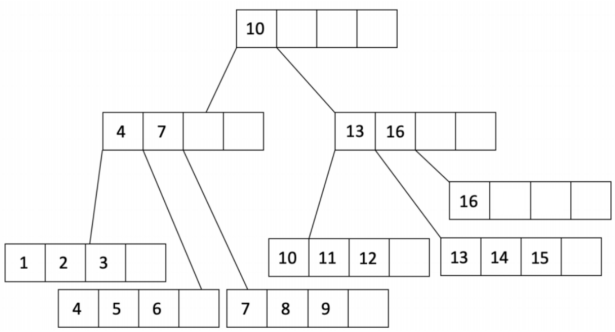
CS150A Homework 1 – Writing

School of Information Science and Technology
May 11, 2024

1 BULK LOADING (10 PTS)

Suppose we were to create an order d=2 B+ tree via bulk-loading with a fill factor of 3/4. Here, fill factor specifies the fill factor for leaves only; inner nodes should be filled up to full and split in half exactly.
We insert keys with all integer values from 1-16 in order. Draw out the final B+ tree. What is its height?



height is 2

2 RELATIONAL ALGEBRA (40 PTS)

As shown in the following figures, there are three instances corresponding to three relations: Products, Customers and Orders. Products Table has 3 attributes: Pname (The name of the product), Pid (The id of the product), Price (The price of the product). Customers Table has 3 attributes: Cname (The name of the customer), Cid (The id of the customer), Region (The

living region of the customer). Orders Table has 3 attributes: Cid (The customer id of the order), Pid (The product id of the order), Quantity (The quantity of the product of the oder).

PRODUCTS

Pname	Pid	Pprice
disks	131	$100
pcs	152	$700
macs	831	$800
printers	255	$120
paper	221	$5

(a) Products Table

CUSTOMERS

Cname	Cid	Region
Bob	1	TX
Harry	2	TX
Linda	3	MA
Martha	4	FL
Lin	5	FL
Leyla	6	CA

(b) Customers Table

ORDERS

Cid	Pid	Quantity
1	152	1
2	152	2
4	131	3
5	255	1
6	831	1

(c) Orders Table

Use relational algebra to describe the following queries.

1. Find the names of products which are cheaper than $200 and with odd id.

   $\pi_{\text{Pname}}(\sigma_{\text{Price}<200 \wedge \text{Pid}\%2=1}(\text{Products}))$

2. Find the quantity of an order ordered by customer named "Martha".

   $\pi_{\text{Quantity}}(\sigma_{\text{Cname}=\text{"Martha"}}(\text{Customers} \bowtie \text{Orders}))$

3. Find the price of products which is ordered by customers from Region "FL".

   $\pi_{\text{Price}}(\sigma_{\text{Region}=\text{"FL"}}(\text{Customers} \bowtie \text{Orders} \bowtie \text{Products}))$

4. Find the regions of customers who have ordered a product with quantity larger than 1.

   $\pi_{\text{Region}}(\sigma_{\text{Quantity}>1}(\text{Orders} \bowtie \text{Customers}))$

5. Find the name of customers who didn't order product called "pcs".

   $\pi_{\text{Cname}}(\text{Customers}) - \pi_{\text{Cname}}(\sigma_{\text{Pname}=\text{"pcs"}}(\text{Customers} \bowtie \text{Orders} \bowtie \text{Products}))$

3 JOINS (10 PTS)

Determine whether each of the following statements is True or False.

a. Page Nested Loops join will always perform at least as well as Simple Nested Loops Join when it comes to minimizing I/Os.

   T

b. Grace hash join is usually the best algorithm for joins in which the join condition includes an inequality.

   F

4 Query Optimization (40 pts)

For the following questions, assume the following:

- The System R assumptions about uniformity and independence from lecture hold.
- We use System R defaults when selectivity estimation is not possible

Table Schema	Records	Pages	Notes
CREATE TABLE Students (id INTEGER PRIMARY KEY, name VARCHAR, age INTEGER, email VARCHAER,)	500	50	Primary key *id* is sequential in this table, starting from 1, and there are no gaps in *id*.
CREATE TABLE Selections (sid INTEGER PRIMARY KEY, student_id REFERENCE Students(id), course_id REFERENCE Courses(id), grade INTEGER,)	25,000	2,500	None
CREATE TABLE Courses (id INTEGER PRIMARY KEY, name VARCHAR, rating INTEGER, credit INTEGER,)	1,000	100	- There is a clustered Alternative 3 index on *id* of height 2. - *id* ranges from 1 to 2500. - *rating* for each course is unique. - *rating* ranges from 1 to 3,000.

1. (5 pts) What is the selectivity of `id <= 100` from the Students table?

   We get the selectivity as $\frac{100-1+1}{500} = \frac{1}{5}$

2. (5 pts) What is the selectivity of `rating = 2500 AND credit < 5` from the Courses table?

   We know that the selectivity of `price = 10000` is $\frac{1}{3000}$ because prices are unique. We don't know $\max(credit)$ and $\min(credit)$, so we can assume it is $\frac{1}{10}$. Therefore, our resulting selectitivy is $\frac{1}{3000} \cdot \frac{1}{10} = \frac{1}{30000}$.

3. (10 pts) After applying the System R optimizer, which query has a lower estimated final cost for the optimal query plan? Please explain the choices you made.

   A. `SELECT * FROM Students`

   B. `SELECT * FROM Selections WHERE sid > 1000`

   C. `SELECT * FROM Courses WHERE id > 1000`

   Choice A would be a full table scan of 50 pages. Choice B would be a full table scan of 2500 pages. Choice C would be an index scan on *id*. We have to scan at least 60 pages because the selectivity of `id > 1000` is $\frac{3}{5}$. Therefore, choice A has the lowest cost.

Join	Left Relation	Left Ordering	Right Relation	Right Ordering	Join Type
(A)	Students	none	Selections	none	PNLJ
(B)	Courses	id	Selections	none	SMJ
(C)	Students	none	Courses	id	BNLJ
(D)	Students	id	Selections	student_id	SMJ

4. (10 pts) Which of the joins from the table above will NOT result in an interesting order for the next pass of the System R algorithm for the following query? Please explain the choices you made.

```
SELECT *
FROM Students as Stu, Selections as Sel, Courses as C
WHERE Stu.id = Sel.student_id AND C.id = Sel.course_id AND C.id < 100
ORDER BY C.id;
```

Choice A and C will both not result interesting orders because they are block nested loop joins. Choice B will not result in an interesting order because the *id* column is not used in a later join or for an order by. Choice D does result in an interesting order because it can be used later on for the order by clause

5. (10 pts) Which of the following join orders will **NOT** be considered for Pass 3 in the System R algorithm for the query above? Please explain the choices you made.

   A. `Students ⋈ (Selections ⋈ Courses)`

   B. `(Students ⋈ Courses) ⋈ Selections`

   C. `(Courses ⋈ Selections) ⋈ Students`

   D. `Selections ⋈ (Students ⋈ Courses)`

   E. `(Students ⋈ Selections) ⋈ Courses`

   F. `Courses ⋈ (Selections ⋈ Students)`

   G. `(Selections ⋈ Courses) ⋈ Students`

Choice A, D, and F are not left-deep joins so they will not be considered. Choice B has a cross join between C and I so it will not be considered