

Tecnologie delle Basi di Dati M

Appello del 25/10/2013

Esercizio 1 (3 punti)

Data la relazione con schema:

Studenti(matricola, cognome, nome, datanascita, luogonascita, indirizzo)

si effettui una stima del numero di pagine necessarie per memorizzare la relazione e del numero di livelli (e di nodi) di un B⁺-tree *denso* costruito sull'attributo composto (cognome, nome). Si supponga che non esistano omonimie, di avere pagine di dimensione 4 KB, di cui 96 B riservati per il page header, e si considerino i seguenti valori:

- Numero di tuple = 40K
- Dimensione *matricola* = 4 byte
- Dimensione *cognome* = 38 byte
- Dimensione *nome* = 38 byte
- Dimensione *datanascita* = 10 byte
- Dimensione *luogonascita* = 30 byte
- Dimensione *indirizzo* = 40 byte
- Dimensione RID = 6 byte
- Dimensione PID = 4 byte
- Percentuale di riempimento foglie = 82%

Esercizio 2 (4 punti)

Date le relazioni con schema:

Esami(matricola, codice, data, voto)

Studenti(matricola, nome, cognome, data, luogo, cdl)

si ottimizzi l'esecuzione della seguente interrogazione SQL:

```
SELECT S.nome, S.cognome, E.voto
FROM Studenti S, Esami E
WHERE S.matricola=E.matricola
AND E.voto >= 30
AND S.data BETWEEN 1/1/1991 AND 31/12/1991
```

tenendo conto che dai cataloghi della base di dati risulta:

- Numero di tuple *Esami* = 100K
- Numero di pagine *Esami* = 2K
- Numero di tuple *Studenti* = 10K
- Numero di pagine *Studenti* = 2K
- Indice clustered su *voto*: numero foglie = 320, numero voti = 16
- Indice unclustered su *E.matricola*: numero foglie = 400
- Indice unclustered su *S.matricola*: numero foglie = 100
- Indice unclustered (TID ordinate) su *data*: numero foglie = 100, valore min = 1/1/1971, valore max = 31/12/1995

Si disegni infine l'albero corrispondente al piano di accesso di costo minimo e stimi il numero di risultati dell'interrogazione.

Suggerimento: per la formula di Cardenas si utilizzino i seguenti valori, validi per P = 2000:

R	$\Phi(R, 2000)$
200	190.37
400	362.62
600	518.47
800	659.49
1000	787.09
1200	902.54
1400	1007.00
1600	1101.52
1800	1187.04
2000	1264.43

R	$\Phi(R, 2000)$
2200	1334.44
2400	1397.79
2600	1455.11
2800	1506.98
3000	1553.91
3200	1596.37
3400	1634.79
3600	1669.55
3800	1701.00
4000	1729.46

R	$\Phi(R, 2000)$
4200	1755.22
4400	1778.52
4600	1799.60
4800	1818.67
5000	1835.93
5200	1851.55
5400	1865.68
5600	1878.47
5800	1890.03
6000	1900.50

Esercizio 3 (5 punti)

Si confrontino il linear hashing e l'extendible hashing, evidenziandone i rispettivi pregi e difetti.

Esercizio 4 (3 punti)

Si illustri quale sia l'effetto di incrementare/decrementare la dimensione della pagina dati su una struttura ad indice multi-dimensionale paginata quale R-tree, discutendone in particolare l'impatto sui costi di inserimento e ricerca.

Soluzione Esercizio 1

Dimensionamento relazione:

Dimensione di ogni tupla = $4 + 38 + 38 + 10 + 30 + 40 = 160\text{B}$

Numero di tuple per pagina = $(4096 - 96)/160 = 4000/160 = 25$

Numero di pagine della relazione = $NT/25 = 40000/25 = 1600$

Dimensionamento indice (cognome, nome):

Numero di chiavi = 40K.

Dimensione di ogni record (foglia) = $6 + 38 + 38 = 82\text{B}$

Dimensione "reale" foglia = $(4096 - 96) \times 0.82 = 3280\text{B}$

Numero di record per foglia = $3280/82 = 40$

Numero di foglie = $40000/40 = 1000$

Dimensione di ogni record (nodo interno) = $4 + 38 + 38 = 80\text{B}$

Numero di record per nodo interno = $(4096 - 96)/80 = 50$

Numero nodi livello 1 = $1000/50 = 20$

Numero nodi livello 2 = $20/50 = 1$

Il B⁺-tree corrispondente si compone quindi di 3 livelli per un totale di 21 nodi interni e 1000 foglie.

Soluzione Esercizio 2

Selettività dei predicati:

Predicato `E.voto >= 30` = $2/16 = 0.125$

Predicato `S.data BETWEEN 1/1/1991 AND 31/12/1991` = $1/25 = 0.04$

Predicato di join = $1/10\text{K}$ (chiave esterna)

Accesso a S:

Costo scan sequenziale = **2000**

Costo indice su `data`: $NL / 25 + \Phi(NT / 25, NP) = 100/25 + \Phi(10\text{K}/25, 2\text{K}) = 4 + 363 = 367$

Costo indice su `matricola`: $1 + 1 = 2$

Numero tuple residue = $NT / 25 = 400$

Accesso a E:

Costo scan sequenziale = **2000**

Costo indice su `voto`: $NL \times 0.125 + NP \times 0.125 = 320/8 + 2\text{K}/8 = 40 + 250 = 290$

Costo indice su `matricola`: $NL / 10\text{K} + NT / 10\text{K} = 400/10\text{K} + 100\text{K}/10\text{K} = 1 + 10 = 11$

Numero tuple residue = $NT / 8 = 12500$

Costi di join:

S esterna: costo = costo indice su `data` + $400 \times$ costo indice `matricola` = $367 + 400 \times 11 = 4767$

E esterna: costo = costo indice su `voto` + $12500 \times$ costo indice su `matricola` = $290 + 12500 \times 2 = 25290$

Il numero di risultati dell'interrogazione è $100\text{K} \times 0.125 \times 0.04 = 500$