Name :
Ubitname:
Person # :

CSE 562 : Assignment 4
Due at the start of the Class on 03/31/2014

Please follow these instructions to answer this assignment:

1. **Print this assignment and use the empty space provided to answer all your questions.**
2. **Answer all the questions legibly. You may loose points if the answers aren't legible.**
3. **Answer all the questions in blue or black ink. Assignments answered with a pencil will not be entertained for re-grade requests.**
4. **All the necessary assumptions are given below, if you still have your own assumptions, please make sure that you state them clearly.**
5. **By submitting this assignment you vouch that this assignment was done individually.**

[Question 1 Total 60 points] Consider the following query:

SELECT *

FROM R, S, T

WHERE R.A = S.A and R.B = T.B AND T.C = S.C

AND R.D = 16 AND S.F = 17

1. **[5 points each for every correct expression]** Give six non-equivalent* relational algebraic tree plans that (I) contain no cartesian product, (II) have pushed selections down, (III) each join is associated with an atomic condition (not a conjunctive condition). [*Note : R ⋈ S and S ⋈ R are assumed to be equivalent queries/expressions for this problem. Or in other words, changing the join order does not make it two different queries/expressions]

2. **[10 points for selecting the optimum plan and 20 points for justifying your answer]** Find the optimum plan, i.e., an algebraic expression where selections and projections are annotated with execution primitives. The execution primitives are SCAN and INDEX for the selection operator and LOOPS, INDEX, MERGE, and HASH for join. Justify your answer

Assume the following schemas and statistics for the relations. The function "V" below gives the number of tuples. For example, V(D,R)=1000 means an index on attribute D in relation R has 1000 different values.

- R(A, B, D, E) has 1 million tuples, an index on D, V(D,R)=1000, V(A,R)=1000000, V(B,R)=1000000.

- S(A, C, F, G) has 1 million tuples, an index on F, V(F,S)=1000, V(A,S)=1000000, and V(C,S)=1000000.

- T(B, C, H, I) has 1 million tuples, an index on B, with V(B,T)=1000000, V(C,T)=1000000.

To estimate the size of the joins use the following information:

- For every tuple of R there is exactly one tuple of S with the same A value and one tuple of T with the same B value. Or in other words R.A and R.B are foreign keys in S at T respectively. Equivalently, the chances that a given S tuple joins a given R tuple are 1/1000000.

- For every tuple of S there is exactly one tuple of R with the same A value and one tuple of T with the same C value.

- For every tuple of T there is exactly one tuple of R with the same B value and one tuple of S with the same C value.

(Blank Space for answering Question 1)

(Blank Space for answering Question 1)

(Blank Space for answering Question 1)

[Question 2 Total 10 points] : Are these two expressions equivalent? Justify your answer.

$\sigma_{p \wedge t}(P \bowtie T) = (\sigma_p P) \bowtie (\sigma_t T)$

where p refers to attributes of P only and t refers to attributes of T only.

(Blank Space for answering Question 2)

[Question 3 Total 30 points] : Consider the following two definitions of the semijoin operator ⋉ which may or may not be equivalent.

- **Direct :** The semijoin ⋉ of relations R and S, written R⋉ S, is the bag of tuples t in R such that there is at least one tuple in S that agrees with t in all attributes that R and S have in common.

- **Indirect :** Let us call a(R) the list of attributes of R. Then, it is

$$R \ltimes S = \prod_{a(R)} (R \bowtie S)$$

where ⋈ stands for the natural join.

1. **[10 points]** Are the two definitions equivalent? Assume bag semantics for the algebra. If the answer is yes provide proof, showing that, given arbitrary R and S, if a tuple t appears k times in the result of R ⋉ S according to the direct definition, then the tuple t will appear k times in the result of R ⋉ S according to the indirect definition. If the answer is no, provide an example with an R table and an S table and show R ⋉ S for the direct and the indirect definition.

2. **[20 points]** Consider the indirect definition of ⋉. Assume that the schema of P is P(A, B, C, D) and the schema of T is T(C, E). Prove the following

$$\prod_{A,B} \sigma_{D=5 \wedge E=6} (P \bowtie T) = \prod_{A,B} [(\prod_{A,B,C,} \sigma_{D=5} P) \ltimes (\prod_{C,} \sigma_{E=6} T)]$$

(Blank Space for answering Question 3)

(Blank Space for answering Question 3)