CSCI 585 - Database Systems (Spring 2014)
Homework Assignment 1
Prof. Shawn Shamsian
Due: 02/19/2014, 11:59PM

The main goal of this assignment is to: (1) Design a conceptual schema using the ER/EER Data model; (2) Convert the schema to tables in an Oracle DBMS; (3) Run sql queries on the database to get the specific data.

1. ER/EER data model (40pt)

Design a conceptual schema based on the specification below. You should submit an ER/EER diagram of your schema design using the notation given in the class. Your diagram should indicate all relevant classes, subclasses, relationships (weak & strong), relationship cardinalities and degrees, total participation, attributes, primary keys. In addition, specify whether each attribute is single-valued or multivalued, stored or derived, and atomic or composite.

In your design, you may make and state reasonable assumptions if they are not provided in the specification.

Please design a database schema for *LinkedIn* (https://www.linkedin.com). It should store and manage the following information.

Every user of this website has his/her own profile. The profile contains basic information: (1) a unique user id. (2) the user's email address. (3) the user's full name consisting of first name and last name. (4) the user's country and zip code. (5) the user's status (employed [job title, whether self-employed, company], job seeker [most recent job title and company, time period] or student [school/university, dates attended]). (6) the user's photo, image, presentation or other relevant documents or files. (7) register date. The profiles of the group and company in the following follow the same way as the user, except some different fields.

A user can only access to LinkedIn with his/her email address. The user may find his/her friends and add connections with them by importing the email contacts with their email address.

A user can join groups, which are comprised of the group id, group name. The user can also follow influencers or companies in the same way. For the table of connections, one field like connection_type may be used to distinguish the general connection, group, influencer and company connection.

A user can share an update with his/her connections, no matter whether it's a profile or status update. The post contains: (1) post id. (2) post type. (3) user id of the post. (4) post content. (5) share with public + twitter, public or connections. (6) date and time. (7) files if any. (8) number of likes. (9) number of comments.

A user can add a comment to other users' post. The comment contains: (1) user id of the sender (2) comment id. (3) date and time sent. (4) comment content. (5) whether the sender likes the post. (6) whether the sender shared the post. (7) post id

2. Designing the tables and using the oracle DBMS (15pt)
To complete the rest of this homework, you need to install Oracle DBMS on your machine. You may refer to Appendix a for instructions for installing Oracle on your machines.

Convert your schema in Part 1 to tables. For each table, specify the table name, its fields and their data types, the primary key and foreign keys (if there are any). For each foreign key (if there is any), specify which table it refers to. You are free to choose proper data types and null-ability assumptions according to your own understanding of the real *LinkedIn* website (if not specified in this project specification). You may want to optimize your table design in this step using 2NF or 3NF described on the textbook since your ER/EER model may not be optimal. But it is fine if your schema is not perfect as long as it works properly for Part 3. Notice that you are NOT allowed to use the schema of the test data tables, otherwise you can only get partial credit for this part.

Submit your table design for this part.

3. SQL query on the database (40pt)
In this part, you are expected to write sql scripts on oracle DBMS based on the test data HW1 data.xls. Note that only **sql** files are accepted. **Please include a drop_db.sql script file in your submission.**

1) Create tables. You are expected to submit a script **createdb.sql** to create a database and insert the test data. Note that there are several tabs in the Excel file. You may be use data of one tab for one or more tables in your database. If there is any attributes which are unavailable in the Excel file, you can make reasonable assumptions and fill them out yourself.

2) Find the full names of users who join the group of University of Southern California (abbrev. as USC). (**find_names.sql**)

3) Find all the post contents since 01/01/2014 of the user whose full name is Jackie Chan. (**find_posts.sql**)

4) Find all the post contents of the user whose full name is Jackie Chan and posts are without files in January. (**find_posts_without_file.sql**)

5) Find the full names of users who send more than 2 comments to Jackie Chan. (**find_senders.sql**)

6) Find email address of users whose connections include both Jackie Chan and Lady Gaga. (**find_email_address.sql**)

7) Get the names of users who commented to the messages posted publicly by USC in January. (**get_comments_names.sql**)

8) Find users (id, full name, email address, register time) in the descent order of time who registered on LinkedIn on January 24th. (**find_registers. sql**)

9) Find all users (id, full name, email address) who are from USA and have at least one group in common. (**find_users_usa. sql**)

10) Find all users that have not posted any status update in the last week and have connections with at least one group which has more than 10 members. (**find_users_group.sql**)

4. Table design question (5pt)
In Part 3, consider a situation, if we need to connect several tables and get specific data, we may use 'in' or 'inner join' in the SQL statement. When a large amount of information is invoked, which one is more efficient? How can you further improve its performance by modifying the table design.

Submission Guidelines
You must submit a compressed folder via DEN (https://www.uscden.net) under the name **<your name>_HW1**.zip that contains the following files:
- ***report.pdf***

In the report, please give solutions to Part 1, Part 2, and Part 4. For Part 1, you can either scan your hand-drawn diagram or use any diagramming software (Dia is recommended, if you have no idea about what diagramming tool to use: http://live.gnome.org/Dia) for this task. For Part 2, you only need to give the solution to the table design question. If you wish to edit your report with MS Word, please be reminded to convert your final report into PDF format.
- ***SQL scripts***

Please submit the scripts mentioned in Part 3. Make sure to name the script as stated in bold.
- ***Readme.txt***

This text file contains your name, USC-ID and your email address.

*****Please do not submit via e-mail. No late submission will be accepted.*****

Appendix
A. Oracle installation guides
The Oracle setup file can be downloaded here.
(http://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html)
Make sure to download the suitable version that works on your operating system.

B. Click here (http://www.oracle.com/pls/db111/portal.portal_db?selected=11) for the official Oracle installation guides for different operating systems.

- If you use mac system, I suggest you install the Oracle database on a virtual machine, please refer to http://dimitrisli.wordpress.com/2012/08/08/how-to-install-oracle-database-on-mac-os-any-version/. If it fails in the end, you may also install MySQL on your system. Please mention it in your submission.

C. Click here (http://www-scf.usc.edu/~jgui/csci585/Oracle_11g_R2_Installation_Guide.pdf) for the Oracle 11g installation tutorial for Microsoft Windows users.