

Corso di Basi di Dati II

Primo Compitino

18 Aprile 2002

1. (Indicativamente, punti 7) Dato il seguente insieme di chiavi:

18, 7, 2, 10, 24, 15, 32, 21, 1, 3, 9, 11, 20, 17, 27, 30, 41, 44

- (a) mostrare il B-albero di ordine 5 ottenuto inserendo un elemento dopo l'altro nell'ordine dato;
- (b) mostrare il B-albero ottenuto dalla cancellazione dell'elemento 1 e poi dell'elemento 11 dal B-albero in (a);
- (c) mostrare il B⁺-albero di ordine 4 ottenuto inserendo un elemento dopo l'altro nell'ordine dato;
- (d) mostrare il B⁺-albero ottenuto dalla cancellazione dell'elemento 18, poi dell'elemento 41, poi dell'elemento 44 dal B⁺-albero in (c).

Per ogni punto, commentare brevemente il procedimento applicato.

2. (Indicativamente punti 7) Dato il seguente insieme di chiavi

18, 7, 2, 10, 24, 15, 32, 21, 1, 3, 9, 11

corrispondente ad un sottoinsieme delle chiavi elencate nell'esercizio 1:

- (a) mostrare la struttura di hash virtuale, con funzioni $H_i(k) = k \bmod 3 * 2^i$ e capacità di ogni pagina 2, ottenuta inserendo gli elementi nell'ordine dato; giustificare l'eventuale duplicazione della tabella;
 - (b) mostrare le operazioni effettuate per ricercare in tale struttura le chiavi 10 e 19.
3. (Indicativamente punti 4) Mostrare la struttura di hash estensibile, con capacità di ogni pagina 2, supponendo di inserire nell'ordine i record le cui pseudochiavi hanno i seguenti valori:
- $$h(r_1) = 1101110, h(r_2) = 1010011, h(r_3) = 0011000, h(r_4) = 0100101, h(r_5) = 1111111, h(r_6) = 1110101$$
4. (Indicativamente punti 2) Si supponga che il B-tree e il B+-tree dell'esercizio 1 e l'organizzazione hash dell'esercizio 2 si riferiscano ai valori dell'attributo *Stipendio* (espresso in centinaia di Euro) di una relazione *Impiegati*(*Nome*, *Mansione*, *Stipendio*, *Dip*).
- (a) Quale tecnica conviene utilizzare se le interrogazioni più frequenti sono del tipo $Stipendio = c$, con c costante?
 - (b) Quale tecnica conviene utilizzare se le interrogazioni più frequenti sono del tipo $Stipendio \geq c$, con c costante?

- (c) Specificare sotto quali ipotesi sul dominio applicativo é possibile utilizzare un indice bitmap per velocizzare le operazioni sull'attributo *Stipendio*. Presentare un'istanza della relazione *Impiegati* e il relativo indice bitmap per la colonna *Stipendio*. Per quali altri attributi della relazione *Impiegati* potrebbe essere ragionevole costruire un indice bitmap? Perché?
5. (Indicativamente punti 3) Si consideri lo schema relazionale IMPIEGATI(nome, qualifica, dipartimento, indirizzo) dove i campi sono tutti e quattro di tipo `char(20)`. Si consideri la seguente query di proiezione

```
SELECT DISTINCT qualifica, nome FROM IMPIEGATI
```

Sapendo che la relazione occupa 1000 pagine, che ci sono 10 pagine di buffer, che l'attributo nome é una chiave candidata e utilizzando la versione ottimizzata dell'algoritmo di proiezione basato sul sorting (in cui l'external merge sort viene modificato per effettuare la proiezione)

