CS 582 — Database Management System II — Spring 2007

Final

Name:

Signature:

Four questions on 4 pages (2–5). 100 points for 30%.
Material for Question 1 is separated from the text.
Please read the question carefully!

1. (30 points) Consider an example of a XML file representing a CD catalog (print out from the Web, copy 1).

1. Define a XML schema called `CATALOG.XSD` which accepts the given XML data file as one of its valid instances.

2. Add to your XML schema (if you have not done so) to ensure that the following constraints are satisfied:

   (a) The CD title is unique.

   (b) The price is a non-negative number.

3. Design a relational database schema that can be used for the same purpose as the XML and the XML Schema for the CATALOG. It suffices if you list the tables with their attributes and the constraints that you would like to enforce.

4. Design a XQuery that returns the list of CDs published in or before 1985 and costed more than 8 US$. The result should look like the following

   ```
   <list_CDs>
     <cd title=''Empire Burlesque'' year=''1985''/>
     <cd title=''Sylvias Mother'' year=''1973''/>
     ....
   <list_CDs>
   ```

   Avoid repetition if it arises.

5. Design a XQuery that lists the artist and his/her company. The result should look like the following

   ```
   <artist_company>
     <artist>
       <name> Bob Dylan </name>
       <company> Bob Dylan </company>
     </artist>
     <artist>
       <name>Bonnie Tyler </name>
       <company> CBS Records </company>
     </artist>
     ....
   <artist_company>
   ```

   Avoid repetition if it arises.

2. (20 points) A school needs to store the information about students, courses, and transcripts. In a stand-alone application, the school would have organize this information into a system consisting of the following relation schemas (different keys are underlined differently):

Student(Id, Name, Address, Status)

Course(CrsCode, Department, CrsName)

Transcript(StudId, CrsCode, Semester, Grade)

Suppose that the school decides to re-organize the information into a distributed database where information about Student and Course is stored at one site (let us call it SC) and the transcript information is stored at another site (T). We also knows that

- The Student table has 15,000 rows.
- The Course table has 1,000 rows.
- The Transcript table has 300,000 rows.
- On average, each graduate student takes 11 courses.
- On average, each graduate level class has 15 students.
- Only course name student name, and address are more than 10 bytes. None of the attributes is more than 50 bytes. Student ID, course code, grade are about the same size.

A special report needs to be done which requires the computation of the transcripts of all students attending the CS582 class (a graduate class in computer science) in Spring 2007. A transcript usually contains the student ID, name, his/her address, the list of courses that the student has taken, and for each course, the course name and grade. Answer the following:

1. Write a relational algebra expression for the query.
2. Without information about the concrete size of each attribute in the database, shows that the following plan for computing the answer is *correct* and is the *best possible option*:

   - **Step 1: (Site T)** Compute the list of studen IDs attending the CS582 class in Spring 2007. (Call this list L1). Send this list to the site SC.
   - **Step 2: (Site T)** Create the transcripts (without student names and course names) for all students taking CS582 in Spring 2007 from the Transcript table. Let us call this list T582. Send T582 to A.
   - **Step 3: (Site T)** Project the course codes from the list T582. Let us call this list C582. Send C582 to SC.
   - **Step 4: (Site SC)** Compute the join of L1 and Student and project out the student id and name. Let us call the result of this operation J1. Send J1 to A.
   - **Step 5: (Site SC)** Compute the join of C582 and Course and project out the course code and name. Let us call the result of this join J2. Send J1 to A.
   - **Step 6:** At site A, compute the join of T582, J1, and J2 as the answer.

   Hint: One possible way to show *correctness* is to show that the relational algebral expression produced in six steps of (2) can be obtained from the relational algebral expression of (1).

3. (25 points) Son's Superstore (SST) is a new retailer chain that operates in several countries including the 50 states of the US. SST specializes in computer and networking equipments. The chain stores the sale information, denoted by `Sale`, together with the following dimensions:

- `Sid` which is a unique store identification, associated with the geographical location of the store; `Sid` is related in the `location dimension table` whose schema is `Location(Continent, Country, State, City, Sid).`
- `Pid` which is the Product ID; `Pid` is stored in the dimension table `Product(Name, Category, Pid).`
- `Tid` which is the time ID and is stored in the dimension table `Time (Year, Month,Week,Tid).`
- `Mid` which is the manager ID and is stored in the the dimension table `Manager(Sid, Mid)`

Answer the following:

1. Design a relation schema that the company can use to store its fact table. Give an example of a row in the fact table.
2. Assume that each dimension table has 50 rows. How many rows will the fact table of the SST chain will have?
3. Design the necessary SQL queries that can be used to create a table of the following form:

	America	*Africa*	*Asia*	*Australia*	*Europe*	*Total by Product*
Pid_1	1000	100	200	50	700	.
Pid_2	.	.	.	.	.	.
Pid_3	.	.	.	.	.	.
Pid_4	.	.	.	.	.	.
Pid_5	.	.	.	.	.	.
.	.	.	.	.	.	.
Total by Continent	.	.	.	.	.	*TOTAL*

Use the CUBE, ROLLUP, etc. operators to shorten your queries at your discretion.

4. (25 points) To decide whether or not a student can continue receiving their loan, a school gathers information about whether or not a student past performance affects the student performance in the current semester. The following is a sample table.

Sid	Single	FailSome	Current GPA	Fail
S1	Yes	Yes	3.50	No
S2	Yes	No	1.50	Yes
S3	No	No	3.50	No
S4	Yes	No	3.50	No
S5	No	Yes	3.50	No
S6	No	Yes	3.75	No
S7	No	Yes	2.50	Yes
S8	Yes	No	1.90	Yes
S9	Yes	No	4.00	No
S10	Yes	No	2.50	No
S11	Yes	No	3.00	No
S12	No	No	3.00	No
S13	No	No	3.00	No
S14	No	No	2.50	No
S15	Yes	No	3.60	No
S16	Yes	Yes	3.50	Yes
S17	Yes	No	3.10	No
S18	No	Yes	2.75	Yes
S19	Yes	No	3.50	No
S20	Yes	Yes	2.00	Yes

Answer the following:

- Compute the decision tree for the school based on the above training table. Show your computation. Use calculator if desired.
- Give an example of a decision rule based on the decision tree.