CSC2141 Project Part 1

Finding Data and Planning Your Database

Due date: Monday, February 5, 11:59PM

Overview

In the first part of the course project, you will identify a dataset that you will use in later parts of the project. Part 1 is scored out of 20 and worth 5% of your final grade. Please refer to the project overview document for general rules and guidelines.

The components of Part 1 are as follows:

- **Introduce** the dataset (5 points).
- **Generate and describe** your dataset (10 points).
- **Create a conceptual model** and **an internal model,** marking possible keys, constraints, and data types (5 points).

Submitting Project Part 1

Please provide your submission as a formatted PDF with the filename '**B########_Project_Part1.pdf**', where 'B########' is your Banner ID.

Upload through the Brightspace portal.

Please consult the section on academic integrity in the project description document.

Important: You will not submit any data or SQL code for Part 1!

Section 1: Introduce the Dataset

In this section you will provide an overview of the data you will be working with. Specifically, you should address the following questions:

1a) What kind of data will you be working with?

1b) Why are these data interesting?

1c) What kinds of information would you like to generate from the database you will build? You can provide a couple of example questions here.

Expectations are roughly **one paragraph each** for a, b, and c.

Section 2: Generate and Describe Your Dataset

You must identify a dataset that has at least four distinct entities (in the conceptual sense) that can be related to one another such as:

- EMPLOYEE, PRODUCT, INVOICE, CUSTOMER
- MOVIE, ACTOR, COUNTRY, GENRE
- ANIMAL, HABITAT, CONTINENT, CLASS

The dataset should allow you to ask questions such as "Which employee has sold the most products between January 1 and December 31, 2019?" or "Which continent has the highest diversity of threatened species?"

Rules for building the dataset

- In building the dataset, you must have at least **one** entity that is represented by real data, most likely retrieved from an online source. So, for example, data for your MOVIE entity might come from a Kaggle dataset. This dataset must have at least 100 rows.
- The rest of your tables can be made up of a combination of real data and generated data. At least one of these tables must also have at least 100 rows. The others can be smaller but should not be trivially small.
- Each of your tables must have several defined attributes.
- Generated data can come from other sources. You can create some yourself, you can use generative AI, or another individual can generate the data for you. A reminder that **you must do all the other project work**, such as describing the datasets, creating the schemas, and eventually writing the queries, yourself.
- Remember that your use of the data must respect the licensing requirements for those data. You must also cite the source of the data.

Expectations are as follows:

2a) *Description of the dataset:* Roughly two paragraphs describing the dataset structure, for instance:

- What types of information are contained within the proposed database?
- How many tables?
- How many entities and attributes in each table?
- What are the logical connections between the tables?

2b) *Dataset generation:* Using a bulleted list, describe how the data were retrieved or generated. Provide URLs, other reference information, prompts used, and anything else of relevance.

Sometimes I find it necessary to try a whole series of prompts before I get what I'm looking for; in such cases I try and build a single prompt that generates new data according to my criteria. For clarity I recommend you do the same if appropriate.

Section 3: Conceptual and Internal Models

In this section you will generate diagrams showing the conceptual and internal models for the database you will build in Part 2 of the project. You do not need to provide any text in this section apart from indicating which is the conceptual model and which the internal. The two figures must be clearly legible.

