	Kendra	The Stand	Phoenix	2013-09...	They are out of something every ti...
4	Lickn	The Duce	Phoenix	2011-02...	Love this place.
5	Vanessa	Sakana Su...	Phoenix	2016-07...	I love the tataki on the specials bo...
6	Andi	Lindstrom ...	Phoenix	2012-11...	Hidden gem of Arcadia. \$16.95 for ...
7	Lindsey	Brat Haus	Scotts...	2015-01...	This is by far one of my favorite pati...
8	Michael	Sushi Roku	Scotts...	2013-10...	50% off today!
9	Michael	Sochu Hou...	Phoenix	2013-01...	Simply..... The best food for the rea...
10	Matth...	Butters Pa...	Scotts...	2018-08...	Love the food: New location: TOO ...
11	Tuyen	Lo-Lo's C...	Scotts...	2011-02...	Be prepared to wait...
12	Lakota	Salty Sow	Scotts...	2018-04...	Music is back !!!!!!! So so so happy !
13	N	Munich Gyr...	Founta...	2014-01...	Crunchy falafel!
14	Rob	The Apple ...	Tempe	2014-09...	Alan is the man! Friendly and helpful.
15	Christ...	YC's Mon...	Tempe	2014-06...	it's kinda dirty.
16	Jon	Restaurant...	Mesa	2016-05...	Habanero salsa, it's on fire! In a go...
17	Joseph	Defalco's I...	Scotts...	2012-07...	Real old world sandwiches. Try the ...
18	Larissa	AT Elle Nails	Phoenix	2013-03...	My favorite downtown nail salon.
19	Nydia	Amy's Bak...	Scotts...	2014-04...	wtf if you haven't eaten here, dont...
20	Amber	Bliss Rebar	Phoenix	2012-07...	Mimosas!! Only \$10 for bottle
21	Gabi	Masterpiec...	Phoenix	2017-08...	Official grand opening today!
22	Victoria	Spirit Airlines	Las Ve...	2013-02...	Here's an insider tip: if you pay for ...
23	Monica	Bliss Spa	Scotts...	2011-09...	New spa hours sun-mon 10am-7 p...

**Business Page:**

Figure 1 – User information window

**B. Business Search:**

Using this application users can search for the businesses which are within a certain state, city, zip and which belong to a certain category. In addition, the application analyzes the properties of each business in the zipcode that belongs to the selected category, and identifies the popular and convenient businesses (businesses with good service). See below for more details.

**Use Cases:**

1. User selects a state, city, and zipcode. When search button is pressed, the businesses in that state/city/zipcode are displayed (see Figure-2). The following information is provided for each business returned in the search result.

