Piani di accesso - treni

Si consideri la query:

```
select idC, tipo, durata
from corsa
where partenza = 'Milano'
and arrivo = 'Torino'
order by durata
```

sulla relazione:

CORSA (idC, tipo, partenza, arrivo, durata)

sulla quale sono costruiti due indici: uno primary, dense, clustered su *idC* (NL_{idC} *= 5662)* e uno unclustered su *durata* (NL_{durata} *= 2842).*

Si calcoli la selettività dei predicati e si determini il miglior piano di accesso per la risoluzione della query tenendo conto dei seguenti dati:

NP = 45000,

NK_{idC} *= 500000,* $NK_{partenza}$ *= 350,* NK_{arrivo} *= 350,* NK_{durata} *= 2000*

Len(idC) = 4 byte, Len(tipo) = 30 byte, Len(partenza) = 20 byte, Len(arrivo) = 20 byte

Len(durata) = 4 byte, Len(TID) = 4 byte

D = 1Kb, u=0.69

Si ipotizzi di utilizzare l'algoritmo Sort-Merge a Z=3 vie per l'eventuale ordinamento del risultato.

Si indichi infine il numero atteso di tuple nel risultato della query.

Soluzione

Selettività dei predicati

p1: partenza = 'Milano' ; p2: arrivo = 'Torino';

$$f(p1) = \frac{1}{350} \qquad f(p2) = \frac{1}{350}$$

Record attesi

Poiché su *idC* è costruito un indice primary dense, il numero di record della relazione coincide con il numero di chiavi distinte di *idC*, quindi ***NT = NK$_{idc}$ = 500000***.

ET (expected tuples); NP$_R$ (numero di pagine necessarie a contenere I dati attesi); ET$_R$ (numero atteso di tuple nel risultato finale);

$$ET = NT \times f(p1) \times f(p2) = 500000 \times \frac{1}{350} \times \frac{1}{350} \cong 4$$

$$NP_R = \left\lceil \frac{ET \times len(select\ list)}{D} \right\rceil = \left\lceil \frac{4 \times 38}{1024} \right\rceil = 1$$

$$ET_R = \mathrm{ET} = 4$$

Costi

C$_{sort}$ (costo di ordinamento con Z-way merge sort); C$_{seq}$ (costo di accesso con scansione sequenziale); C$_{idC}$ (costo di accesso tramite indice su idC) ; C$_{durata}$ (costo di accesso tramite indice su durata); C$_{tree}$ (costo di lettura dell'indice);

$$C_{sort} = 2 \times NP_R \times \lceil log_Z NP_R \rceil = 2 \times 1 \times \lceil log_3 1 \rceil = 0$$

$$C_{seq} = NP + C_{sort} = 45000$$

Se operiamo una scansione sequenziale va considerato anche il costo di ordinamento in quanto il file dati non è già ordinato sull'attributo *durata*. Il costo di ordinamento è tuttavia trascurabile in ogni caso, in quanto nullo.

Accedere tramite l'indice costruito su *idC* non conviene in quanto non ci sono predicati su questo attributo; l'accesso all'indice comporterebbe solamente un overhead di costo per la scansione dell'albero e delle sue foglie.

$$C_{idC} = \lceil 1 \times NL_{idC} \rceil + \lceil 1 \times NP \rceil + C_{sort} + C_{tree} > C_{seq}$$

Accedere tramite l'indice costruito su *durata* è ulteriormente peggiorativo: innanzitutto non ci sono predicati su questo attributo, dunque saranno lette tutte le tuple della relazione; la lettura delle foglie dell'indice può portare a leggere le stesse pagine più volte inutilmente.

$$C_{durata} = \lceil 1 \times NL_{durata} \rceil + \lceil 1 \times NK_{durata} \rceil \times \Phi\left(\frac{NT}{NK_{durata}}, NP\right) + C_{tree} \gg C_{seq}$$