MSIT 3860

Fall 2022

Final Project Instructions

This is it!

- You have learned a lot this semester, now it's time to put it all together.
- Imagine you are a data engineer, recently hired by Yelp to support one of their data science teams.
- Your first task is to build a relational database for experimenting with new machine learning tools. The database will mirror the structure of Yelp's main production database, but with only a subset of the total data.
- In addition to setting up the database, your manager also wants you to document it in a data dictionary and ERD so the data scientists understand its structure.
- Finally, a colleague on the Partner Insights teams needs to answer some questions about Yelp's users and business partners. The rest of your team is busy, so your manager would like you to use the new database to answer those questions.
- This is a big opportunity to make an impression on your new manager! Next stop – Chief Information Officer!

What you will be doing

- Download Yelp sample data files from Canvas
- Create the database
- Write SQL commands to build the tables
- Create an ERD
- Create a data dictionary for a user with no knowledge of the data
- Make a presentation answering 1) several questions a business stakeholder might want to know, and 2) questions about your experience working with the data

What you will be handing in

- A final report presentation. Use the file “final-report-skeleton.pptx” as a template. At minimum, answer all the prompts in the template. You may add additional graphics and text.
- A complete data dictionary covering all the tables you create. Create this in Microsoft Excel, using the Northwind data dictionary as an example.
- An ERD, created in draw.io and exported to PNG.
- A SQL file containing commands to set up the database, including creation of all tables, primary keys, foreign keys, and indexes.
- A SQL file containing all queries you ran to answer the questions in the report template.

Yelp sample data files

File	Contents
businesses.csv	Main data on businesses
businessattributes.csv	Data on additional business attributes (e.g., what kind of parking they have)
businesscategories.csv	Business categories (e.g., Doctor, Restaurant, etc.)
businesshours.csv	Hours of operation for each day of the week
reviews.csv	User reviews of businesses
tips.csv	User tips on a business (e.g., suggestions)
users.csv	Main data on Yelp users
userfriends.csv	Yelp users can have friends, like Facebook. This data maps users to their friends.
usereliteyears.csv	Yelp users can have an “elite” status. This data lists the years each user was elite.

businesses.csv

Sample row of data:

```
Pns2l4eNsfO8kk83dixA6A,"Abby Rappoport, LAC, CMQ","1616 Chapala St, Ste 2",Santa Barbara,CA,93101,34.4266787,-119.7111968,5.0,7,0
```

Columns:

Column	Value	Example
0	Business id (22 characters)	`Pns2l4eNsfO8kk83dixA6A`
1	Business name	`"Abby Rappoport, LAC, CMQ"`
2	Street address	`"1616 Chapala St, Ste 2"`
3	City	`Santa Barbara`
4	State (string up to 3 characters)	`CA`
5	Postal code (string up to 9 characters)	`93101`
6	Latitude	`34.4266787`
7	Longitude	`-119.7111968`
8	Average reviews (stars), a numeric value between 0.0 and 5.0	`5.0`
9	Number of reviews	`7`
10	Is the business open (treat as an integer equal to 0 or 1)	`0`

businessattributes.csv

Sample row of data:

`Pns2l4eNsfO8kk83dixA6A,byappointmentonly,True`

Columns:

Column	Value	Example
0	Business id (22 characters)	`Pns2l4eNsfO8kk83dixA6A`
1	Attribute name	`byappointmentonly`
2	Attribute value (treat this as a string)	`True`

One business can have many attributes.

businesscategories.csv

Sample row of data:

```
Pns2l4eNsfO8kk83dixA6A,Doctors
```

Columns:

Column	Value	Example
0	Business id (22 characters)	`Pns2l4eNsfO8kk83dixA6A`
1	Category name	`Doctors`

One business can have many categories.

businesshours.csv

Sample row of data:

`mpf3x-BjTdTEA3yCZrAYPw,Tuesday,08:00:00,18:30:00`

Columns:

Column	Value	Example
0	Business id (22 characters)	`mpf3x-BjTdTEA3yCZrAYPw`
1	Day of the week	`Tuesday`
2	Opening time (ISO-formatted)	`08:00:00`
3	Closing time (ISO-formatted)	`18:30:00`

One business can be open on multiple days per week.

reviews.csv

Sample row of data:

```
saUsX_uimxRlCVr67Z4Jig,8g_iMtfSiwikVnbP2etR0A,YjUWPpI6HXG530lwP-fb2A,3.0,0,0,0,"Family diner. Had the buffet. Eclectic assortment: a large chicken leg, fried jalapeño, tamale, two rolled grape leaves, fresh melon. All good. Lots of Mexican choices there. Also has a menu with breakfast served all day long. Friendly, attentive staff. Good place for a casual relaxed meal with no expectations. Next to the Clarion Hotel.",2014-02-05 20:30:30
```

