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

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

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) In our semi-structured data, what is the difference between a dataverse and a dataset?

h) If a database has queries that are analytical, rather than transactional, should we prefer duplication or partitioning?

i) Suppose we have the schema `R(a,b) S(b,c)`. When are R.a and S.c `NULL` in a `FULL OUTER JOIN` of R and S?

j) In NoSQL (semi-structured) data bases, what is heterogeneity and why does the problem arise?

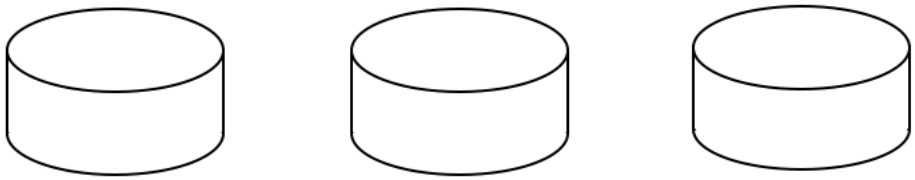
k) Why would a DBMS prefer uncorrelated subqueries?

l) Give an example of a safe, non-linear recursive datalog rule.

m) We have the schema R(a, b), S(x, y). Given the following query, draw a parallel query plan that reduces the amount of **data** sent over the interconnection network. Assume all attributes are INTs and B(R) = B(S).

```
SELECT R.a
FROM R, S
WHERE R.b = S.x AND
      S.y = 100
```

R has been hash partitioned on R.b and S is block partitioned



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

b) Draw two logical plans that accomplish this goal.

c) 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.

4 Cost Estimation

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