

Assignment 1 – Relational model

Let's model data from Stack Exchange. You can find the datasets available for download here: <https://cs.rit.edu/~mmior/csci-620/askubuntu.tar.gz>. Specifically, we will be looking at data from Ask Ubuntu, a site similar to Stack Overflow. The database system we are going to use in this assignment is [PostgreSQL](#).

Your tasks

1. Provide an ER diagram to represent the Stack Exchange data. This does *not* mean your ER diagram should match the data files. You should aim for entity sets to represent things which will be stored in a database, which may not correspond to **any particular data file**. (Note specifically that there is no corresponding “account” for the AccountId of users, but this AccountId can just be stored as an attribute of the user. This is commonly the case, as it is here, when the ID refers to some external system and you cannot create a foreign key relationship.) **(25 points)**
2. Create a relational model to store Stack Exchange information based on your ER diagram. Write SQL scripts to create all tables in the database including primary and foreign keys. **You must provide these scripts as executable code.** (A text file of SQL statements is sufficient.)

You need to explore the dataset to decide the sizes of the attributes in advance and make sure to use the appropriate data type for each column. Provide a description in your report on how you solved each of these issues. **(20 points)**

3. Provide a description of the contents of the files in the Stack Exchange dataset. This should explain the purpose of each file and its contents. **(5 points)**
4. Provide a program to load the Stack Exchange data into the database you created in Q2. Your program needs to load the whole database in approximately four hours using **commodity hardware** (actual runtime will likely be much less). You may skip inserting any rows that would violate foreign key constraints. Provide a description of how you solved these issues and report your timings. This program may simply be a text file of SQL statements. **(45 points)**
5. Provide a program that connects to the previous database and creates a transaction to insert three rows of data. Force an error in row #2 so that the transaction aborts. Your program should check to ensure that the database is in the same state as before. (That is, row #1 should not have been inserted.) **(5 points)**