W4111-002 Database Systems
Fall 2014
December 16, 2014

Final Exam. 40% of final grade.
(Closed Book Exam. No calculators or notes permitted.)

This exam consists of 170 points, one per minute of the exam. There are 7 questions, and this exam is 12 pages long. If the percentage grade on this exam is higher than the grade on the midterm, then the midterm grade will be replaced with the final exam grade. Show all work, since partial credit may be given. Questions must be answered in the spaces provided on this sheet; the spaces correspond to the expected length of answers. If you need additional space (due to erasures or large handwriting etc.) use the back of the page and indicate (in the answer slot) that your answer is continued on the back of the page. *We will not read excessively long answers.* Make sure that your name and UNI (or email address) appear below.

Name: _____ UNI/Columbia email: _____

1. (20 points) In at most two sentences each, explain the meaning [1 point] and relevance [1 point] of the following terms as they relate to database systems:

   (a) Checkpoint .....

   (b) Object-Relational database .....

   (c) Views .....

   (d) Log .....

   (e) Update anomaly .....

   (f) Assertion .....

   (g) Isolation .....

   (h) Disk block .....

   (i) Deadlock .....

   (j) Atomicity .....

2. (15 points) In at most two sentences each, answer the following short questions:

   (a) Suppose a user Mary grants a privilege to another user Joe with the grant option. Joe then grants the permissions to Alice and Bob with the grant option. Mary revokes the grant option (but not the permission itself) from Joe. What changes result from the revoke statement? .....

   .....

   .....

   .....

   (b) Is every scheme that is in Boyce-Codd Normal Form (BCNF) also in Third Normal Form (3NF)? Why or why not? .....

   .....

   .....

   .....

   (c) Let R and S be tables with n rows and m rows respectively. What is the smallest possible number of rows in $R \bowtie S$? What is the largest possible number of rows in $R \bowtie S$? .....

   .....

   .....

   .....

   (d) Suppose that a commonly-asked query is optimized once, and never re-optimized, to save optimization time. Identify three (or more) kinds of changes that, if they take place after optimization time but before execution time, would impact the choice of an optimal plan. .....

   .....

   .....

   .....

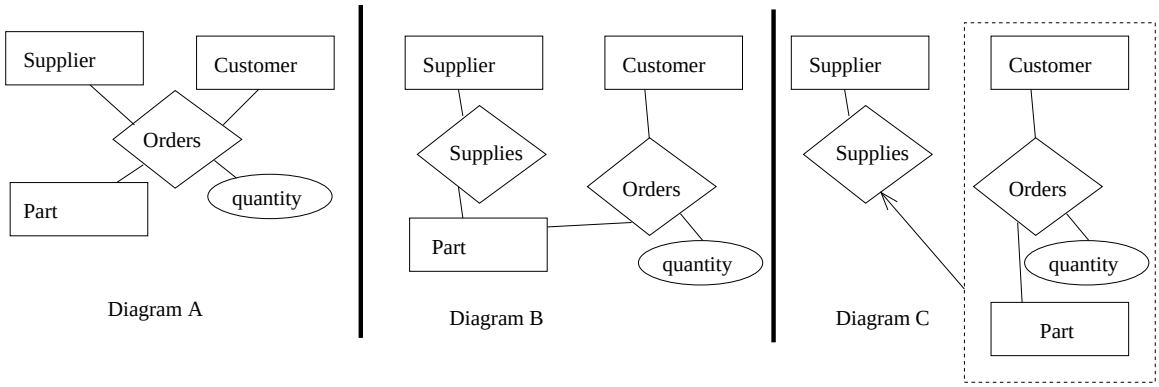
   (e) What is wrong with the following statement by a database administrator? "Whenever I create a new database, I create indexes on all columns of all tables. That way, I never have to respond to user requests to add new indexes, and the indexes are there to help the optimizer find good query plans." .....

   .....

   .....

   .....

3. (20 points) Below are three ER-diagrams. Explain the differences between the three diagrams in terms of the information that can (and cannot) be represented. Explanations should be in plain English, and should not use any technical ER-diagram terminology.



