

## CMPT506 – Advanced Database System

### Homework 2

**Due by 4pm Sunday 4/12/2016 – Submit your softcopy to blackboard and your hardcopy at the start of the class.**

1. [10 pts] Consider a file with 10,000 blocks and three available buffer blocks. Assume that the external sort-merge algorithm is used to sort the file.
  - a. [3 pts] Calculate the initial number of runs produced in the first pass.
  - b. [3 pts] How many passes will it take to sort the file completely?
  - c. [4 pts] How many buffer blocks are required to sort the file completely in two passes?
2. [20 pts] Consider the following SQL queries on Online Book Store Application.
  - a. [3 pts] Transform these SQL queries into relational algebra expressions.
  - b. [3 pts] Draw the initial query trees for each of these expressions.
  - c. [4 pts] Derive their optimized query trees after applying heuristics on them. You can just show the last optimal query tree but briefly explain the heuristics applied to arrive to the optimal solution.

```
2.1) SELECT ISBN, Book_title, Category, Price
FROM BOOK B, PUBLISHER P
WHERE P.P_ID = B.P_ID AND Pname='Bright Publications'
AND Price>30;
```

```
2.2) SELECT ISBN, Book_title, Year, Page_count, Price
FROM BOOK B, AUTHOR A, AUTHOR_BOOK AB
WHERE B.ISBN = AB.ISBN AND AB.A_ID = A.A_ID
AND Aname = 'Charles Smith';
```

#### 3. [15 pts] Query Execution

Consider two relations,  $R(A,B,C)$  and  $S(B,D,E)$ .  $R$  has 1140 tuples. Each block can hold 15 tuples per block. If it is not specified, assume we have a large enough buffer to perform the operation.

Assume that  $S$  is the smaller relation and  $R$  is the larger relation.

- a. Assume we have a memory buffer that can hold 25 blocks ( $M=25$ ), and the cost of joining  $R$  and  $S$  using a block nested-loop join is 228. How many blocks are used to store the tuples in  $S$ ?
- b. What is the cost of joining  $R$  and  $S$  using a simple sort-merge join, using the  $B(S)$  you found in question 1?
- c. What is the cost of joining  $R$  and  $S$  using a hash-based join, using the  $B(S)$  you found in question 1?

#### 4. [15 pts] Cost Estimation

Note that  $T(R)$  is the number of tuples in relation  $R$ , and  $V(R,A)$  is the number of distinct values of the attribute  $A$  in relation  $R$ . Consider a database with three relations:  $R(A,B,C)$ ,  $S(B,D)$ ,  $Z(C,E)$ .

We have the following statistics:

$T(R)=500$     $V(R,A)=10$     $V(R,B)=100$     $V(R,C)=50$

$T(S)=5000$     $V(S,B)=400$     $V(S,D)=100$

$T(Z)=200$     $V(Z,C)=100$     $V(Z,E)=25$

For each query, estimate the number of tuples returned. You must also write the formula you use to calculate the number of tuples (in terms of  $T$ 's and  $V$ 's):

- a. Select Z.E Where Z.E = 10
- b. Select R.B, Z.E From R,Z Where R.C = Z.C
- c. Select R.A, S.D From R,S Where R.B = S.B and R.A = 3

### 5. [15 pts] Query Execution

Consider the following two relations:

Customer(cid, name, phone, etc.)

Order(oid, orderDate, cid, etc.)

Customer (C) has 10,000 tuples, with 25 tuples fitting on a block.

Order (O) has 5,000 tuples, with 50 tuples fitting on a block.

There is no index on any attribute of the relations. You may assume the table blocks are contiguous. In your answer always provide the general formula.

- (a) Suppose the memory buffer has 101 blocks, compute the cost of using a block-nested loop join to join the above two relations.
- (b) Suppose we wanted to join the two relations using a block-nested loop join and limit the cost to 900. What is the smallest value M can be?
- (c) What is the cost of joining Customer and Order using a hash-based join?

### 6. [25 pts] Join Algorithms

Consider two relations R and S. The tuples in each relation are listed in the following table.

R S

7 8

2 4

9 2

8 1

3 3

9 2

1 7

3 3

6

We want to do the natural join of R and S based on different join algorithms. For each algorithm listed as following, give the join results in the order that they would be output by the corresponding join algorithm.

- (a) [7.5 pts] The one tuple at a time nested-loop join algorithm. Suppose R is used for the outer loop and S is used for the inner loop.
- (b) [7.5 pts] The merge-sort join algorithm.
- (c) [10pts] The hash join algorithm. We assume only two hash buckets exist, numbered 0 and 1, respectively. The hash function hashes even values to bucket 0 and odd values to bucket 1. Moreover, we assume that in the join phase of the hash join algorithm, **R** is used as the “load” relation and **S** is used as the “stream” relation (i.e., First load R bucket 0, and for each entry scan through S bucket 0 to find matches. Then do the same for the bucket 1). Furthermore, we assume the content of a bucket are read in the same order as they were written.