Week 11 Exercises: Query Optimization

Assume that performing a random probe on a B+ tree index has a cost in number of pages equal to the depth of the tree and for the same on a Hash index, an average cost of 1.2 pages.

Exercise 11.1 Consider a relation with this schema:

Employees(*eid:* **integer**, *ename:* **string**, *sal:* **integer**, *title:* **string**, *age:* **integer**)

Suppose that the following indexes, all using Alternative (2) for data entries, exist: a hash index on *eid*, a B+ tree index on *sal* with fanout 100, a hash index on *age*, and a clustered B+ tree index on *<age, sal>* with fanout 50. Each Employees record is 100 bytes long, and you can assume that each index data entry is 20 bytes long, and that the index leaf page size is the same as the relation table's data page size. The Employees relation contains 10,000 pages, and each data page contains 20 employee tuples. There are 11 buffer pages.

1. Consider each of the following selection conditions and, assuming that the reduction factor (RF) for each term that matches an index is 0.1, compute the cost of the most selective access path for retrieving all Employees tuples that satisfy the condition:
   (a) *sal* > 100
   (b) *age* = 25
   (c) *age* > 20
   (d) *eid* = 1000
   (e) *sal* > 200 ∧ *age* > 30
   (f) *sal* > 200 ∧ *age* = 20
   (g) *sal* > 200 ∧ *title* ='CFO'
   (h) *sal* > 200 ∧ *age* > 30 ∧ *title* ='CFO'
2. Suppose that, for each of the preceding selection conditions, you want to retrieve the average salary of qualifying tuples. For each selection condition, describe the least expensive evaluation method and state its cost.
3. Suppose that, for each of the preceding selection conditions, you want to compute the average salary for each *age* group. For each selection condition, describe the least expensive evaluation method and state its cost.
4. Suppose that, for each of the preceding selection conditions, you want to compute the average age for each *sal* level (i.e., group by *sal*). For each selection condition, describe the least expensive evaluation method and state its cost.
5. For each of the following selection conditions, describe the best evaluation method:
   (a) *sal* > 200 ∨ *age* = 20
   (b) *sal* > 200 ∨ *title* ='CFO'
   (c) *title* ='CFO' ∧ *ename* ='Joe'

Exercise 11.2 For each of the following SQL queries, for each relation involved, list the attributes that must be examined to compute the answer. All queries refer to the following relations:

Emp(*eid:* **integer**, *did:* **integer**, *sal:* **integer**, *hobby:* **char(20)**)
Dept(*did:* **integer**, *dname:* **char(20)**, *floor:* **integer**, *budget:* **real**)

1. SELECT COUNT(*) FROM Emp E, Dept D WHERE E.did = D.did
2. SELECT MAX(E.sal) FROM Emp E, Dept D WHERE E.did = D.did

3. SELECT MAX(E.sal) FROM Emp E, Dept D WHERE E.did = D.did AND D.floor = 5
4. SELECT E.did, COUNT(*) FROM Emp E, Dept D WHERE E.did = D.did GROUP BY D.did
5. SELECT D.floor, AVG(D.budget) FROM Dept D GROUP BY D.floor HAVING COUNT(*) > 2
6. SELECT D.floor, AVG(D.budget) FROM Dept D GROUP BY D.floor ORDER BY D.floor

Exercise 11.3 You are given the following information:

> Executives has attributes *ename*, *title*, *dname*, and *address*; all are string fields of the same length.
> The *ename* attribute is a candidate key.
> The relation contains 10,000 pages long.
> Each Executives record is 100 bytes long. You can assume that each index data entry is 25 bytes and that the index fanout is always 50, and that the index leaf page size is the same as the relation table's data page size.
> There are 10 buffer pages.

1. Consider the following query:

SELECT E.title, E.ename FROM Executives E WHERE E.title='CFO'

Assume that only 10% of Executives tuples meet the selection condition.
(a) Suppose that a clustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan? (In this and subsequent questions, be sure to describe the plan you have in mind.)
(b) Suppose that an unclustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan?
(c) Suppose that a clustered B+ tree index on *ename* is (the only index) available. What is the cost of the best plan?
(d) Suppose that a clustered B+ tree index on *address* is (the only index) available. What is the cost of the best plan?
(e) Suppose that a clustered B+ tree index on <*ename, title*> is (the only index) available. What is the cost of the best plan?

2. Suppose that the query is as follows:

SELECT E.ename FROM Executives E WHERE E.title='CFO' AND E.dname='Toy'

Assume that only 10% of Executives tuples meet the condition *E.title ='CFO'*, only 10% meet *E.dname ='Toy'*, and that only 5% meet both conditions.
(a) Suppose that a clustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan?
(b) Suppose that a clustered B+ tree index on *dname* is (the only index) available. What is the cost of the best plan?
(c) Suppose that a clustered B+ tree index on <*title, dname*> is (the only index) available. What is the cost of the best plan?
(d) Suppose that a clustered B+ tree index on <*title, ename*> is (the only index) available. What is the cost of the best plan?
(e) Suppose that a clustered B+ tree index on <*dname, title, ename*> is (the only index) available. What is the cost of the best plan?
(f) Suppose that a clustered B+ tree index on <*ename, title, dname*> is (the only index) available. What is the cost of the best plan?

3. Suppose that the query is as follows:

SELECT E.title, COUNT(*) FROM Executives E GROUP BY E.title

(a) Suppose that a clustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan?
(b) Suppose that an unclustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan?
(c) Suppose that a clustered B+ tree index on *ename* is (the only index) available. What is the cost of the best plan?
(d) Suppose that a clustered B+ tree index on <*ename, title*> is (the only index) available. What is the cost of the best plan?
(e) Suppose that a clustered B+ tree index on <*title, ename*> is (the only index) available. What is the cost of the best plan?

4. Suppose that the query is as follows:

SELECT E.title, COUNT(*) FROM Executives E WHERE E.dname > 'W%' GROUP BY E.title

Assume that only 10% of Executives tuples meet the selection condition.
(a) Suppose that a clustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan? If an additional index (on any search key you want) is available, would it help produce a better plan?
(b) Suppose that an unclustered B+ tree index on *title* is (the only index) available. What is the cost of the best plan?
(c) Suppose that a clustered B+ tree index on *dname* is (the only index) available. What is the cost of the best plan? If an additional index (on any search key you want) is available, would it help to produce a better plan?
(d) Suppose that a clustered B+ tree index on <*dname, title*> is (the only index) available. What is the cost of the best plan?
(e) Suppose that a clustered B+ tree index on <*title, dname*> is (the only index) available. What is the cost of the best plan?