4. (20 points) In each of the following subquestions, a decomposition of a scheme $ABCD$ into a subscheme is described. In each case there is a different set of functional dependencies. For each part, identify the answer to each of the following yes/no questions. (For these parts, just a yes or a no is sufficient.)

   - Is the decomposition lossless?
   - Does the decomposition preserve dependencies?
   - Is the resulting scheme in 3NF?
   - Is the resulting scheme in BCNF?

   Finally, in one sentence explain when (if ever) the decomposition is a good choice.

   (a) BCD and AD where $ABC \to D$ and $D \to A$.

   Lossless? .....
   Preserves dependencies? .....
   3NF? .....
   BCNF? .....
   Good choice? When? .....
   .....

   (b) BCD and AD where $BC \to D$ and $A \to D$.

   Lossless? .....
   Preserves dependencies? .....
   3NF? .....
   BCNF? .....
   Good choice? When? .....
   .....

   (c) ABC and CD where $AB \to C$ and $C \to D$.

   Lossless? .....
   Preserves dependencies? .....
   3NF? .....
   BCNF? .....
   Good choice? When? .....
   .....

   (d) ABC and ABD where $AB \to C$ and $AB \to D$.

   Lossless? .....
   Preserves dependencies? .....
   3NF? .....
   BCNF? .....
   Good choice? When? .....
   .....

5. (12 points) Consider a relational database table Projects(ID,State,Budget) that records the identifier and budget of all government projects in each state. You are the DBA of this database system. You may assume that every state always has at least one active project. Several senators have told you that they want to keep track of the current total budget in each state. You are considering two options:

   - Creating a view StateProjectsView defined by the following SQL statement:
     `Select State, sum(Budget) as Total From Projects Group By State`
     Users would then be asked to reference the view if they want to see the total budget.
   - Creating a new table called StateProjectsTable with columns State and Total. You would also create triggers that, on every insertion, deletion, and update to the Projects table, made the corresponding change to that state's total in the StateProjectsTable table. Users would then be asked to reference the StateProjectsTable if they want to see the total budget.

   (a) (6 points) Explain the trade-offs you would need to consider when determining which of the two scenarios is better. For each solution, outline (in one sentence) a scenario when it is the preferred solution. .....

   .....

   .....

   .....

   .....

   .....

   .....

   .....

   (b) (2 points) Are triggers executed within the same transaction as the triggering update, or in a separate transaction? .....

   .....

   (c) (4 points) Suppose the database system did the wrong thing, i.e., did the opposite of your answer for part (b) above. Use the example above to explain what could go wrong. .....

   .....

   .....

   .....

   .....

   .....

   .....

   .....

6. (36 points) Consider a hospital that wishes to record and keep up to date records of the patients within the hospital. Suppose that current patient/room information is kept in a table `PatRoom(patient-id,room-id)` where the two columns are unique identifiers for patients and rooms. Sometimes patients change rooms, which means that the room-id for a patient would be updated. Suppose that a patient can be assigned to just one room, and that a room can contain just one patient. There are separate `Patient(patient-id,...)` and `Room(room-id,...)` tables that keep additional information about patients and rooms, respectively. Some patients (e.g., those in the emergency room) may not be assigned to hospital rooms, and some rooms may be empty. The hospital periodically determines its current room occupancy level by running the query `Q`:
`Select count(*) from PatRoom`.

   (a) (6 points) What integrity constraints should be associated with the `PatRoom` table? Write them in SQL. .....

   (b) (6 points) Suppose that a patient with patient-id 123 is moved from room R45 to room R67. Suppose that this change is recorded in the database via two transactions: one to delete (123,R45) from the `PatRoom` table, and another to subsequently insert (123,R67) into the table. Explain what is wrong with this approach. .....

(c) (6 points) Suppose as before that a patient with patient-id 123 is moved from room R45 to room R67. Suppose that this change is recorded in the database via a single transaction that deletes (123,R45) from the `PatRoom` table, and inserts (123,R67) into the table. Is this sufficient to meet all of the hospital's needs? *Hint: What happens if two patients switch rooms?* .....

