CptS -451 Introduction to Database Systems
Spring 2022

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 with designing your database schema. In Milestone2 you will revise your database schema and create it on the PostgreSQL DBMS.

Milestone Description:

Task 1: Parsing the Yelp JSON data

Clone the following GitHub repo which includes the following files:

https://github.com/WSU-CptS451-Spring2022/YelpDataset/

- `yelp_CptS451_2022.zip` : Yelp dataset you will use in your term project - this is a subset of the original Yelp dataset.
- `parseJSON_sample.py` : Sample JSON Parser program (written in Python).
- `yelpdata_sampleoutput.zip` : Sample parsed output of the given Yelp dataset.

Clone the repository and unzip the `yelp_CptS451_2022.zip` file. 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 sample JSON Parser program provides example code for:

- reading JSON objects form a file and extracting certain key and value pairs from JSON objects,
- writing 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 an array or another JSON object (for example: hours,attributes), therefore you need to recursively parse those nested objects until you extract all data stored in JSON objects. You will write the code for parsing business, tips, user, and checkin JSON objects.

- In `yelp_business.json` : Parse all keys. Flatten the nested attrite and hours values.
- In `yelp_user.json` : Parse all keys.
- In `yelp_tip.json`: Parse all keys.
- In `yelp_checkin.json` : Parse all keys.

About check-in data: Each check-in object has a "date" key whose value includes the timestamps of the check-ins to the corresponding business. All check-in timestamps are included in a long string, separated by commas. You need the split this string and extract the check-in timestamps for the business.

About attributes in business JSON objects: make sure to **recursively** parse all business attributes at all nesting levels. You should not assume a particular nesting level – some later business objects in the file may have attribute objects with deeper nesting levels than the first few business objects.

As explained above, in this milestone you will simply write the parsed data to a text file. And in milestone-2, you will revise your parse code and generate SQL INSERT statements using the parsed data. The GitHub repository includes samples of parsed data output of the given Yelp dataset - `yelpdata_sampleoutput.zip` The sample output files are provided as examples; you don't need to match the format of the given output files.

Task 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 schema should be precise but 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.

Milestone-1 Deliverables:

1. (35%) Source code for parsing all JSON data. Only submit your source code, **not the data files**. Name your file "*<your-team-name>*_parseJSON.*py*"
2. (65%) The E-R diagram and relations (CREATE TABLE statements) for your database design. To create your ER diagram, I suggest you to use draw.io tool (https://www.draw.io/). You may also use your favorite drawing tool (e.g., Visio, Word, PowerPoint). Should be submitted in .pdf format. Name the diagram "*<your-team-name>_ER_v1.pdf*" and the SQL statements "*<your-team-name>_schema.sql*".

Create a zip archive *"<your-team-name>_milestone1.zip*" that includes the below 3 files and upload it to the milestone1 dropbox on Canvas.

- *<your-team-name>*_parseJSON.*py*
- *<your-team-name>_ER_v1.pdf*"
- *<your-team-name>_ER_v1.sql*"

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