Homework #6
due Thursday, December 2 , 2010 at 2 pm

Database Systems, CSCI-4380-01

Each student must work on this homework alone.

1 Homework Description

Question 1. Estimate the cost of following query plans. For each plan, the implementation of each operation, the amount of buffer allocated to each operation and the size of the input relations are given. Assume $R(A, B, C, D, E)$, $S(A, F, G, H)$, $B(R) = 5,000$ and $B(S) = 10,000$, $Tuples(R) = 100,000$ and $Tuples(S) = 10,000,000$.

(a) $R \bowtie S$ using block-nested loop join for $\bowtie$ using $M = 101$.

(b) $(R \bowtie S) \bowtie T$ using block-nested loop join between R and S using $M = 101$, and the join with T using block-nested loop join also using $M = 101$. Assume $B(T) = 5,000$ and $B(R \bowtie S) = 30,000$.

(c) $\delta(R)$ using sort based duplicate removal where $M = 200$ buffer pages are allocated for sorting and $M = 1$ buffer page is allocated for duplicate removal. (Note that you can do duplicate removal combined with sorting as well, this question is simplifying the logic for you).

(d) $R \cup S$ using sort based union where $M = 200$.

(e) $\sigma_{A=5 \text{ and } B>10 \text{ and } B<100} R$ using sequential scan of R.

(f) $\sigma_{A=5 \text{ and } B>10 \text{ and } B<100} R$ using index I1 on $R(B, A)$ of height 2 where each node (except root) contains about 200 entries and there are $2,000$ tuples for $A = 5$ and $50,000$ tuples for $B > 10$ and $B < 100$ and 500 tuples for the whole selection condition.

Question 2. Estimate the size of the following relations in terms of number of tuples and number of disk pages (or memory blocks). Assume each attribute takes 40 bytes and each disk block is 4K bytes (assume 4,000 bytes for simplicity). For example, given a relation with 5 attributes, each tuple is about 200 bytes and a disk page takes about 4000/200=20 tuples.

Suppose $R(A, B, C, D)$ has 10,000 tuples, and $VALUES(R.A) = 10,000$, $VALUES(R.B) = 2,000$, $VALUES(R.C) = 5,000$ and $VALUES(R.D) = 200$. Suppose $S(E, F, G)$ has 1,000,000 tuples, $VALUES(S.E) = 10,000$, $VALUES(S.F) = 1,000$ and $VALUES(S, G) = 5,000$.

(a) $R.A = 5$

(b) $R.B = 10$ and $20 \leq R.A$ and $R.A \leq 100$

(c) $20 \leq R.A$ and $R.C = 100$ and $50 \leq R.D$

(d) $S.F = 10$

(e) $S.E = 1$ or $S.G = 2$

(f) $\pi_{B,F}(R \bowtie_{A=E} S)$

Question 3. For each query above in Question 2, discuss the best index that would impact the query. Write one sentence to explain why this is a good index to create.

Also, assuming that each query is executed with the same frequency, what is the most important single index to create?

Question 4 (Bonus). Improve the running time of the following query (Q3 from Homework 3):

```
SELECT
   DISTINCT
   u.id AS uid
   , u.userName
   , u.email
FROM
   friends f
   , friends f2
   , users u
WHERE
   (f.user_id = 10
   and f.friend_id = f2.user_id
   and u.id = f2.friend_id)
   or
   (f.user_id = 10
   and f.friend_id = f2.friend_id
   and u.id = f2.user_id)
   or
   (f.friend_id = 10
   and f.user_id = f2.user_id
   and u.id = f2.friend_id)
   or
   (f.friend_id = 10
   and f.user_id = f2.friend_id
   and u.id = f2.user_id)
EXCEPT
SELECT
   DISTINCT
   u.id AS uid
   , u.userName
```

```
   , u.email
FROM
   friends f
   , users u
WHERE
   (f.user_id = 10
    and u.id = f.friend_id)
   or
   (f.friend_id = 10
   and u.id = f.user_id)
ORDER BY uid ASC;
```

First, find the running time of this query running

```
EXPLAIN ANALYZE query ;
```

Then, you can make this query faster by either rewriting it or by creating one or at most two indices, or a combination of both. Assume only the keys for the primary keys exist and use the database created in your own user from the previous homeworks. In CSC4380, this query takes 2683.541 ms. Document what you have done and how it helped speed up the query. Show the query plan before and after your index creation to prove the database is actually using your index if you created indices.

2 Deliverables

Turn in a single text file (.txt or PDF) containing all your answers.