

Tecnologie delle Basi di Dati M

Appello del 24/7/2013

Esercizio 1 (3 punti)

Data la relazione con schema:

Personale(codice, cognome, nome, datanascita)

si effettui una stima del numero di livelli (e di nodi) di un B⁺-tree *denso* e di uno *sperso* costruiti sull'attributo (cognome). Si supponga 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 = 250K
- Numero di pagine = 5K
- Numero di chiavi (cognome) = 50K
- Dimensione cognome = 30 byte
- Dimensione RID = 8 byte
- Dimensione PID = 6 byte
- Percentuale di riempimento foglie = 80%

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 = '30L'
      AND S.luogo in ('Bologna', 'Ferrara', 'Modena')
```

tenendo conto che dai cataloghi della base di dati risulta:

- Numero di tuple Esami = 480K
- Numero di pagine Esami = 1.6K
- Numero di tuple Studenti = 48K
- Numero di pagine Studenti = 2K
- Indice clustered su voto: numero foglie = 320, numero voti = 16
- Indice unclustered (TID ordinate) su E.matricola: numero foglie = 200
- Indice unclustered su S.matricola: numero foglie = 100
- Indice clustered su luogo: numero foglie = 100, numero luoghi = 1K

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 = 1600:

R	$\Phi(R, 1600)$
10	9.97
20	19.88
30	29.73
40	39.52
50	49.24
60	58.91
70	68.51
80	78.06
90	87.54
100	96.97

R	$\Phi(R, 1600)$
110	106.34
120	115.65
130	124.90
140	134.09
150	143.23
160	152.31
170	161.33
180	170.29
190	179.21
200	188.06

R	$\Phi(R, 1600)$
210	196.86
220	205.60
230	214.30
240	222.93
250	231.51
260	240.04
270	248.52
280	256.94
290	265.31
300	273.63

Esercizio 3 (5 punti)

Si confrontino le diverse tecniche per la gestione di overflow in area primaria per le strutture hash (metodi di concatenamento/indirizzamento aperto), evidenziandone pregi e difetti.

Esercizio 4 (3 punti)

Il protocollo di recovery del sistema ARIES si basa sull'assunzione che l'operazione di checkpoint sia un'azione atomica, assunzione ovviamente non realistica. Vi saranno, perciò, due record di inizio e fine checkpoint ed è quindi possibile che alcune transazioni scrivano record nel log durante l'effettuazione del checkpoint. Si illustri come debbano essere trattati tali record in un'eventuale fase di analisi seguita ad un crash del sistema.

Soluzione Esercizio 1

Dimensionamento indice denso:

Numero di chiavi = 50K, mediamente ci sono $250K/50K = 5$ tuple per ogni valore di chiave.

Dimensione di ogni record (foglia) = $30 + 5 \times 8 = 70$ B

Dimensione "reale" foglia = $(4096 - 96) \times 0.80 = 3200$ B

Numero di record per foglia = $3200/70 = 46$

Numero di foglie = $50000/46 = 1087$

Dimensione di ogni record (nodo interno) = $6 + 30 = 36$ B

Dimensione "reale" nodo interno = $4096 - 96 = 4000$ B

Numero di record per nodo interno = $4000/36 = 111$

Numero nodi livello 1 = $1087/111 = 10$

Numero nodi livello 2 = $10/111 = 1$

Il B⁺-tree corrispondente si compone quindi di 3 livelli per un totale di 11 nodi interni e 1087 foglie.

Dimensionamento indice sparso:

Dimensione di ogni record (foglia/nodo interno) = $6 + 30 = 36$ B

Dimensione "reale" foglia = $(4096 - 96) \times 0.80 = 3200$ B

Numero di record per foglia = $3200/36 = 89$

Numero di foglie = $5000/89 = 57$

Dimensione "reale" nodo interno = $4096 - 96 = 4000$ B

Numero di record per nodo interno = $4000/36 = 111$

Numero nodi livello 1 = $57/111 = 1$

Il B⁺-tree corrispondente si compone quindi di 2 livelli per un totale di 1 nodo interno e 57 foglie.

Soluzione Esercizio 2

Selettività dei predicati:

Predicato E.voto = '30L' = $1/16 = 0.0625$

Predicato S.luogo in ('Bologna', 'Ferrara', 'Modena') = $1/1000 = 0.001$ per ogni valore di luogo

Predicato di join = $1/20K$ (chiave esterna)

Accesso a S:

Costo scan sequenziale = **2000**

Costo indice su luogo: $3 \times (NL \times 0.001 + NP \times 0.001) = 3 \times (100 \times 0.001 + 2000 \times 0.001) = 3 \times (1 + 2) = 9$

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

Numero tuple residue = $3 \times NT \times 0.001 = 144$

Accesso a E:

Costo scan sequenziale = **1600**

Costo indice su voto: $NL \times 0.0625 + NP \times 0.0625 = 320 \times 0.0625 + 1600 \times 0.0625 = 20 + 100 = 120$

Costo indice su matricola: $NL/48K + \Phi(NT/48K, NP) = 1 + 10 = 11$

Numero tuple residue = $NT \times 0.0625 = 30000$

Costi di join:

S esterna: costo = costo indice su luogo + $144 \times$ costo indice matricola = $9 + 144 \times 11 = 1593$

E esterna: costo = costo indice su voto + $30000 \times$ costo indice su matricola = $120 + 30000 \times 2 = 60120$

Il numero di risultati dell'interrogazione è $480K \times 0.0625 \times 3 \times 0.001 = 90$