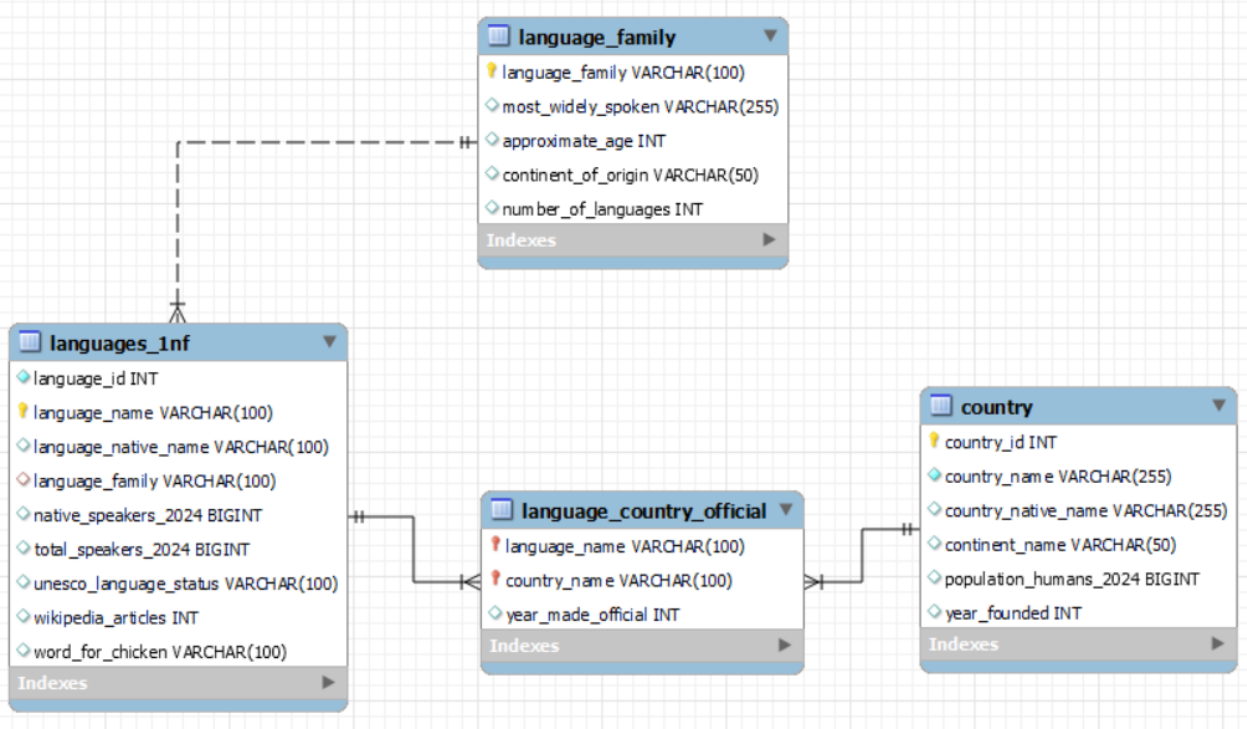
There are many resources you can use to draw models. For the course I have been using a combination of tables in PowerPoint and MySQL Workbench: in the latter case I created the database by choosing "Database" -> "Reverse Engineer" in the application. Another program I like is draw.io, which is good at drawing connections between visual elements.

3a) Provide a conceptual model on one page.

3b) Provide an internal model on a separate page.

Conceptual model examples can be found in the lecture slides.

Here is an example of an internal model:



Rubric

	Excellent (100%)	Very Good (80%)	Acceptable (60%)	Borderline (40%)	Unacceptable (0% - 20%)	Notes
Introduce the dataset	Clearly written, reasons for choosing the dataset are clear and compelling. Associated questions are appropriate to the data at hand.	Clearly written, reasonable connection between the questions and the dataset.	Some aspects of the section are not completely clear. Connection between questions and dataset is plausible.	Significant clarity issues. Connections between dataset and questions are not well articulated but still possible.	Writing is unclear. Poor to no connection between data and questions of interest.	
Describe the dataset	The content in the database are clearly laid out and easy to understand. The number of entities, attributes, and tables is clear. Logical connections between the tables are comprehensively explained.	The contents of the database are well laid out and easy to understand. The number of entities, attributes, and tables is clear. Logical connections between tables are largely sensible but some are missing or unclear.	The contents of the database are laid out at an inadequate level of detail. The number of entities, attributes, and tables is clear. Logical connections between tables exist are present but not well explained.	The contents of the database are poorly laid out. Incomplete or inaccurate descriptions of the number of attributes. Logical connections between tables are incomplete and/or unconvincing.	The contents of the database are unclear. The number of entities, attributes, and tables is missing, unclear, or incorrect. Connections between tables make no sense.	
Generate the dataset	Methods by which dataset were constructed are clear. The reader could generate the dataset (or something very similar to it) by following the procedures laid out in this section.	Methods for dataset construction are reasonably clear, although the reader may not be completely able to reconstitute the dataset.	Methods for dataset construction are provided, but the reader would have serious difficulty regenerating the dataset. Important details are missing or unclear.	Methods for dataset construction are incomplete and it is unclear how the data were generated in all cases.	Methods for dataset construction are disorganized and/or lacking.	Citations and licensing information for all publicly available datasets must be provided.
Conceptual and Internal models	Diagrams are clear and complete, with appropriate levels of detail shown in both models. Model structure is perfectly aligned with prior descriptions of the dataset.	Diagrams are clear and complete, with appropriate levels of detail. Model structure is well aligned with prior descriptions of the dataset in the project.	Diagrams are mostly clear and complete, with small errors (lack of clarity, incorrect designation of attribute types, etc). Model structure is clearly related to prior descriptions in the project but the two do contain some small inconsistencies.	Diagrams lack clarity due to poor graphical design choices. Model structure can be related to prior descriptions but many inconsistencies exist.	Diagrams are absent or do not adhere to model standards.	