Columns:

Column	Value	Example
0	Review id (22 characters)	saUsX_uimxRlCVr67Z4Jig
1	User id (22 characters)	8g_iMtfSiwikVnbP2etR0A
2	Business id (22 characters)	YjUWPpI6HXG530lwP-fb2A
3	User rating (stars, a numeric value between 0.0 and 5.0)	3.0
4	Number of users marking the review as useful	0
5	Number of users marking the review as funny	0
6	Number of users marking the review as cool	0
7	The review text	Family diner. Had the buffet…
8	Review date/time	2014-02-05 20:30:30

One business can have many reviews.
One user can write many reviews.

tips.csv

Sample row of data:

```
AGNUgVwnZUey3gcPCJ76iw,3uLgwr0qeCNMjKenHJwPGQ,Avengers time with the ladies.,2012-05-18 02:17:21,0
```

Columns:

Column	Value	Example
0	User id (22 characters)	AGNUgVwnZUey3gcPCJ76iw
1	Business id (22 characters)	3uLgwr0qeCNMjKenHJwPGQ
2	Text	Avengers time with the ladies.
3	Date/time the tip was left	2012-05-18 02:17:21
4	Number of compliments the tip received	0

One business can have many tips.
One user can leave many tips.

users.csv

Sample row of data:

```
qVc8ODYU5SZjKXVBgXdI7w,Walker,585,2007-01-25
16:47:26,7217,1259,5994,267,3.91,250,65,55,56,18,232,844,467,467,239,180
```

Columns:

Column	Value	Example
0	User id (22 characters)	qVc8ODYU5SZjKXVBgXdI7w
1	Name	Walker
2	Number of reviews the user has left	585
3	Date/time the user joined Yelp	2007-01-25 16:47:26
4	Number of useful votes sent by the user	7217
5	Number of funny votes sent by the user	1259
6	Number of cool votes sent by the user	5994
7	Number of fans	267
8	Average rating of all the user’s reviews (numeric between 0.00 and 5.00)	3.91
9	Number of hot compliments received by the user	250
10	Number of more compliments received by the user	65

users.csv (continued)

Column	Value	Example
11	Number of profile compliments received by the user	55
12	Number of cute compliments received by the user	56
13	Number of list compliments received by the user	18
14	Number of note compliments received by the user	232
15	Number of plain compliments received by the user	844
16	Number of cool compliments received by the user	467
17	Number of funny compliments received by the user	467
18	Number of writer compliments received by the user	239
19	Number of photo compliments received by the user	180

userfriends.csv

Sample row of data:

```
qVc8ODYU5SZjKXVBgXdI7w,NSCy54eWehBJyZdG2iE84w
```

Columns:

Column	Value	Example
0	User id (22 characters)	qVc8ODYU5SZjKXVBgXdI7w
1	Friend id (22 characters)	NSCy54eWehBJyZdG2iE84w

One user can have many friends.

usereliteyears.csv

Sample row of data:

```
qVc8ODYU5SZjKXVBgXdI7w,2007
```

Columns:

Column	Value	Example
0	User id (22 characters)	`qVc8ODYU5SZjKXVBgXdI7w`
1	Year	`2007`

One user can have many elite years.

How to approach this project

- Start early! Do not wait until a week before the project is due (20 December) to get going. I am here to answer questions, but my time is limited. So again, start early.
- The database name, table names, variable names, and data types are up to you. This slide deck has all the information you need to make decisions on things like data types.
- I will be grading you on database and query optimization. Optimize your table structure by choosing the most efficient data types possible (i.e., don't use a TEXT when a VARCHAR will do).
- Pay attention to details. Add comments to your SQL files to explain what you did, make sure your data dictionary is easy to read, and put together a professional-looking final presentation.
- Don't cheat yourself. This is a real sample data set provided by Yelp for instructional and data science purposed. You can probably find a lot of what you'll need for this project online. I will catch obvious plagiarism (trust me), but you can probably slip some stuff by me**. THIS DOES NOT DO YOU ANY GOOD.** This project is a check for you as well – are you ready to move on to other courses? Are you ready for future job interviews? By copying stuff off the Internet you are robbing yourself of an important opportunity to test yourself.

Example of how to think about table design

- Look at slide 9, which explains what's in the businesshours.csv file.
- You know you'll need a table to hold this data. What do you want to call it?
- The slide describes the data that will be in the file. Each line in a file holds a single row of data, separated by commas. Text data that contains commas will be enclosed in double-quotation marks.
- Look at slide 9. What attributes will your table need? What is the most efficient (smallest) data type each attribute could have?
- Slide 9 also tells you that "One business can be open on multiple days per week." What does this imply about the relationship between businesses and businesshours?
- Do you need a primary key on this table? How about indexes to speed up data operations? Do you want to set up any foreign key relationships to protect data integrity?