1 Short Answer

a) What is a foreign key and what constraints does this key enforce (in SQL)?

b) In SQL++, list two commands that are part of the Data Definition Language

c) Give two transactions and a schedule that illustrates the phantom problem.

d) Suppose there is a table R(a,b) where B(R) = 1000, T(R) = 50000, V(R,a) = 300 and V(R,b) = 200. Values of R.b range from -20 to 80. What ist the minumum number of disc accesses it would take to retrieve all records in R where 30¡R.b¡60. Why? Assume R.b has an unclustered index.

e) Does the block nested loop join approach require indexes? Why or why not?

f) For which of the following does a semi-structured approach have the largest benefit over RDBMS? Many-to-one, many-to-many, one-to-many or one-to-one? Why?

g) Suppose we are designing two databases to handle transactions under analytical and transactional data usage. Under which do we expect deadlock to be more difficult to prevent? Why?

h) Is it possible for a schedule to be serial, but not conflict-serializable? Explain.

i) Will a Map/Reduce job take more time if a straggler is a Map task or a Reduce task? Explain.

j) What does it mean if a schedule is unrecoverable? Give an example.

k) Suppose we have the following schema:

```
Person(pid,name,phonenum,city)
Organization(oid,name,address)
Membership(pid,oid,donationamt,startyear,birthday)
```

Which attributes in which tables would we expect to be functionally dependent? List all that apply.

l) From the schema above, which of these tables are entities and which are relationships? If a table is a relationship, identify if it is many-to-many, one-to-one or many-to-one.

2 SQL

For this question, use the following schema:

`Company(cid,country,name)`

`Product(pid,name,description)`

`Store(sid,cid,location)`

`Sells(sid,pid,price)`

a) Provide the E/R diagram for this table. Assume all relations are many-to-many.

b) Write a SQL query to find the names of all companies who sell a product that costs more than $50. You may assume price is an INT.

c) Draw two logical plans that accomplish this goal.

d) Name four pieces of information you would need to find the cheapest physical plan, in terms of disc accesses.

3 Datalog

Suppose we have the following predicates in a datalog database:

`Person(pid,name)`

`Parent(parentid,childid)`

a) Are these predicates extensional or intenstional?

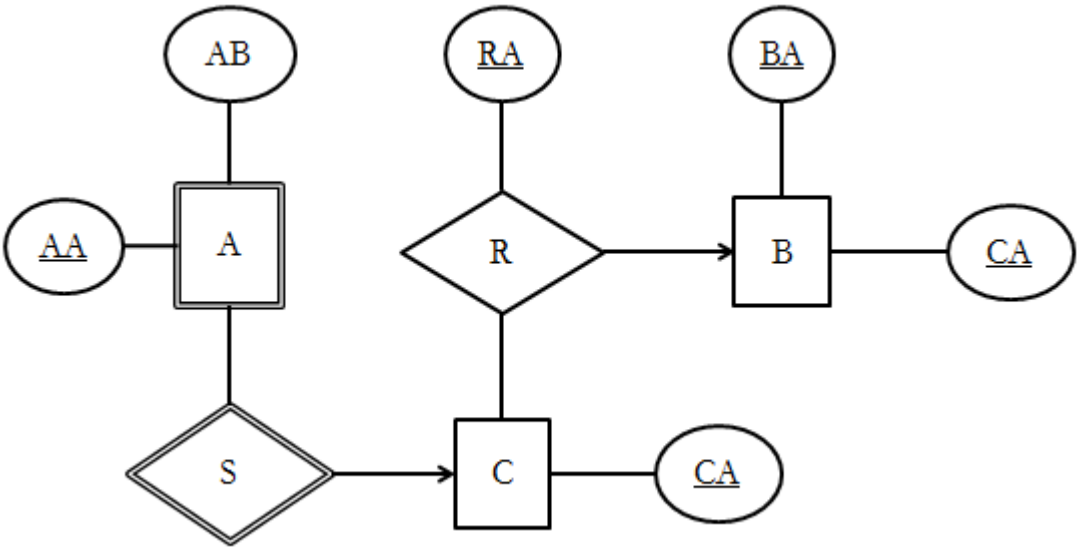
b) Define a predicate SC(x,y) which is true if x and y are second-cousins, i.e. they share a great-grand parent. Use intermediate predicates as necessary.

