

Lab 2: Database Creation

Due dates:

CREATE TABLE statements:

MWF section: Thursday, September 28, 11:59 p.m.

TR section: Friday, September 29, 11:59 p.m.

Full lab (hard deadline):

MWF section: Monday, October 2, 2017, 8:00am

TR section: Tuesday, October 3, 2017, 9:30am

Lab Assignment

Assignment Preparation

This is an individual lab. Each student will complete all work required in the lab, and submit all required materials **exactly as specified** in this assignment.

For this assignment, you will be using your MySQL account.

You will be using the MySQL server hosted at `cs1vm74.csc.calpoly.edu` and the MySQL client available on the CSL machines.

For the purposes of this and future assignments, a **SQL script** is a text file that contains a sequence of SQL statements and MySQL commands, as well as SQL comments. Typically, these files should receive a `.sql` extension. You will be required to prepare and submit a number of SQL scripts for this lab.

The Datasets

Starting with this lab, and for most of the remaining labs in the course you will be working with several datasets created by Dr. Alex Dekhtyar specifically for this course. These datasets are not large, but they are sufficiently diverse in content, scope and structure.

All datasets are available from the course web page:

<http://users.csc.calpoly.edu/~ebuckale/365-Fall2017>

Each dataset comes with a README file, which contains the exact specifications of all data files included in the dataset, and briefly explains the meaning of the dataset. Before starting your work on a dataset, **please study carefully the README file and make sure you understand the structure of the dataset!** All data files in each dataset are stored in the CSV (comma-separated values) format. All text values are enclosed in single quotes. A .zip file for each dataset is available.

Brief descriptions of each of the course datasets are given below.

CSU dataset. Type: multidimensional statistical data. This dataset contains various statistics about the California State University system. Historic information, such as annual enrollments and graduations is included, as well as information about enrollments by discipline at all campuses in a single year, and information about faculty lines¹ and campus fees.

CARS dataset. Type: normalized² dataset. This database stores information about the properties (such as number of cylinders, mileage per gallon, engine displacement, etc.) for over 400 domestic and import cars produced between 1970 and 1982. Information is split into several files: starting with lists of continents and countries, going on to the lists of automakers and the models and makes they were producing.

BAKERY dataset. Type: OLTP (on-line transaction processing) dataset. This dataset records information about one month of sales from a small bakery to a list of its dedicated customers. The dataset captures the notions of a transaction (a single purchase) and market baskets (each purchase may contain more than one item).

MARATHON dataset. Type: universal table. This dataset consists of a single CSV file documenting the performance of participants of a half-marathon race. The performance is tracked for the entire race, as well as within each gender/age category.

STUDENTS dataset. Type: simple normalized. This is a variation of the dataset you have encountered (different names, no buses, no GPA) in Lab 1. The dataset consists of a list of students assigned to grades and classrooms, and a separate list of teachers assigned to classrooms. This is the simplest dataset and can be used as “staging grounds” for most of the activities in this and future labs.

¹This is information about the number of Full Time faculty equivalents at each campus, including part-time faculty added up, e.g. two half-time faculty would count as one.

²In CSC 366 you will get an opportunity to study the formal meaning of the term “normalized” when applied to databases. An informal explanation: a “normalized” database is a database where information is split into a large number of tables, reducing redundancy in data.

AIRLINES dataset. Type: graph. The dataset stores information about a number of airlines, and the flights these airlines have between 100 different airports. It can be viewed as a multicolored graph with 100 nodes representing airports, edges representing direct flights and edge colors representing the airlines running the flights.

WINE dataset. Type: normalized (somewhat). The dataset lists ratings of a variety of single-grape California wines of different vintages as given by the Wine Spectator magazine. The dataset consists of a list of wines complete with their ratings on a 100 point scale, their reported sales prices and the production volumes. Two additional lists: appellation/American Viticultural Areas (AVAs) and grape varieties are available as well.

INN dataset. Type: simple OLTP. The dataset contains information about one year's (2010) worth of completed hotel reservations for a small Bed & Breakfast inn. The inn contains 10 uniquely named rooms - each with its own set of features. The dataset lists all reservations that *commenced* 2010³ and specifies for each room reservation the checkin and checkout dates, the name of the person making the reservation and the number of adults and kids staying in the room.

KATZENJAMMER dataset. Type: normalized⁴. The dataset follows the recording and performing career of an all-female Norwegian band Katzenjammer⁵. The four members of the band, Sloveig, Marianne, Anne-Marit and Turid turn their live shows into a cascade of instrument changes. The dataset documents the albums the band released, the songs on the albums, and the typical ways in which the songs are performed live: the instruments each band member plays, the vocals, and their position on the stage.

Examples of the band's performances can be found in abundance on Youtube.

