# CMPT 454 (Spring 2010)
# Assignment 3: Query Processing and Query Optimization

**Due: Wednesday, 2010-03-24, 11:29AM (at the beginning of class)**

**Instructions on Assignment Submission:** Hard copy only!

- Option 1: drop off your assignments in the assignment box (with label **CMPT 454**) in CSIL.
- Option 2: bring your assignments to the class on the due day, and the instructor will collect the assignments **at the beginning of class**.

Please write legibly or typeset your answers using your favorite word processor. Late assignments will not be accepted unless there is a documented medical reason.

**Problem** 1.

**(20 points) [Parse Tree]**

Consider the simple SQL grammar discussed in the class, provide corresponding parse trees for the following two queries on relations $R(a, b)$ and $S(b, c)$.

(1) SELECT $a, c$ FROM $R, S$ WHERE $R.b = S.b$;

(2) SELECT $b$ FROM $R$ WHERE $a$ IN (SELECT $a$ FROM $R, S$ WHERE $R.b = S.b$ AND $S.c$ LIKE $'\%2010'$);

**Appendix:** The simple SQL grammar we used in the class is listed as following:

| | | |
|---|---|---|
| <Query> | ::= | SELECT <SelList> FROM <FromList> WHERE <Condition> |
| <SelList> | ::= | <Attribute> , <SelList> |
| <SelList> | ::= | <Attribute> |
| <FromList> | ::= | <Relation> , <FromList> |
| <FromList> | ::= | <Relation> |
| <Condition> | ::= | <Condition> AND <Condition> |
| <Condition> | ::= | <Attribute> IN (<Query>) |
| <Condition> | ::= | <Attribute> = <Attribute> |
| <Condition> | ::= | <Attribute> LIKE <Pattern> |

**Problem** 2.

**(20 points) [Relational Algebra]**

We have two relations $R(a, b, c)$ and $S(c, d, e, f)$. Consider the following SQL query:

SELECT $a, b, e, f$
FROM $R, S$
WHERE $R.a < 35$ AND $R.c + S.d \geq 59$ AND $R.c = S.c$;

(1) Write down the corresponding relational algebra expression (use only selection, projection and natural join operators).

(2) Transform the relational algebra expression from the previous sub-question into another equivalent one such that the selections and projections are performed as early as possible.

**Problem** 3.

**(30 points) [Cost Estimation]**

Consider the following 4 relations $W$, $X$, $Y$, and $Z$. The statistics of tuples and distinct attribute values in each relation are listed in the following table.

| $W(a,b)$ | $X(b,c)$ | $Y(c,d)$ | $Z(d,e)$ |
|---|---|---|---|
| $T(W) = 100$ | $T(X) = 200$ | $T(Y) = 300$ | $T(Z) = 400$ |
| $V(W,a) = 20$ | $V(X,b) = 50$ | $V(Y,c) = 50$ | $V(Z,d) = 40$ |
| $V(W,b) = 60$ | $V(X,c) = 100$ | $V(Y,d) = 50$ | $V(Z,e) = 100$ |

Estimate the sizes of relations that are the results of the following expressions.

(1) $W \times Y$;

(2) $\sigma_{c=20}(Y)$;

(3) $\sigma_{d>10}(Z)$; (use Solution 2 for the inequality)

(4) $\sigma_{(a=1)\wedge(b>2)}(W)$; (use Solution 2 for the inequality)

(5) $\sigma_{c=20}(Y) \bowtie Z$;

(6) $W \bowtie X \bowtie Y \bowtie Z$;

**Problem** 4.

**(30 points) [Join Algorithm]**

Consider two unary 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.

(1) The naïve (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;

(2) The merge-sort join algorithm;

(3) 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 step Phase II of the hash join algorithm, $R$ is used as the

"load" relation and $S$ is used as the "stream" relation. Furthermore, we assume that bucket 0 is read first and the content of a bucket are read in the same order as they were written.