Advanced Database Management Systems

Date: Friday 23rd January 2015
Time: 09:45 - 11:45

Please answer BOTH Questions

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text
1. a) Consider the following SQL DDL statements:

```sql
create table advisor
    (s_ID varchar(5),
i_ID varchar(5),
primary key (s_ID),
foreign key (i_ID) references instructor (ID)
on delete set null,
foreign key (s_ID) references student (ID)
on delete cascade);
create table student
    (ID varchar(5),
name varchar(20) not null,
department_name varchar(20),
tot_cred numeric(3,0) check (tot_cred >= 0),
primary key (ID),
foreign key (department_name) references department
on delete set null);
create table instructor
    (ID varchar(5) primary key,
name varchar(20) not null,
department_name varchar(20),
salary numeric(8,2));
```

and the following SQL query over the above given schemas:

```sql
select distinct i.name
from advisor a, instructor i
where i.ID = a.i_ID and i.salary > 75000;
```

i) State, in English, what the query above returns. (1 mark)

ii) Write the expression that represents the direct translation of the above query into relational algebra. (3 marks)

iii) Given the rules for logical rewriting discussed in this course unit, show the effect they would have when applied to the direct translation of the above query by drawing the operator tree that represents its canonical form. (9 marks)

b) Ignoring whether they are in canonical form or not, assume that the following two queries \( Q \) and \( Q' \) are running, in continuous, reactive mode, in a stream query processor (you can take all joins to be the default window-based joins, and \( a \) and \( b \) to be integers):

\[
Q = \pi_{R.c} (\sigma_{R.r > a} (R \bowtie S))
\]

and

\[
Q' = \pi_{T.t} (T \bowtie (\sigma_{R.r > b} (R \bowtie S)))
\]

i) Draw the operator graph that represents the two queries above running together. (9 marks)

ii) Now, assume that, in \( Q \), the selection predicate over the result of joining \( R \) and \( S \) is modified to \( R.r > b \) and draw the operator graph in this modified case. (4 marks)
c) Briefly describe one advantage and one drawback of using semijoins in distributed query evaluation? (2 marks)

d) Briefly explain what fundamental property of data streams imposes a modified semantics on joins and gives rise, e.g., to window-based joins. (2 marks)
2. a) Assume the following 5-way join for the subparts of this item:

\[ A \bowtie B \bowtie C \bowtie D \bowtie E \]

In what follows, let \( XY \) be a shorthand for \( |X \bowtie Y| \). Now, further assume that the estimates for the 2-way join output sizes are ordered as follows:

\[ DE < BD < BE < AE < AD < BC < CD < AC < CE < AB \]

i) Assume the greedy join ordering algorithm discussed in the course. State which, amongst the 2-way joins above, will be placed as the left-deepest join given the ordering of cardinalities above and briefly explain, in terms of the algorithmic logic used, why that one is the 2-way join selected.

(3 marks)

ii) State how many passes through the loop in the greedy join ordering algorithm will take place in this case and briefly explain why this many passes.

(3 marks)

b) Consider the notion of data appliances introduced in this course unit.

i) Briefly explain what characteristics of this kind of data management technology the word appliance is aiming to highlight.

(2 marks)

ii) Briefly explain, in terms of costs incurred in complex DBMS environments, what motivates the idea of data appliance.

(2 marks)

iii) Assume that you are working as a member of a team that has been assigned the task of assessing the benefits of switching the analytical workload of the organization to a data appliance product. At one point in the discussions, Yu, a member of the team, argues that “the response-time estimates for the queries are wrong as they assume no index is built to support them”. Rajiv, the team member who came up with the estimates replies to Yu as follows: “It’s an appliance. This should tell you there’s nothing wrong with the estimate.” At this point, the leader of the team asks you for your opinion. State whether you agree with Yu or Rajiv and briefly explain your reasons for doing so using the technical notions introduced in this course unit.

(4 marks)
c) Assume the following sensor network topology:

In the figure above, a vertex denotes a sensor node whose name is the vertex label (with B denoting the base station), an edge denotes that the nodes involved are within wireless communication range of one another (i.e., are one-hop neighbours), and the edge label denotes the quality of the communication link (the higher the label value, the better the quality). Using the basic flooding algorithm discussed in this course unit, show the step-by-step derivation of a routing tree for the above network topology. (8 marks)

d) The parts below are about data integration.

i) State two reasons why schema matching is hard. (2 marks)

ii) State one reason why using multiple matchers is considered advantageous. (1 mark)

iii) State what the correlation is (either positive or negative) between (life)time and (integration) quality in the case of pay-as-you-go data integration and in the case of mediator-based data integration and briefly explain why. (4 marks)

iv) A DBMS and a data integration system can be thought of as the first (i.e., bottommost) and second layers of a proposed four-layer architecture of which dataspaces are the last (i.e., topmost) layer. In this architecture, what kind of system constitutes the third layer, the one between data integration and dataspaces? (1 mark)