Your Lab 2 assignment should be performed for each of the datasets in the list above. We separate the datasets into two tiers:

1. Tier 1. Datasets that result in tables with only `INTEGER`, `FLOAT` (or `DOUBLE`) and `VARCHAR` attributes. These datasets are:

- STUDENTS
- CSU
- CARS
- AIRLINES
- WINE

³Some reservations made at the end of the year end in 2011.

⁴In the course we use the Version 2.0 of this dataset, created in September of 2016.

⁵<http://www.katzenjammer.com>

- KATZENJAMMER
2. Tier 2. Datasets in this tier have at least one table with DATE attributes. Handling dates and times in MySQL is a topic of conversation for one of the later classes this week. Because of it, you will first work on the Tier 1 datasets. Tier 2 datasets are:
- BAKERY
 - MARATHON
 - INN

The Task

For this lab you will design a relational database for each of the datasets, and will instantiate it with the data provided to you.

Database design and creation.

All tasks outlined in this section must be performed on all datasets.

You need to do the following:

1. For each dataset, create a relational database to store its data. The following rules must be followed:
 - (a) The tables of the database must match the files of the dataset one for one.
 - (b) You are allowed to choose any (hopefully meaningful and non-offensive) names for all relational tables and columns in them.
 - (c) You must properly detect and declare all constraints, including primary key, candidate key (SQL's UNIQUE), and referential integrity/foreign key constraints.
2. Write and test a SQL script for creation of each database (one script per database). These scripts *shall only contain* the CREATE TABLE statements.
3. Write and test a SQL script for deleting all tables from each database (one script per database).
4. Prepare and test SQL scripts for populating each database with the data available from the .csv files. (Hint: use any available programming/scripting language to convert the .csv file into a list of SQL statements for adding records to the database. There is a relatively simple way to solve this problem (almost) forever with a single Python script, for example). You should have one SQL script per database table. Each script you produce *must contain only* INSERT statements. For this lab you must use *one INSERT statement per tuple*. (This is

to make sure that if you have an error in the script, you can debug it efficiently).

5. **Note:** While your table/column names can be any of your choice, the names of all scripts you must prepare and submit are defined in this assignment. Please refer to the **Submission Instructions** section for the specs for proper script naming. **Be aware that incorrect names for submitted files, EVEN incorrect capitalization will lead to significant deductions in your lab score!**
6. Write and test the scripts for checking the contents of each database. A simple statement to check the contents of a database table is

```
SELECT * FROM <Table>;
```

This statement will result in all records in a table printed in their entirety. The query

```
SELECT COUNT(*) FROM <Table>;
```

will output the number of tuples in the table.

WINE dataset: Special Notes

To make your life a bit more interesting, when creating the database for the WINE dataset, please, take into account the following information:

1. The list of appellations/AVAs contained in the `appellations.csv` file, lists a state for each appellation/AVA entry⁶. Similarly, there is a “State” column in the file `wine.csv`, describing the state of the origin of the wine. We note that `wine.csv` does record the appellation/AVA for each wine (the “Appellation” column), and that the state of origin of the wine is essentially the state of the wine’s appellation. Therefore, there is a duplication of this information in the CSV files in the dataset.
2. The last column of the `wine.csv` file, “Drink” is designed to contain the advice of the wine raters on when the wine is best consumed. Most of the entries contain a text value ‘now’. Some entries contain a number representing the year when the wine will start being at its best.

For your WINE database you should be as follows:

- The state information is included as part of the appellation/AVA information.
- State information IS NOT included with the wine score (including appellation information will suffice).

⁶In the current dataset all appellations/AVAs are from California, but state information still shall be recorded for each appellation/AVA.

- the drink advice column IS NOT included with the wine score information.

Note 1: There are a number of other interesting features in this dataset regarding the primary/foreign key identification. Be careful!

Note 2: There is some missing data from this dataset - some wines do not have production volume information. You have to deal with this properly.

Note 3: In winemaking, an “appellation” is essentially a region from which the grapes used in making the wine originate. The most broad appellation in our dataset is the entire state of California. According to US law, county names are legal/valid descriptions of grape origins. US recognizes other regions as legal/valid specifications of grape origins. Such regions are called AVAs, a.k.a., American Viticultural Areas. In California, AVAs can cross county borders and also be embedded into one another. For example, “Russian River Valley” AVA is located inside “Sonoma Valley” AVA, while “Lodi” AVA spans multiple counties. For simplicity, in the latter cases, `appellations.csv` file lists the main county of an AVA. Some AVAs are larger than a county. E.g., “Central Coast” and “North Coast” AVAs span multiple counties. For such AVAs, `appellations.csv` lists the value ‘N/A’ in the “County” column.

The file `appellations.csv` lists all grape origin specifications that can appear on wine labels. The last column specifies whether the origin specification is an AVA or not.

