CSE 444 Practice Problems

DBMS Architecture

1. **Data Independence**

   (a) What is physical data independence?

   (b) What properties of the relational model facilitate physical data independence?

(c) What is logical data independence?

(d) How can one provide a high level of logical data independence with the relational model?

2. **High-Level DBMS Architecture**

   You should know the key components of a relational DBMS. Please see the lecture notes for an overview.

   You should also be able to discuss what you implemented in the labs. For example, we could ask you for a high-level description of your buffer manager.

3. **Data Storage and Indexing**

   You should be able to show what happens when one adds/removes data to/from a B+ Tree. Please see the web quizzes for some good examples.

4. **Data Storage and Indexing**

Suppose we have a relation `R(a,b,c,d,e)` and there are at least 1000 distinct values for each of the attributes.

Consider each of the following query workloads, independently of each other. If it is possible to speed it up significantly by adding up to **two** additional indexes to relation R, specify for each index (1) which attribute or set of attributes form the search key of the index, (2) if the index should be clustered or unclustered, (3) if the index should be a hash-based index or a B+-tree. You may add at most two new indexes. If adding a new index would not make a significant difference, you should say so. Give a brief justification for your answers.

(a) 100,000 queries have the form: select * from R where $b <?$
10,000 queries have the form: select * from R where $c =?$

(b) 100,000 queries have the form: select * from R where $b <?$ and $c =?$
10,000 queries have the form: select * from R where $d =?$
1,000 queries have the form: select * from R where $a =?$

(c) 100,000 queries have the form: select a, c from R where $b <$?
10,000 queries have the form: select * from R where $d <$?

5. **Relational Algebra and Query Processing**

Consider three tables R(a,b,c), S(d,e,f), and T(g,h,i).

(a) Consider the following SQL query:

```
SELECT   R.b
FROM     R, S, T
WHERE    R.a = S.d
         AND S.e = T.g
         AND T.h > 21
         AND S.f < 50
GROUP BY R.b
HAVING count(*) > 2
```

For each of the following relational algebra expressions, indicate if it is a correct translation of the above query or not.

i. $\pi_{R.b}(\sigma_{\text{TOTAL}>2}(\gamma_{R.b,count(*)\rightarrow\text{TOTAL}}(\sigma_{T.h>21 \text{ AND } S.f<50}(R \bowtie_{R.a=S.d} (S \bowtie_{S.e=T.g} T)))))$

CORRECT INCORRECT

ii. $\pi_{R.b}(\sigma_{\text{TOTAL}>2}(\gamma_{R.b,count(*)\rightarrow\text{TOTAL}}(R \bowtie_{R.a=S.d} ((\sigma_{S.f<50}(S)) \bowtie_{S.e=T.g} (\sigma_{T.h>21}(T))))))$

CORRECT INCORRECT

iii. $\pi_{R.b}(\sigma_{\text{TOTAL}>2}(\sigma_{T.h>21 \text{ AND } S.f<50}(\gamma_{R.b,count(*)\rightarrow\text{TOTAL}}(R \bowtie_{R.a=S.d} (S \bowtie_{S.e=T.g} T)))))$

CORRECT INCORRECT

(b) For the following SQL query, show an equivalent relational algebra expression. You can give the expression in the same format as we used above or you can draw it in the form of an expression tree or logical query plan.

```
SELECT R.b
FROM R, S
WHERE R.a = S.d
AND R.b NOT IN (SELECT R2.b FROM R as R2, T WHERE R2.b = T.g)
```

(c) A user just connected to a database server and submitted a SQL query in the form of a string. Give **four** important steps involved in evaluating that SQL query **and the order** in which they are performed. You only need to name the steps. No need to explain them.

(d) What is the difference between a logical and a physical query plan?

6. **Relational Algebra and Query Processing**

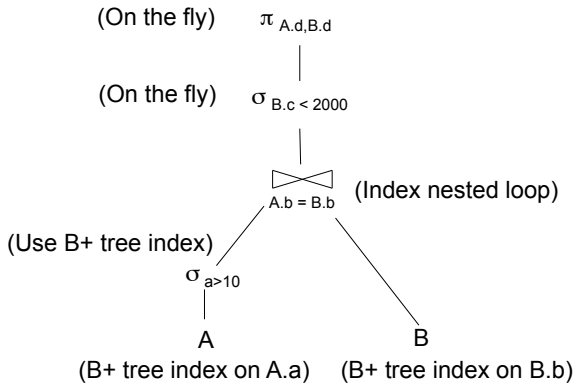
Consider four tables R(a,b,c), S(d,e,f), T(g,h), U(i,j,k).

(a) Consider the following SQL query:

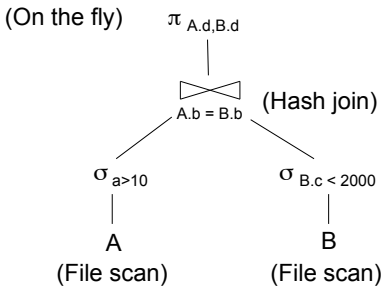
```
SELECT   R.b, avg(U.k) as avg
FROM     R, S, T, U
WHERE    R.a = S.d
         AND S.e = T.g
         AND T.h = U.i
         AND U.j = 5
         AND (R.c + S.f) < 10
GROUP BY R.b
```

Draw a *logical* plan for the query. You may chose any plan as long as it is correct (i.e. no need to worry about efficiency).

(b) Consider the following two physical query plans. Give **two** reasons why plan B may be faster than plan A. **Explain** each reason.



Plan A



Plan B

7. **Operator Algorithms**

Relation R has 90 pages. Relation S has 80 pages. Explain how a DBMS could efficiently join these two relations given that only 11 pages can fit in main memory at a time. Your explanation should be **detailed**: specify how many pages are allocated in memory and what they are used for; specify what exactly is written to disk and when.

(a) Present a solution that uses a hash-based algorithm.

(b) Present a solution that uses a sort-based algorithm.