03.03.2025 Week 3 Exercises: Query Optimization

Exercise 1

Consider the relations R(A, B), S(B, C), T(A, B, C).

1. Find a counter-example to show that:
   $\sigma_{A=\text{'a'}}((R \bowtie S) \cup T) \neq (\sigma_{A=\text{'a'}}(R) \bowtie S) \cup T$
2. Give the left hand side and the right hand side of the above expressions for your counter-example.
3. Modify the right hand side of the expression to make it equivalent to the left hand side.

Exercise 2

Consider the relation with schema EMP(ssn, name, age, jobcode) and the following statistics:

- There are n = 10,000 records in the file.
- There are 40 distinct values for *age* ranging from 21 to 60.
- There are 10 distinct values for *jobcode*.

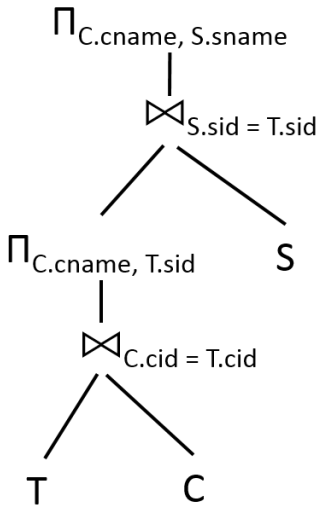
Assume that all attributes (and index references) have the same size and you can fit 10 records per data page. There is an unclustered B-tree index on *age* and a clustered B-tree index on *jobcode*, both of height 2. Describe the most efficient way to execute each of the following queries and estimate the corresponding I/O cost:

1. `SELECT count(*) from EMP where jobcode = 'programmer'`
2. `SELECT count(*) from EMP where age > 40`
3. `SELECT count(*) from EMP where jobcode = 'programmer' and age > 40`

You can make any assumptions that you find necessary about the data distributions.

Exercise 3

Consider the following schema:
Students S(sid, sname)
Taken T(sid, cid, grade, description)
Courses C(cid, cname)
Think of one way to improve the following query plan by applying a heuristic optimization rule.



Exercise 4

Consider the following schema:
Students S(sid, sname)
Taken T(sid, cid, grade)
Courses C(cid, cname)

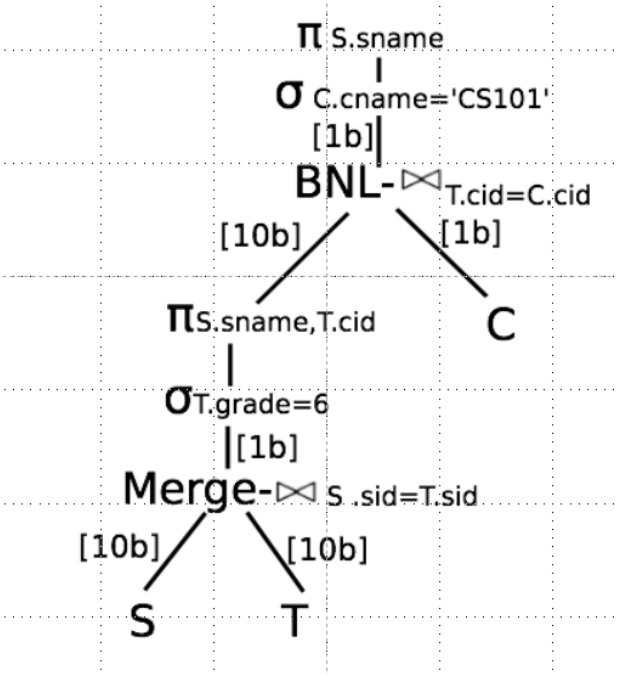
First Part

Consider first the query:

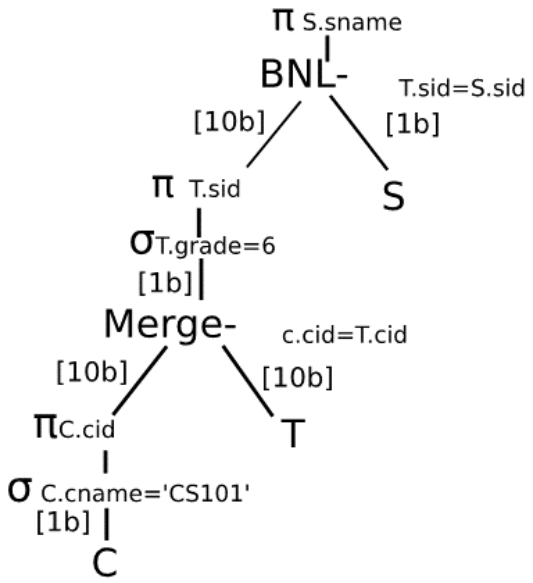
```
SELECT S.sname
FROM S, T, C
WHERE S.sid = T.sid and T.cid = C.cid
and  C.cname = 'CS101' and T.grade =  '6'
```

Estimate the time it takes to execute the above query for the two query plans show below. Consider only Block Nested Loop Join and Merge Join as the physical implementation of the join operators. The ids (*sid, cid*) are 4 bytes long, the names (*sname and cname*) are 60 bytes long and the *grade* is 4 bytes long. There are 33 buffer pages and the page size (excluding the header) is 1024 bytes. The Student relation contains 64,000 tuples, the Taken relation contains 128,000 tuples and the Courses relation contains 40 tuples. The relations S and T are ordered according to the sid column. There is a total of 1,000 grades of 6, 20 grades of 6 for the course CS101, and a total of 200 students taking the course CS101. It takes 0.1ms to transfer a page from disk to main memory. There are no indexes. Assume that the relations are stored on SSDs *i.e.*, they have no seek costs.

Provide the steps of your estimation. The number of buffers per operator are shown in brackets.



(a) The 1st plan



(b) The 2nd plan

Second Part

Consider now the following query:

```
SELECT *
FROM S, T, C
WHERE S.sid = T.sid and T.cid = C.cid
```

Consider only Hash Join and Block Nested Loop Join as the physical implementation of the join operators. Assume the following join costs:

- HJ of S and T = 30
- BNL of S and T = 60
- HJ of C and T = 40
- BNL of C and T = 50
- HJ of the result of S⋈T and C = 40
- BNL of the result of S⋈T and C = 30
- HJ of the result of C⋈T and S = 50
- BNL of the result of C⋈T and S = 40

Do not consider plans that contain the join between C and S, as they will result in a Cartesian product. Depict how the System R query optimizer constructs *iteratively* the best query plan.