.....

.....

.....

.....

.....

.....

.....

(d) (6 points) Suppose that to switch a pair of patients, we use the following single transaction.

```
 i. Delete (P1,R1)
ii. Delete (P2,R2)
iii. Insert (P1,R2)
iv. Insert (P2,R1)
```

Suppose also that at the same time that this transaction is submitted, the query `Q` above is concurrently run to determine the room occupancy levels. Explain what might go wrong (in the absence of any concurrency control) if the whole of query `Q` were executed between steps (ii) and (iii) above in a schedule. .....

.....

.....

.....

.....

.....

.....

.....

.....

(e) (6 points) Suppose that strict two-phase locking were used by the hospital's transaction-processing system. Would the interleaved schedule in part (d) above be possible? If not explain why not. If so, explain why this observation does not violate correctness properties of strict 2-phase locking. Assume that to execute a query, a shared lock on each row needed by the query is obtained. .....

(f) (6 points) If the transaction were instead written as

```
 i. Update PatRoom set room-id=R2 where patient-id=P1
ii. Update PatRoom set room-id=R1 where patient-id=P2
```

and query `Q` was scheduled between the two update statements, then `Q` would appear to give the correct count of rooms occupied in the absence of concurrency control. Explain why such a schedule would not be allowed under strict two-phase locking, and why this is appropriate even though `Q` would give the right answer without locking. .....

7. (47 points) In a census database the schema has the following tables:

   - Person(person-id,date-of-birth,gender,zip-code,job-code)
   - Regions(zip-code,area)
   - Jobs(job-code,title,education-level)

   The person table records information about every person in the USA, about 300 million rows of data. Regions are determined by their zip-code, and have a total area within the zip code. Jobs are categorized based on job-codes, and have titles and an associated education level. The census database is gathered once every ten years and is not updated once it has been built.

   (a) (8 points) Write the following query in SQL: List all zip-codes in decreasing order of density, where density is the total population within that zip-code divided by the area of the region. You may assume that all zip-codes have a non-zero population. . . . . .

   (b) (8 points) Write the following query in SQL: List all zip-codes that have the following property: for each job-code in the Jobs table, the zip code has one or more people with that job-code. In other words, find zip-codes in which every job is represented. . . . . .

(c) (4 points) Suppose that a B+-tree index is created on the date-of-birth attribute of the Person table. If a record-identifier takes 8 bytes, a date-of-birth takes 4 bytes, and a disk block is 1000 bytes, how many leaf nodes would you expect given that each node is 2/3 full? .....

.....

.....

.....

.....

.....

(d) (3 points) If the branching factor in this B+-tree was about 50, how many levels would the B+-tree have? Make sure to count the root and leaf levels in your answer. Round to the nearest integer. .....

.....

.....

.....

.....

.....

(e) (8 points) Suppose we are interested in answering a query of the form

```
Select gender, count(*)
From Person
Where date-of-birth < C
Group By gender
```

where `C` is some threshold supplied by the user. Under what circumstances would the index on date-of-birth be useful in answering this query? Explain how the database system would decide whether to use the index. (No other indexes are available.) .....

.....

.....

.....

.....

.....

.....

.....

(f) (6 points) Suppose we are interested in a slightly different query of the form

```
Select count(*)
From Person
Where date-of-birth < C
```

where `C` is some threshold supplied by the user. Under what circumstances would the index on date-of-birth be useful in answering this query? .....

.....

.....

.....

.....

.....

.....

.....

(g) (10 points) Consider the following relational algebra expression (attribute names are replaced by their first letter):

$$\pi_t \sigma_{e>12 \ AND \ d<01-01-1960 \ AND \ \text{Person.j=Jobs.j}} \, (\text{Person} \times \text{Jobs})$$

Use equivalences of relational algebra to rewrite this expression into an equivalent expression that is more efficient: (a) intermediate results should contain as few rows and columns as possible, and (b) cross-products should be avoided. .....

.....

.....

.....

.....

.....

.....

.....

.....