COMP 3311: Database Management Systems

Lecture 19 Exercises
Query Optimization

Exercise 1: Given relation R(A, B, C)

Assume: R contains 10,000 tuples in 1,000 pages.
A has 50 distinct values in the range 1…50.
B has 100 distinct values in the range 0…100.

Estimate the size (number of tuples) of each of the following operations *assuming uniform distribution and attribute independence*.

a) $\sigma_{A=10}R$

b) $\sigma_{A=10 \wedge 20<B}R$

c) $\sigma_{C=1}R$

d) $\sigma_{C=10 \wedge A=10}R$

e) $\sigma_{C=10 \wedge A=10 \wedge 20<B}R$

Exercise 2: Consider the relation Sailor(sailorId, sName, rating, age) and the query:

```
select sName
from Sailor
where rating=7
      and age=40;
```

n_{Sailor} = 10,000 B_{Sailor} = 1,000 pages $bf_{Sailor} = \lceil 10{,}000 / 1{,}000 \rceil = 10$

V(rating, Sailor) = 10 (10 distinct rating values)

V(age, Sailor) = 100 (100 distinct age values)

Estimate the cost of the following alternative plans to process the query assuming uniform distribution and attribute independence. *Ignore the cost of searching any indexes.*

a) file scan

b) binary search
cost to search on rating

cost to search on age

c) single B^+-tree index (on either attribute)
index on rating

index on age

d) multiple B^+-tree indexes (on both rating and age)

Name: (1) _____/_____ **Student#:** (1) _____ **Date:** _____
Family/Given (PRINT) Given/First (PRINT)

Name: (2) _____/_____ **Student#:** (2) _____
Family/Given (PRINT) Given/First (PRINT)

NOTE: You are highly encouraged to do this exercise with a partner.

COMP 3311: Database Management Systems

Lecture 19 Exercises
Query Optimization

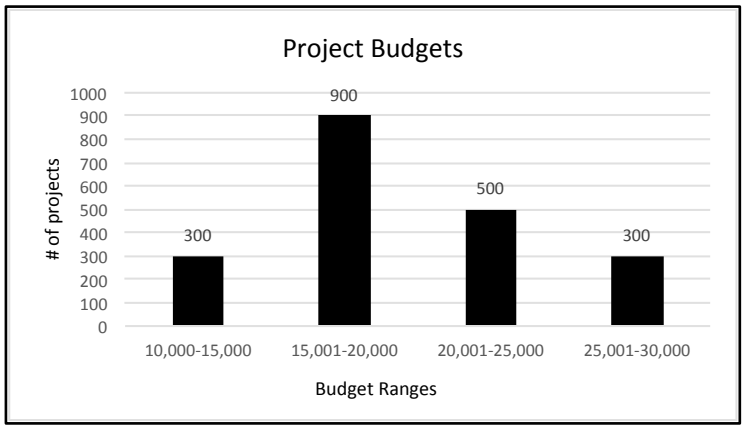
Exercise 3: Employee(empId: 4 bytes, name: 35 bytes, title: 2 bytes, salary: 5 bytes, deptId: 4 bytes)
Department(deptId: 4 bytes, projectId: 4 bytes, name: 25 bytes, location: 7 bytes)
Project(projectId: 4 bytes, title: 20 bytes, budget: 6 bytes, report: 970 bytes)

Employee:	50 bytes/tuple;	20,000 tuples	Page size: 4,000 bytes
Department:	40 bytes/tuple;	500 tuples	Memory buffer pages: 12
Project:	1,000 bytes/tuple;	2,000 tuples	

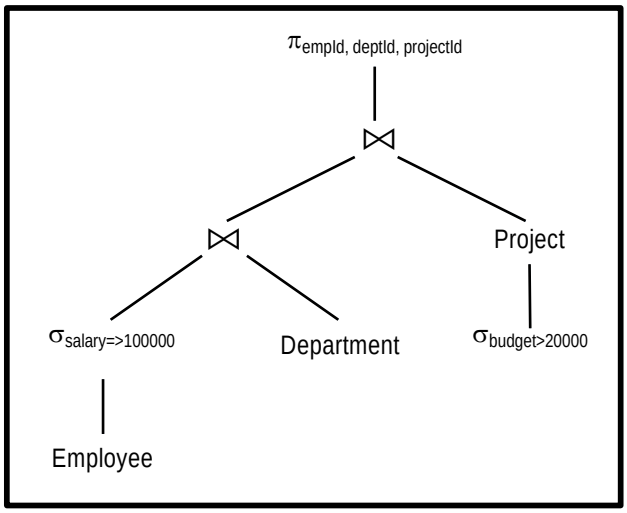
Employee salaries: uniformly distributed in the range 10,000 to 110,000.

Project budgets: distributed in the range 10,000 to 30,000 according to the histogram below.

- There is a clustering B^+-tree index with 3 levels on salary for Employee.
- There is a hash index on deptId for Department, which is ordered on deptId.
- There is a hash index on projectId for Project, which is ordered on projectId.




```
select E.empId, D.deptId, P.projectId
from Employee E, Department D, Project P
where E.deptId=D.DeptId
      and D.projectId=P.projectId
      and salary=>100000
      and budget>20000;
```



a) Use the relational algebra tree to estimate the output size of the query in tuples.

b) Evaluate the query using the relational algebra tree and the steps given below. The goal is to minimize the average number of page I/Os. Where possible, use pipelining rather than materialization (i.e., keep intermediate results in memory where possible). Assume the file organizations and indexes described above. For each step, give the strategy used and the average case page I/O cost. Give the total page I/O cost to process the query and the estimated result output size in pages.

Step 1: $\sigma_{salary=>100000}$Employee $\Longrightarrow$ result A

Strategy:
Cost:

Step 2: result A $\bowtie$ Department $\Longrightarrow$ result B

Strategy:
Cost:

Step 3: $\sigma_{budget>20000}$Project $\Longrightarrow$ result C

Strategy:
Cost:

Step 4: result B $\bowtie$ result C
Strategy:
Cost:

Total page I/O cost: Output result size in pages:

You must upload this completed exercise sheet to Canvas by **11:00 p.m. today.**