Lab 2: Database requirements analysis and design

Due date: Friday, April 19, 7:00pm.

Note: This lab involves a number of deliverables that are due before April 14. The due date above is for the final set of the deliverables.

This is a team lab. Each team submits one set of deliverables. Each team member receives the same grade for the assignment.

Course Project Information

This quarter, CSC 366 comes with a team course project. General information about the project is provided below.

Customer. Our customer this quarter is CalMatters, a California investigative journalism non-profit. CalMatters is the sponsor of the Digital Democracy project (http://www.digitaldemocracy.calmatters.org), a Cal Poly project providing access to annotated transcripts of California legislative hearings and related information.

Project: CalAccess LEAKS (Lobbying, Election Activity Knowledge-base System). The Digital Democracy Project Database (DDDB) is a large relational database containing information about the legislative hearings, their full annotated transcripts, as well as information about all the individuals (and organizations they represent) who testified in the legislative hearings. It also tracks the means of influence-seeking, such as campaign contributions, gifts to lawmakers and behests on their behalf. DDDB uses several publicly available datasets and databases to ingest and maintain current its contents. It also stores a variety of original information not available from other data sources.

The Digital Democracy project tracks Lobbyists (a “Lobbyist” is a special-purpose designation in California, requiring an official registration) and lobbying activities/advocacy within the legislative process on behalf of various organizations seeking to influence the passage of different laws. It also tracks campaign donations received by the Lawmakers.

In California, the source of this information is the publicly available CalAccess Database maintained by the Office of the Secretary of State. CalAccess database consists of several spreadsheets (distributed in TSV format) documenting all the public disclosure forms that must be submitted to the Secretary of State Office. The Digital Democracy project uses CalAccess as one of the data sources, however the project’s usage of CalAccess is far from the most effective.

In this project each team will design a database to store information about Lobbyists, lobbying activities, California state (and local) elections and candidates, and the financial transactions information associated with the elections. Each team will build software that extracts the necessary information from a variety of CalAccess spreadsheets, and puts it in the database it designed. In addition, each team will develop a small data warehouse to host some aggregate information that can be extracted from the primary LEAKS database, and will create software to populate the LEAKS data warehouse.

Full details of the project will be provided in a series of requirements specifications made available to you throughout the quarter. The first part of the documentation is made available now.

Lab Overview

In the course of this lab, you will
• receive an overview of the customer requirements via direct communication with the customer (April 9);
• prepare an initial design of the database for the project (April 11 - 18);
• elicit additional requirements for the database via direct communication with the customer (April 18);
• extend the database design to model the customer requirements in full (April 18 - 19).

The lab assignment spans four lab periods, and has four distinct steps, each of which serves a different goal and yields a set of deliverables. This document specifies your assignment for each of the periods as well as final deliverable.

The overview of the lab structure is below. Details of each step are discussed in the followup Sections.

April 9, 2022: 3:10 - 3:30pm: Groups are announced. Thomas Gerrity joins us for an initial presentation. Initial project documentation is released. During the follow-up lab periods, the groups will study the documentation start the initial database design.

April 11: 4:40 - 6:00pm: Initial database design. We will work on identifying major components of the database model and on discovering the key relationships between them. This is a guided team activity - we will work together on some aspects, each team will work on some initial prompts, we will eventually form some sort of consensus on the important aspects of the database model for the application.

April 12 – April 18: Independent team work (we will have a double lecture on Tuesday, April 16, so all team meetings during this time will have to take place outside of class time) on the conceptual design of the database. On April 16, we will release additional project documentation detailing some of the forms you will be extracting the information from.

April 18: 3:10 - 4:30pm: Q&A session with the customer. Each team will have an allocated portion of time for questions. Questions/answers must be recorded and posted later. (Note: the instructor will be out of town, the Q&A session will be moderated by Dr. Khosmood). The subsequent (April 18, 4:40-6:00pm) lab period is reserved for team meetings.

April 18 - April 19 teams work on the full database design, finalize the main deliverable document and submit it.

Assignment

April 9: Group Creation and Initial Presentation

We will spend the first few minutes of the class on April 9 on group creation. Group lineup will be announced by the instructor at the beginning of the class. We will ensure that the group lineups are feasible (e.g., that we are not missing three people on a single 4-person team). Teams will contain three to four people.