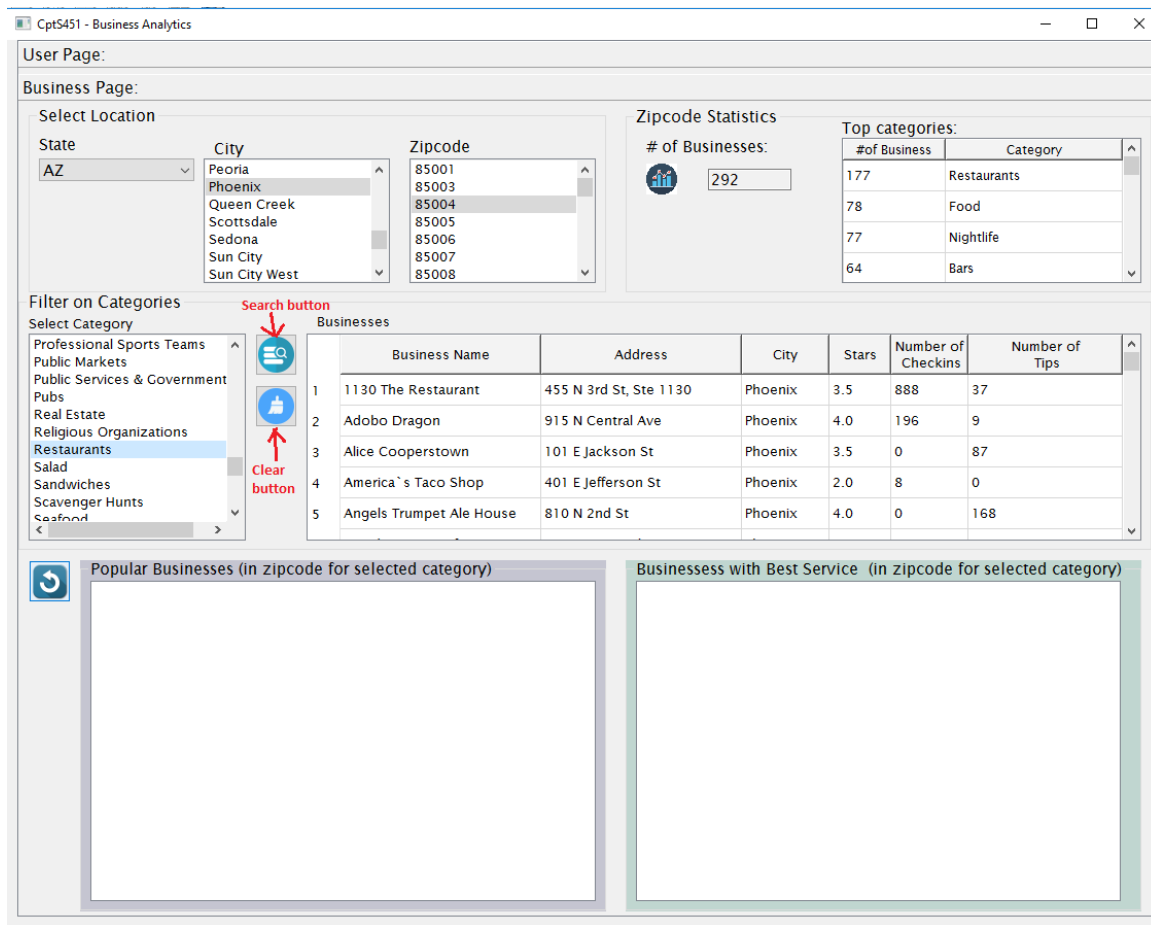
- Business name
- Address and city
- Business rating (stars)
- # of reviews provided for the business
- Average rating (stars) of the reviews provided for the business
- Total number of check-ins

(Note: You should (i) query the tip table to calculate the number of tips (ii) query the check-in table to calculate the number of check-ins for each business and (iii) update those attribute values in the business table.)

The screenshot shows a web application titled "CptS451 - Business Analytics". It has a "User Page:" header and a "Business Page:" section. Under "Select Location", there are three dropdown menus: "State" (set to "AZ"), "City" (set to "Phoenix"), and "Zipcode" (set to "85004"). To the right, "Zipcode Statistics" shows "# of Businesses:" as 292. Further right, "Top categories:" lists categories and their counts: Restaurants (177), Food (78), Nightlife (77), and Bars (64). Below these, "Filter on Categories" shows a list of categories like "Accessories", "Active Life", etc. A "Businesses" section contains two empty result boxes: "Popular Businesses (in zipcode for selected category)" and "Businesses with Best Service (in zipcode for selected category)".

**Figure 2 – Searching for the businesses in a Phoenix, AZ, 85004**

2. When the user selects a zipcode (in use case-1), the following information for the selected zipcode will be displayed (see Figure-2 above):
  - Total number of businesses in the zipcode,
  - The number of businesses per category (in the zipcode)
3. The user might refine the results by specifying a business category. The search will return the businesses which belong to the selected category (see Figure-3).