Extra Credit

For extra credit do the following:

- Find a data collection on the web. The data you will be using must come from a source that either explicitly allows its use for non-commercial purposes, or the nature of the data is such that such use is permissible. Examples of such sources include (but are not limited to) government statistical data (you, the taxpayer, have paid for this data to be collected), any public domain data collections, any scientific data that was published, or any data you assemble yourself from various sources.
- Create a database schema, extract the data that matches the schema, and insert the data into the database.
- Submit SQL scripts for database creation, database population, and removal of all tables of the database (one per table). Additionally, submit a `README` file describing the nature of the database and the meanings (if necessary) of the table columns.

Note. Because some students in the past chose this part of the assignment as the means of trolling the instructor, extra credit **will not be given** for

any submissions that deliberately contain curse words or other bad language in the names of the dataset, tables, attributes, or table contents. The instructor will be the final arbiter of what constitutes offensive material.

Submission Instructions

General Instructions

Please, follow these instructions exactly. Up to 25% of the Lab 2 grade will be assigned for conformance to the assignment specifications, **including the submission instructions.**

Please, **name your files exactly as requested** (including capitalization and any typos present in instructor's filenames), and submit all files **in a single archive**. Correct submission simplifies grading, and ensures its correctness.

Please include your name and Cal Poly email address in all files you are submitting. If you are submitting code/scripts, include, at the beginning of the file, a few comment lines with this information. Files that cannot be authenticated by observing their content will result in penalties assessed for your work.

Specific Instructions

You must submit all your files in a **single** archive. Accepted formats are **gzipped tar** (`.tar.gz`) or **zip** (`.zip`). The file you are creating for your final submission must be named `lab2.<extension>`, i.e., either `lab2.tar.gz` or `lab2.zip`.

Inside it, the archive shall contain nine directories named **AIRLINES**, **CSU**, **CARS**, **BAKERY**, **KATZENJAMMER**, **MARATHON**, **STUDENTS**, **WINE** and **INN** (same as the dataset names). In addition, the root of the directory must contain a **README** file, which should, at a minimum, contain your name, Cal Poly email, and any specific comments concerning your submission.

INCORRECT SUBMISSION. A submission that unpacks a single directory (e.g., `lab02/`) which, in turn, contains the nine directories described above in it is considered to be **INCORRECT**. Also, please make sure the file name capitalization and spelling for all files is correct!

If you are submitting extra credit assignments, put the directory/directories for your dataset(s) in ALLCAPS here as well. Follow the same naming conventions for your extra credit SQL scripts as described below.

Each directory shall contain all SQL scripts built by you for the specific dataset. The scripts shall be named as follows.

Filenaming conventions:

- database creation scripts. The filename shall be `<dataset>-setup.sql`. Here, `<dataset>` is the name of the directory in ALL CAPS. E.g., for

the CARS dataset, the filename will be `CARS-setup.sql`.

- table deletion scripts. The filename shall be `<dataset>-cleanup.sql`.
E.g., for the CARS dataset, the filename will be `CARS-cleanup.sql`.
- table population scripts. The filename shall be `<dataset>-build-<table>.sql`.
Here `<table>` is the name of the `.csv` file (**not** the name of the table that you are building).

For example, for the table populated with the list of car makers in the CARS database, the filename will be `CARS-build-car-makers.sql`.
Use **the same capitalization** as the filename in the dataset: e.g., `CSU-build-Campuses.sql` and `INN-build-Rooms.sql`, `CARS-build-cars-data.sql` and `BAKERY-build-items.sql`.

- testing scripts. These are scripts described in item 6 of Part 1 assignment. The filename shall be `<dataset>-test.sql`. E.g., `CARS-test.sql`

Submission of CREATE TABLE statements

To simplify and speed-up grading of the lab, you will submit your `CREATE TABLE` statements ahead of time (as well as at the end with the full assignment). The `CREATE TABLE` statements submission shall be done as follows.

- You will submit nine files: the `<DATASET>-setup.sql` files described above.
- Place these files into a separate directory, change into this directory, and run the `zip` or `tar` and `gzip` commands.
- Name the submitted file `lab02-setup.zip` or `lab02-setup.tar.gz`.
- Make sure your name is placed in a comment at the top of each file you are submitting.
- Submit using the appropriate following handin command:

```
$ handin ebuckale mwflab02-setup <file>
$ handin ebuckale trlab02-setup <file>
```

Submission of the main assignment

Submit the full Lab 2 assignment using the following `handin` command:

```
$ handin ebuckale mwflab02 <file>
$ handin ebuckale trlab02 <file>
```

Note: `<DATASET>-setup.sql` files **must be placed** in the appropriate directories in your full Lab 2 submission.

Testing

Your submission will be tested by running all scripts you supply and checking the produced output for correctness. I may also use some extra scripts to verify the correctness of the databases you have constructed. Thus you see the importance of adhering to the naming conventions for all files. As you all know, a computer script is not forgiving of files names that differ from expectations.

A test script will be released to you in the middle of the week. **Please**, do not submit your work without first running the test script on it.

If you are aware of any bugs or incorrect behavior of your SQL scripts, I strongly suggest that you mention it in the README file.