CptS -451 Introduction to Database Systems
Online
Spring 2019

Project Milestone-1

Summary:

In this milestone you will parse the Yelp JSON data and develop a simple database application. The goal of this exercise is to get you started in database programming early on. In Milestone3 you will develop a larger application with all required features.

Milestone Description:

1) Download the Yelp dataset from https://www.eecs.wsu.edu/~arslanay/CptS451/project/yelp_dataset/Yelp-CptS451-2019.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.

Download the JSON Parser program (Python) from Blackboard (*ProjectResources\Milestone1\JSON Parsing Code*). This programs:

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

Please make sure that you can run this program successfully and parse all files (`yelp_business.json`, `yelp_review.json`, `yelp_user.json`, `yelp_checkin.json`) in the given Yelp data. The code assumes that the data files are in the current directory. In milestone-2, you will update this code and directly insert the data onto a PotgreSQL database.

Parsing Check-in Data: The check-in objects include information about the number of check-ins for a particular business . The "time" check-in JSON objects are in the form of:

"day": {"hour": number of checkins ,….}

For example `"Friday":{"20:00": 5,"21:00": 10}` shows that there are 5 check-ins between 20:00pm and 20:59pm and 10 check-ins between 21:00pm and 21:59pm on Friday. (time values are based on 24hour clock (i.e., military time))

2) i) Design a database schema that models the database for the described application scenario 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) (i) Download the "milestone1DB.csv" file from the link
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*". The schema of the *business* table should comply with the columns of the CSV file, i.e., there should be an attribute for each column of the CSV file. Please define the type and domain of each attribute based on the possible values that appear in the corresponding column.

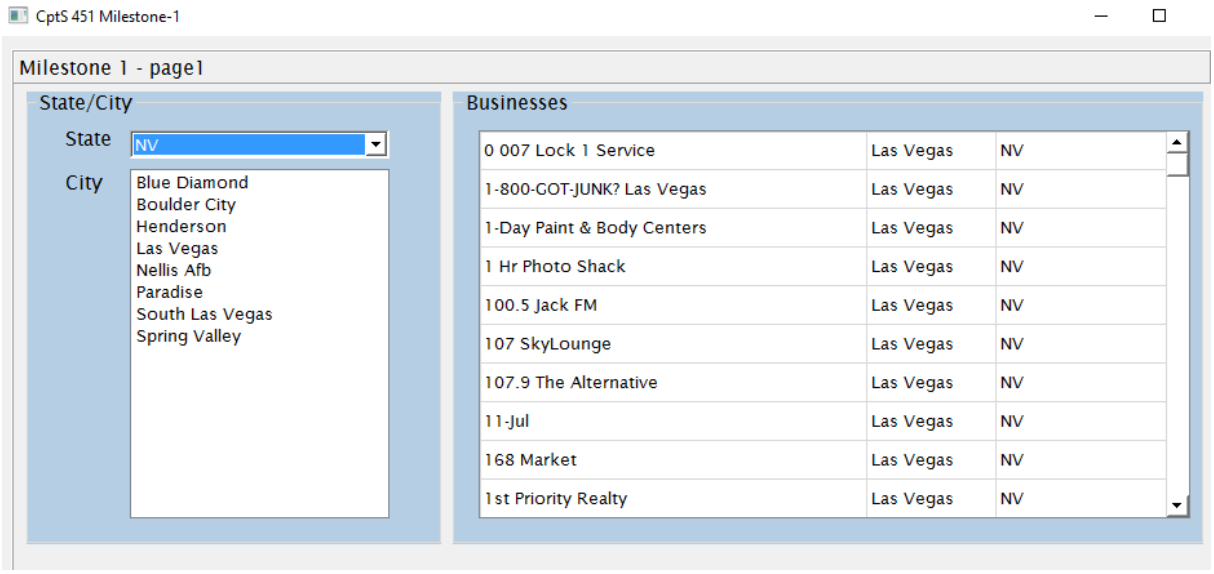
The "`milestone1db.csv`" file includes 3 columns: *name* (name of the business), *state*, and *city*.

Import the CSV file into this table by executing the following statement in the PostgreSQL command line. Please replace <path> with the directory path for the `milestone1DB.csv` file.

```
\copy business (name,state,city) FROM '<path>/milestone1DB.csv'
DELIMITER ',' CSV
```

(ii) Write a simple Python application which connects to the *Milestone1DB* database and runs simple queries on the *business* table. A sample screenshot for your milestone1 application is shown below. The application will:

- list the states that appear in business table and allow user to select a state;
- when a state is selected, the zipcodes in that state will be listed;
- when a zipcode is selected the list of the businesses will be listed.



A video tutorial on how to establish connectivity with the PostgreSQL in Python using `psycopg2` is available on Blackboard.

You need to run the following queries on the *business* table:

```
SELECT DISTINCT state
FROM business
ORDER BY state;
```

```
SELECT DISTINCT city
FROM business
WHERE state= <selected state>
ORDER BY city;

SELECT name
FROM business
WHERE city= <selected city> AND state= <selected state>
ORDER BY name;
```

The skeleton pyqt5 code is available at the link:
https://gist.github.com/arslanay/d61b1636d563e18f2680e39803f12282

Milestone-1 Deliverables:

1. (50%) The E-R diagram for your database design and the DDL SQL (i.e., CREATE TABLE) statements for creating the tables.
   a. Name the ER file "*<your-name>_ER_v1.pdf*" (Should be submitted in .pdf format.)
   To create your ER diagram, I suggest you to use Edraw Max (https://www.edrawsoft.com/download-edrawmax.php) . You may also use your favorite drawing tool (e.g., Visio, Word, PowerPoint).
   b. Name the file that includes the SQL DDL statements "*<your-name>_relations_v1.sql*"

2. (50%) Source code for your application. Only submit your source code, **not the data files**.

Create a zip archive *"<your-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.

References:

1. Yelp Dataset Challenge, http://www.yelp.com/dataset_challenge/
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