Database Systems, CSCI 4380-01
Homework # 7
Due Monday May 2, 2016 at 11:59:59 PM

Introduction.

In this homework, you are allowed to work in groups of at most 2. If you do, please write the name of your group members in your solution file. Turn in a single submission (PDF or text) per group.

This homework is on the underlying mechanics of query optimization. Read carefully in each question about assumptions you can make.

`TUPLES(R)=100,000, TUPLES(S)=2,000,000, TUPLES(T)=500,000`

`PAGES(R)=5,000, PAGES(S)=10,000, PAGES(T)=3,000`

Attribute	N_DISTINCT
R.A	10,000
R.B	5,000
R.C	20
S.C	500
S.D	2,000
S.E	20
T.F	500,000
T.G	50,000
T.H	200

Question 1. Given the above statistics, find the estimated cardinality (number of tuples) of the following queries. (use 1/3 for range selectivity).

```
Q1. SELECT * FROM R WHERE R.A=10 AND R.B<4;
Q2. SELECT * FROM R,S WHERE R.B=S.C AND (S.C<4 OR S.D=5)
Q3. SELECT * FROM S WHERE S.C=S.D AND S.E=3
```

Question 2. Estimate the cost of the following join orders (plans 1-3) assuming the following:

- All joins are block-nested loop joins with M=201. Inner and outer joins are fixed as shown.

- All joins are pipelined, i.e. the output of the lower join results is used as the input for the outer relation of the next join.

- Assume that after a join, you can store about 200 tuples per memory block (page).

To accomplish this, you must find the cardinality and size of the lower join in terms of number of pages to see how many times the last relation need to be read.

```
Q. SELECT * FROM R,S,T WHERE R.B=S.C AND S.D=T.F AND R.C=T.H
```

```
Plan 1: (R join S) join T
Plan 2: (R join T) join S
Plan 3: (S join T) join R
```

**Question 3.** What is the cost of sorting S given the following memory: (a) M=200, (b) M=50?
Question 4. What is the cost of sort merge join as given in Plan 1 below for the following query, assuming you have M=201 for each join:

```
Q. SELECT * FROM R,S,T WHERE R.B=S.C AND S.D=T.F AND R.C=T.H
```

```
Plan 1: (R join S) join T
```

Note that you can combine the merge and sort steps if all the necessary blocks fit in memory. You can disregard duplicate values (whether they will all fit in memory or not). Same in Question 2, assume that after a join, you can store about 200 tuples per memory block (page).
You will sort R on R.B and S on S.D. Once you have completed merge sort for (R join S), you need to sort the results again by S.D this time (the more selective attribute). You will also sort T on T.F to match.
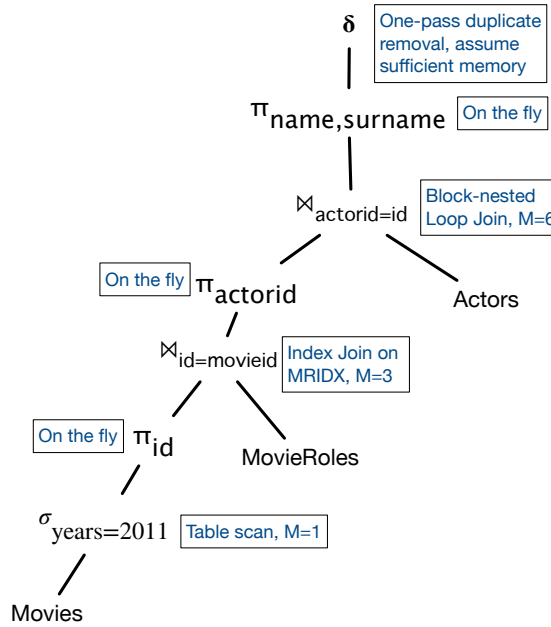
Question 5. Finally, estimate the full cost of the above query plan and show details of your work. This question is based on the IMDB database we used in the homeworks. All relevant details are given below. Use the estimates based on the values given below, not the actual values.

- A disk page is 1K bytes. An integer is 4 bytes and a tuple address is 12 bytes.

- TUPLES(Movies)=4,708, PAGES(Movies)=70, N_DISTINCT(Movies.years)=282

- TUPLES(MovieRoles)=265,107, PAGES(MovieRoles)=2,044,
  N_DISTINCT(MovieRoles.actorid)=149,374, N_DISTINCT(MovieRoles.movieid)=4,340

- Index MRIDX on MovieRoles(movieid,actorid), DEPTH=2 (root, internal and leaf nodes).
  Assume that the root is already in memory.

- TUPLES(Actors)=149,374, PAGES(Actors)=2,044