c) Can this be solved using SQL? Why or why not?

d) Define a predicate cousins(x,y,n) where x and y are people and n is the level of their relationships. Siblings should have n = 0, cousins have n = 1, second cousins n=2 and so forth. For any pair of people (x,y) only one tuple cousins(x,y,n) should exist, even though first cousins must also share great-grandparents.

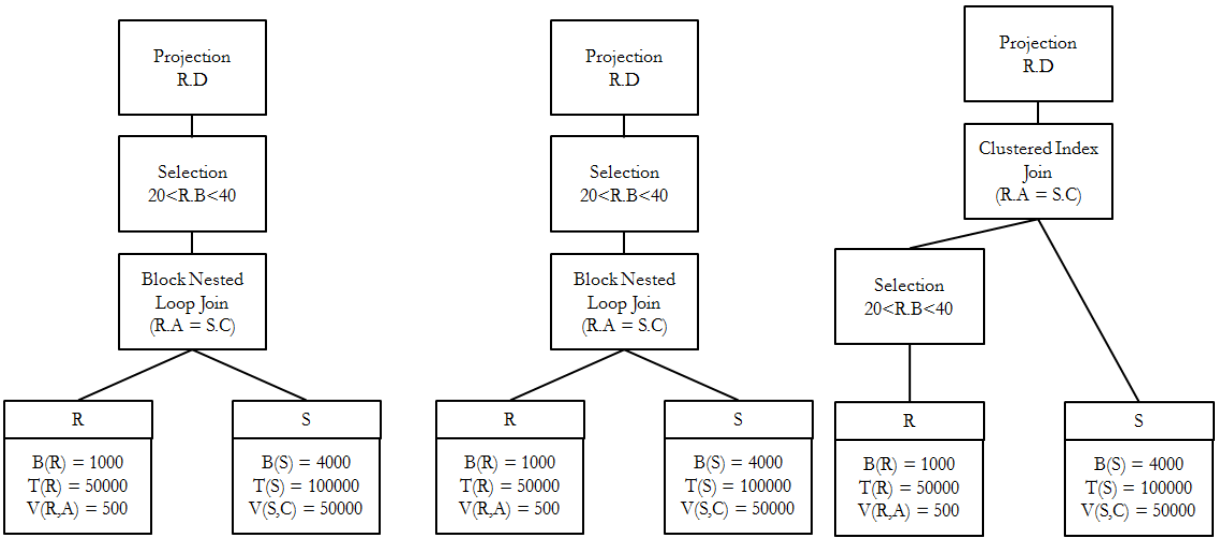
4 E/R Diagrams

Produce a schema for the following E/R Diagram. Indicate if there are any foreign keys or if anything must be non-null.



5 Cost Estimation

Give the cost for the following physical plans. Assume the following clustered indexes (R.A and S.C) and no unclustered indexes:



6 Transactions

Suppose a scheduler is trying to schedule the following transactions.

a) Transaction 1: $r(A), w(B), r(A)$

b) Transaction 2: $r(B), w(B), w(A)$

c) Transaction 3: $w(A), w(B)$

For each schedule, indicate whether it is:
a.) a valid schedule given the transactions
b.) conflict-serializable,
c.) serializable or
d.) serial
Keep in mind, each schedule may be multiple.

a) $r_1(A), w_3(A), r_2(B), w_2(B), w_1(B), w_2(A), r_1(A), w_3(B)$

b) $r_2(B), r_1(A), w_1(B), w_2(B), w_2(A), r_1(A), w_3(A), w_3(B)$

c) $r_1(A), r_2(B), w_3(A), w_1(B), w_2(B), w_3(B), r_1(A), w_2(A)$

d) Which of the above schedules results in a dirty or inconsistent read? Which transactions are affected?