- quante sottoliste ordinate vengono prodotte al primo passo? ognuna quante pagine é lunga? qual é il costo in termini di operazioni di I/O della prima fase?
 - quanti passi di merge é necessario fare per ottenere il risultato della proiezione? qual é il costo in termini di operazioni di I/O della fase di merge?
 - si supponga di avere un indice B+ clusterizzato su qualifica, tale indice può essere utilizzato per effettuare più efficientemente la proiezione? e se l'indice fosse di tipo hash? e se l'indice fosse non clusterizzato?
 - si supponga ora di avere un indice B+ clusterizzato su (nome,qualifica), tale indice può essere utilizzato per effettuare più efficientemente la proiezione? e se l'indice fosse di tipo hash? e se l'indice fosse non clusterizzato?
6. (Indicativamente punti 5) Si consideri lo schema relazionale R(A,B,C,D,E,F) e la seguente interrogazione su tale schema:

```
SELECT A,B
FROM R
WHERE A BETWEEN (10,50) AND B IN ("a","b","c")
      AND (C = "k" OR D = "h") AND E = "e" AND F > 10
```

con R relazione con 5000 tuple memorizzate su 500 pagine, su cui sono definiti i seguenti indici:

- I_1 indice B+ clusterizzato sull'attributo *A*, tale che $\text{Max}(A,R) = 120$, $\text{Min}(A,R) = 0$, con 120 entrate e memorizzato su 10 foglie
- I_2 indice B+ non clusterizzato sull'attributo *B*, con 100 entrate e memorizzato su 10 foglie
- I_3 indice B+ non clusterizzato sull'attributo *C*, con 150 entrate e memorizzato su 12 foglie
- I_4 indice B+ non clusterizzato sull'attributo *E*, con 100 entrate e memorizzato su 10 foglie
- I_5 indice hash non clusterizzato sull'attributo *F*, tale che $\text{Max}(F,R) = 20$, $\text{Min}(F,R) = 0$.

Considerando un ottimizzatore quale quello del System R che utilizza al più un indice come cammino di accesso ad una relazione:

- mostrare quali piani di esecuzione dell'interrogazione vengono considerati
- stimare il costo di esecuzione di ognuno di essi
- individuare il piano di accesso di costo minimo, che sarà quindi scelto dall'ottimizzatore.

7. (Indicativamente punti 6) Si consideri il seguente schema relazionale

IMPIEGATI(eid:integer, did: integer, sal: integer, hobby: char(20))
DIPARTIMENTI(did:integer, nomedip: char(20), piano: integer, telefono: char(10))
BILANCIO(did:integer, budget: real, uscite: real, entrate: real)

e la seguente interrogazione

```
SELECT nomedip, budget
FROM IMPIEGATI,DIPARTIMENTI,BILANCIO
WHERE IMPIEGATI.did = DIPARTIMENTI.did AND
      DIPARTIMENTI.did = BILANCIO.did AND piano = 1 AND
      sal > 59000 and hobby = "giardinaggio"
```

- (a) identificare un piano logico di esecuzione (cioé un albero rappresentante un'espressione algebrica) che rifletta l'ordine delle operazioni che un ottimizzatore "ragionevole" sceglierebbe
- (b) elencare gli ordini di join (cioé l'ordine in cui viene effettuato il join di coppie di relazioni per produrre il risultato dell'interrogazione) che un ottimizzatore basato sul system R considererebbe
- (c) supponendo di avere le seguenti informazioni aggiuntive:
 - indici B+ non clusterizzati su IMPIEGATI.did, IMPIEGATI.sal, DIPARTIMENTI.piano, DIPARTIMENTI.did e BILANCIO.did
 - i salari degli impiegati variano da 10000 a 60000, gli hobby diversi degli impiegati sono 200 e la compagnia occupa due piani dell'edificio
 - nella base di dati ci sono 50000 impiegati e 5000 dipartimenti (ognuno con le relative informazioni sul bilancio)
 - l'unico metodo di join a disposizione é l'index nested loop
- i. per ogni relazione di base stimare quante tuple saranno selezionate dalla relazione se tutti i predicati non di join vengono applicati prima di iniziare l'elaborazione dei join
- ii. sulla base della risposta alla domanda precedente, qual é l'ordine di join di minor costo stimato e quale il suo costo?