Home work 1

Due date: 14 January, 2019, 11:55pm IST

Instructions

1. This homework must be completed *individually.*
2. Use the file template.tex to fill in your answers. Rename the tex file as EntryNumber.tex (eg. 2018MCS456.tex). Generate your pdf from this file.
3. Your pdf should be named EntryNumber.pdf (eg. 2018MCS456.pdf). Please follow this naming convention strictly.
4. You should upload *both* the pdf file as well as the tex file onto moodle.
5. No deadline extensions are allowed for this homework. Late submissions will not be considered.
6. Use PostgreSQL 11 for your homework. See `https://www.postgresql.org/download/` for instructions on how to download and install it on your OS. .
7. How to start postgres command line:
   (a) Linux: Run `sudo -u postgres psql` on terminal. Enter sudo password when asked.
   (b) Windows: Open 'cmd' and navigate upto 'bin' folder of postgres installation. Type `psql -U postgres` to get postgres prompt.
8. If you have questions, *post them on piazza* and tag your folder with 'homework1'. We will continue to answer questions until Jan. 13, 2019, but will not do so beyond that. You may have to wait a few hours to have questions answered.
9. **Penalty:** The maximum penalty for this Homework is a deduction of **2 marks** at the end of the semester from your final score out of **100**. To avoid all penalties:
   (a) Solve **all** questions. We will check your submissions individually, and incomplete and shoddy work will be penalized.
   (b) Submit your answers *in the format specified.*
   (c) Submit your files *on time.*

Database Design

1. You must have heard of the website `https://www.linkedin.com/` used for professional networking. It uses a database in the background to store the data. You have to reverse engineer a substantial portion of the database and show:
   (a) The ER diagram, with a brief explanation for each entity and relationship. You may draw the diagram by hand and insert a picture of it. However, you will lose marks if the diagram is unreadable.

       In addition to the diagram, list the entities and their attributes as well as the key in tabular form so that it is easy to read.
   (b) Give the set of relations which can be derived from the ER diagram (Relational Model).

(c) Identify all keys and FDs in your relational design.

(d) Show sample data (5-10 tuples) for each of your tables (this should be real data from some snapshot(s) of the website).

(e) Ensure that you have at least one of each the following: weak entity set, non-binary relationship, hierarchical relationships, constraints such as type of relationship and referential constraints. List these in a separate table.

2. Choose 3 entities, each of which contains at least 3 attributes from the ER diagram above. Construct a universal relation from them. Step by step, decompose these relations into 2NF, 3NF and BCNF. Along with each decomposition, give an example of a redundancy it eliminates (that is, a set of example values in the universal relation, which is then projected out in the decomposition).

3. Visit `https://www.kaggle.com/datasets`. Download 3 datasets of varying sizes available in csv format (no. of tuples should be between: 50,000-100,000 tuples, duplicate the data if needed).

   (a) Create a simple schema (a single table is fine) to store this information in Postgres.

   (b) You should insert the data in following modes: i) bulk load (find out how to do this), ii) generate a .sql file containing insert statement for each tuple (you can programmatically generate .sql file from csv file containing insert statements) and execute it to insert the data, iii) bulk load using JDBC or any of the available Python drivers for PostgreSQL, and, iv) inserting each tuple one by one using JDBC or Python.

   (c) Report in an easy-to-read tabular form, the following information: configuration of the machine, name of the datasets, sizes of the raw datasets, time to bulk load, time to load using inserts. Give a brief explanation of how you performed the three kinds of data loading operations.

4. Repeat the exercise above, but with datasets of sizes between 5GB and 10GB. Explain how you generated this data (that is, if you had to duplicate the data and how many times). Report the same statistics as above.