601.315 Databases, Spring 2022
Project Phase C: Setup and Cleanup

Due: Tue, 12 Apr at 11pm. Use of late days *is NOT* permitted.

Part 1: Identifying and Cleaning Up Data. Process the data you have identified for your database (that is, put it in text files in the format you want, clean it up as needed, etc.) so that it can be uploaded to your database via a setup script. You should avoid tedious data cleaning "by hand" and instead opt to write a script to automate the work. Briefly document the this process (which data came from which URL and in what format, what downloaded columns you needed to remove, which downloaded files contained duplicate values or anomalies and how you removed them, what splitting of files was needed, etc.) in a plain text file named *process.txt* so that your steps could be replicated by someone else if needed. At the end of this process, complete data for each one of your relations should exist in a plain text file named *relation-name.text* (e.g., the text file named *course.txt* would contain the data to be uploaded into a relation named *course*, the text file named *faculty.txt* would contain the data to be uploaded into a relation named *faculty*, etc.).

Part 2: Creating a Test Data Set. Next, make copies of each of your data text files which are significantly smaller versions with, say, 8-15 tuples of data each. To name these files, append *-small* to each text filename prior to the *.txt* extension. For example, if *course.txt* contains all data for the full *course* relation, then *course-small.txt* will contain the pared-down version which contains only 8-15 tuples for the relation. These will be files you use to make a small database for testing purposes. Soon, you'll write queries in MySQL that allow you to answer questions including those you posted in Phase A, and you'll test them on your small database populated with these files before using them on your full database. Don't simply use the first 8-15 lines in each of the full data files here; give some thought to interplay between tuples that exist in different relations, and make your small data set *adversarial*, or intentionally tricky to handle. Even in this small version of your database, you'll want the queries you test to yield interesting results. A good check on the quality of your test data is to determine the output that you'd expect to see if you asked 6-8 of the most interesting questions you posed in Phase A. You might need to fabricate some test data here so that you can meaningfully test in the next phase.

Part 3: Setup Script for Small Database. Now, write a MySQL script named *setup-small.sql* that, when run on dbase.cs.jhu.edu, will create your database. Your script should create each of your relations with appropriate primary key, foreign key, and other constraints as needed, then populate your database with the *-small* versions of your data. Test it out on the dbase server, and make modifications as needed so that no error messages are reported as the script runs.

Part 4: Cleanup Script. Now, write a MySQL script named *cleanup.sql* that executes commands which completely delete all relations (not just tuples) in your database. After your deconstruction script executes, you should be able to re-run one of your setup scripts without any errors due to, for example, the attempted creation of a duplicate table. The same cleanup script should be effective in wiping out a database created by either *setup-small.sql* or *setup.sql*, described in the following part.

Part 5: Full Database Setup Script. Finally, write a MySQL script named *setup.sql* that, when run on dbase.cs.jhu.edu, will create a database with the same table structure as described by *setup-small.sql*, but populates the tables using your full data files. This should require only minimal tweaks to the script you created in Part 3 above. Once again, test your script out on the dbase server, and make modifications as needed so that no error messages are reported as the script runs.

Part 6: README File. Lastly, create a plain text file named *README*, which contains partner names and JHEDs and paragraphs as needed describing any issues you encountered during Phase C, or specific or general concerns you have about your project at this point.

Deliverables. Make sure that each of the following files is included, and named as specified.

- *README*
- *process.txt*
- *setup-small.sql*
- *cleanup.sql*
- *setup.sql*
- A set of shortened data files ending with *-small.txt*, one for each relation in the finalized schema you laid out in Phase B. The set must include all files needed for *setup-small.sql* to run to completion without errors.
- A separate set of full data files ending with *.txt*, one for each relation in the finalized schema you laid out in Phase B. The set must include all files needed for *setup.sql* to run to completion without errors.

The top of your `README` and *process.txt* plain text files must list the names and JHED IDs of each partner. Likewise, the first several lines of the script (*.sql*) files should be comment lines specifying each partner's name and JHED.

Submit your work via Gradescope by the deadline listed above for Phase C. One partner will submit the work as a team submission upload, and will indicate all partner names. Therefore, only one partner should submit.

Looking Ahead: In Phase D, you'll need to hand in thoroughly-tested SQL DML statements representing the queries you need to write to answer the questions you posed in English during Phase A. Once you complete Phase C, feel free to get started crafting those queries and testing them.