Following team formation, our customer, Thomas Gerrity, will present an overview of your project, followed by a short, informal Q&A session. The main purpose of the presentation is to give you a good understanding of what the project entails. After the presentation, we will release the customer-supplied documentation.

Assignment. Class attendance is required for everyone. If you are unable to attend, please PM the instructor on Slack.

Deliverables. Each team shall come up with a name (due at the end of the lab period or as soon as you come up with it - I need to create slack channels for each team). Over the course of the week, each team shall go over the documentation received from the customer, and will prepare, in writing a list of questions for the Q&A session with the customer (Due April 19).
April 11: Initial Design

The main task for this lab is for each team to build an Entity-Relationship model of the database for the application you are building in the course. We will build this model in two stages. First, on April 11, you will work out the overall database model, and will determine the most important components (entity sets and relationship sets) of the model. This will be done via a set of instructor-guided activities during the April 11 lab period. After that, you will refine the model to incorporate specific requirements that the customer placed on the identified components.

Assignment: Each team shall fully participate in the model-building activities that occur during the April 11 lab period. As the result of these activities, each team should produce a draft initial (simplified) Entity-Relationship model for the customer database by the end of the lab period.

Deliverables: The key deliverable of this stage is the simplified E-R model of the customer database. Each team will produce a draft of the model by the end of the April 11 lab period. This deliverable is informal (you will revise it later) and will be a result of brainstorming activities of your team during the April 11 lab period.

The model description shall consist of the following:

1. List of entity sets. For each entity set specify its attributes, and the primary key. Identify weak (if any) entity sets, specify their owners.

2. List of relationship sets. For each relationship set, identify participating entity sets, list any attributes, identify its type (one-to-one, one-to-many, many-to-many), and whether it is the defining relationship set for a weak entity set.

3. An E-R diagram of the model.

After the lab period is over, create a text document representing your initial E-R model. The first two parts of the E-R model description should be plain (or formatted) text. The E-R diagram shall be drawn using some graphics editor or presentation program. Eventually it will be embedded in your document, but at this stage (since it is an early draft that you will modify many times) you can keep the E-R diagram in a separate file. I use xfig to create diagrams that are embedded into latex/PostScript files. I use MS PowerPoint or Google Slides to create diagrams that are embedded into MS Word documents/Googledocs documents.

The document you create shall be shared with the instructor (see submission instructions below).

April 18: Question and Answer Session

During the April 12 - April 18 period, each team shall continue revising and improving its E-R model design. Our April 16 class is a double lecture, so all team activities shall be conducted outside the class time.

On April 16 we will release additional documentation identifying the specific forms stored in the CalAccess database that you will be extracting the information from. This documentation should not have significant effect on the overall structure of the E-R model, but it shall provide you additional information about the attributes for different entity sets and relationship sets in the model.

The customer documentation released to you contains initial information about the database and the software application you will be building this quarter. The documentation reflects the information our customer chose to provide. This documentation, however, may be both (a) unclear and (b) incomplete. Your next step is to analyze the documentation, formulate additional questions to the customer, and solicit and record the responses.
Preparation. Before the April 18 lab, each team shall study the materials provided by the customer. The goal of each team is to identify any aspects of the application domain, information about objects to be stored in the database, customer needs (w.r.t. the database), use cases to be implemented, and so on that require additional clarification. Each team shall prepare a list of questions that the team wants to ask the customer during the Q&A session. Note, that in addition to the customer, you may direct your questions at the instructor.

Please note: initial list of questions must be prepared by each team independently of other teams. However, once lists are prepared, I allow for teams to compare them. I expect that many questions will coincide or be similar (it is ok if multiple groups have similar/same questions). It is also ok for one group to come up with a question, no other group has asked — there is no need for other groups to add it to their lists, since each question only needs to be asked once, and all teams benefit from the answer.

Each team shall record a list of the questions as a googledocs document.

Q&A session. Each team will be given about 10 minutes of time to ask their questions and receive answers from the customer and the instructor. We will set the order of questioning by a simple lottery at the beginning of the lab. I strongly suggest that each team sits and works together during the Q&A. Please note the following:

- Listen carefully to the questions other groups are asking. It is very likely, that some of your team's questions will be posed by teams who get to ask questions earlier.
- If your group runs out of questions yield your time to the next group.
- If your group has unanswered questions left after its Q&A period, wait until all groups ask their questions. If the question has still not been asked, you will have a chance to ask it at the end of the lab.
- You are certainly allowed to ask questions that are not on your list - some questions might occur to you as a result of things said during the Q&A session. However, make sure you document all the questions you asked and all the answers you received.