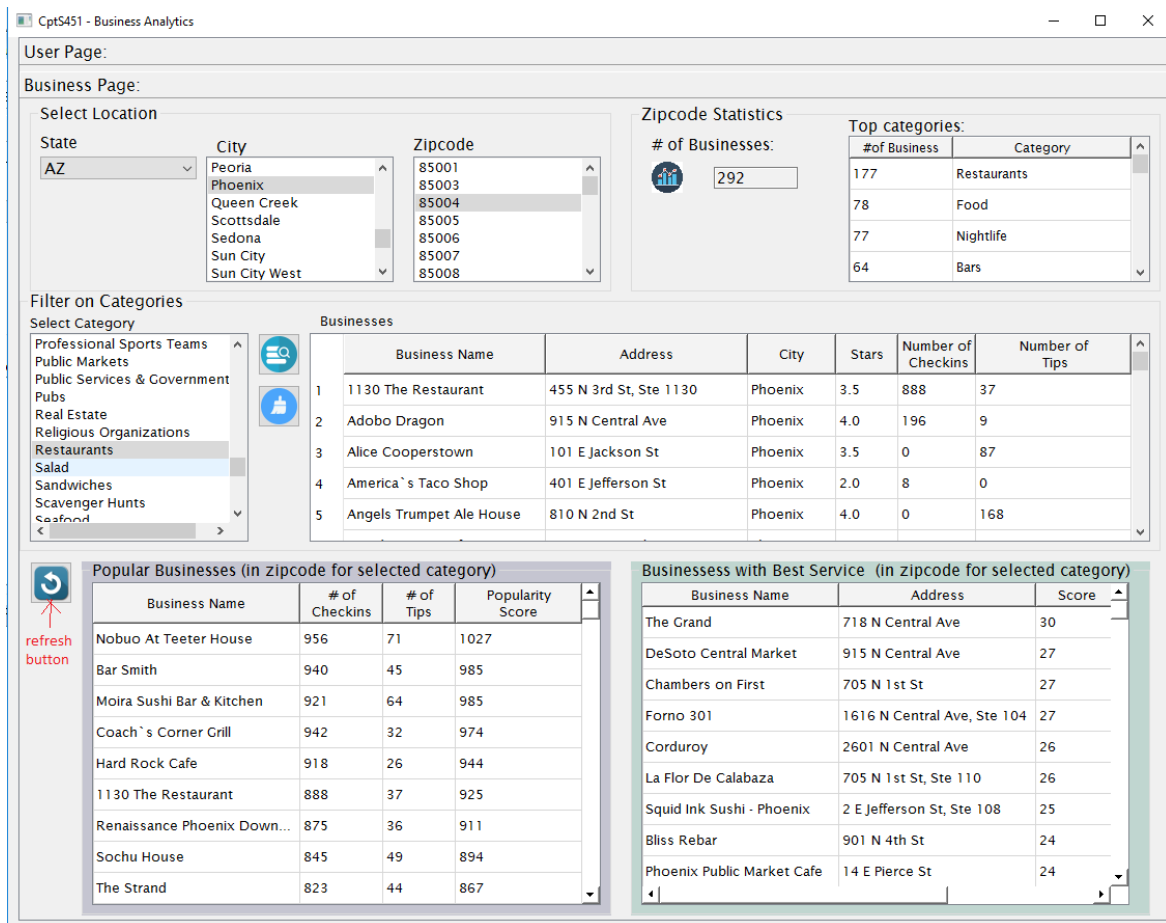
**Figure 3 – Searching for the businesses with category ‘Restaurants’ in Phoenix, AZ, 85004.**

4. Another major use case for this application is analyzing the business properties and identifying the businesses that may be classified in one of the following groups.
  - Popular businesses that seem to attract more customers and that have more positive reviews compared to other businesses in the same category and zipcode.
  - Businesses which provide more comprehensive and convenient service to their customers.

In the above description, the terms “popular”, “comprehensive”, and “convenient” are vaguely defined. You need to propose and formulate your own metrics for classifying the businesses into these two groups and specify which information you will use in your analysis. You will write a 2 page long paper where you describe your proposed metrics for friend suggestions and for identifying popular and convenient businesses. And in your application, you will implement these metrics (as SQL queries) and query the business data. Note that you may pre-analyze the Yelp data and extract some additional features/details about the businesses, store those extracted information in the database, and use them in your queries.

Figure-4 shows an example classification of businesses in the selected zipcode. The sample implementation doesn’t apply any specific metrics but classify the businesses based on the values of certain business attributes. Your results for the same zipcode will be different in your application. Also, you may display different set of attributes in your results.





**Figure 4 – When the “refresh” button is pressed, the popular and convenient businesses in the selected zipcode are displayed.**

Please note that all data should be kept in the database and should be retrieved from the database when needed. You are not allowed to create internal data structures to store the data.

## Appendix-B

### Yelp's Academic Dataset

Yelp has made available a dataset which contains user reviews for **192.6K** businesses from United States, Canada, UK, and Germany. The purpose was to provide a real-world data set to promote research in various areas of research. The dataset includes 6 types of data objects: *business*, *review*, *user*, *tip*, *check-in*, and *photos*. Every object contains a 'type' field, which tells whether it is a *business*, a *user*, or a *review*. *Business* objects contain basic information about local businesses. *Review* objects contain the details of the reviews by users for the businesses. *Review*'s `user_id` associates the reviews with the *user* objects. Similarly, *review*'s `business_id` associates each review with the *businesses*.

Detailed description of the data objects is available at:

<https://www.yelp.com/dataset/documentation/main>

In your project, you will only parse *business*, *user*, *tip*, and *check-in* objects.

Usage of this dataset is governed by the Academic Dataset Terms of Use.