Each team is responsible for recording the answers to all questions its members have asked (whether those questions were prepared or were asked on the spot).

After each team had a chance to ask questions, if any additional questions remain, a free-form Q&A period will begin and will last until the end of the lab period. Any questions that went unasked, or unanswered during the Q&A session should be submitted in writing (via email) to the instructor. The questions will be forwarded to the customer, who, in turn, will provide answers prior to the Friday, April 12 lab.

Deliverables. After the Q&A session, all teams, in collaboration, shall create the project knowledge base. The project knowledge base will be represented this year as a single GoogleDoc file created by the instructor and shared (with editing permissions granted) to all students in the class. The structure of the knowledge base is left up to the teams (generally speaking, you want to break the entire knowledge base by topic: e.g., group information about one type of data in one place, and information about another type of data — in another). A single question can appear in multiple places in the knowledge base.

The knowledge base is one of the two main deliverables for this lab. The initial state of the knowledge base shall be completed by the main Labs 2 due date. The intent is to have a useful and accessible resource for your further development.

April 11- April 19: Database Design

Assignment. The main task of this lab is for each team to prepare an initial full database design document. Teams shall use the time between the April 11 class, and the April 19 (in and out of the lab) to complete the assignment to revise and enhance the initial E-R model.
Deliverable. Each team is given time until Friday, April 19, 7pm to finalize and submit the database model document.

The design document prepared and submitted by each team shall be a GoogleDoc shared with the instructor (see instructions below). At a minimum, it shall contain the following information:

- Team name, list of all team members (names, Cal Poly email addresses)
- List of entity sets for the proposed database.
- List of attributes for each proposed entity set.
- Identification of primary keys for each proposed entity set.
- Identification of all weak entity sets in the proposed database, and of their discriminating attributes.
- Identification of any other entity set constraints for the proposed entity sets.
- List of relationship sets for the proposed database. For each relationship set, the following must be indicated:
  - All participating entity sets.
  - All clarifying relationship set attributes.
  - Type of the relationship set (one-to-one, one-to-many, many-to-many).
  - If the relationship set is an identifying one for a weak entity set.
- List of other relationship set constraints.
- List of any class hierarchies and/or aggregates in the database model.
- List of relationship sets (with all the information as above) associated with the aggregates.
- E-R diagram of the proposed database design embedded into your document.

The design document must created as a GoogleDocs file. See comments above about the software to use to draw the E-R diagram. The E-R diagram shall contain all entity sets and relationship sets. It shall also contain all key attributes for all entity sets and all relationship set attributes. Other attributes can be omitted from the diagram for clarity. All constraints that can be shown on the diagram, shall be shown.

Deliverables Overview and Submission Instructions

Overall, the lab has the following formal deliverables:

<table>
<thead>
<tr>
<th>No.</th>
<th>Deliverable</th>
<th>Due:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List of questions for Q&amp;A session</td>
<td>April 18, 3:00pm</td>
</tr>
<tr>
<td>3</td>
<td>Project knowledge base</td>
<td>April 19, 7:00pm</td>
</tr>
<tr>
<td>4</td>
<td>Database Design document</td>
<td>April 19, 7:00pm</td>
</tr>
</tbody>
</table>

Submission. The project knowledge base file is the responsibility of everyone. This document will be shared with you on April 11. No "submission" activities are needed - just make sure that your team participates in editing it.

Each team shall create a GoogleDocs directory for the project and share it with the instructor’s gmail address dekhtyar@gmail.com. Name your directory CSC366-Project-TeamName, where ”TeamName” is the actual name of your team. Include in the directory a short README document with the list of students and email addresses. For Lab 2, create a Lab2 directory inside your project directory, and place in it your design document, the list of questions, and any other documents you create for this lab.
By the **April 19, 7pm deadline**, one person of each team shall send me a PM on Slack informing me that the Lab 2 deliverables are ready. (you can share the project directory with me at the same time, or any time earlier).

Please, keep the soft copies of all submitted documents. When you edit your Lab 2 submissions during Lab 3, make sure that you create separate documents for your Lab 3 deliverables, while still maintaining a copy of your original Lab 2 submission. This is IMPORTANT!