Indici Hash

Lorenzo Vaccarecci

12 Aprile 2024

1 Funzione Hash

- distribuzione uniforme delle chiavi nello spazio negli indirizzi
- distrinuzione casuale delle chiavi (eventuali correlazioni tra i valori delle chiavi non devono tradursi in correlazioni tra gli indirizzi generati)

Una funzione hash è detta perfetta se non vengono prodotti trabocchi. Può essere sempre definita disponendo di un'area primaria con capacità complessiva pari al numero dei record da memorizzare. Le funzioni hash operano su insiemi di chiavi intere (se le chiavi sono alfanumeriche si può associare un id prima di applicare la trasformazione).

1.1 Metodo della divisione

La chiave numerica viene divisa per M e l'indirizzo è ottenuto considerando il resto:

$$H(k) = k\%M$$

Per avere una buona distribuzione delle chiavi è opportuno che M sia un numero primo oppure ≤ 20.

1.2 Costi

- In assenza di overflow, il costo di accesso a indice è **costante**.
- In presenza di overflow, il costo non è facilmente determinabile.
  - Quanti blocchi di overflow per il blocco acceduto?
  - Dove sono memorizzati?

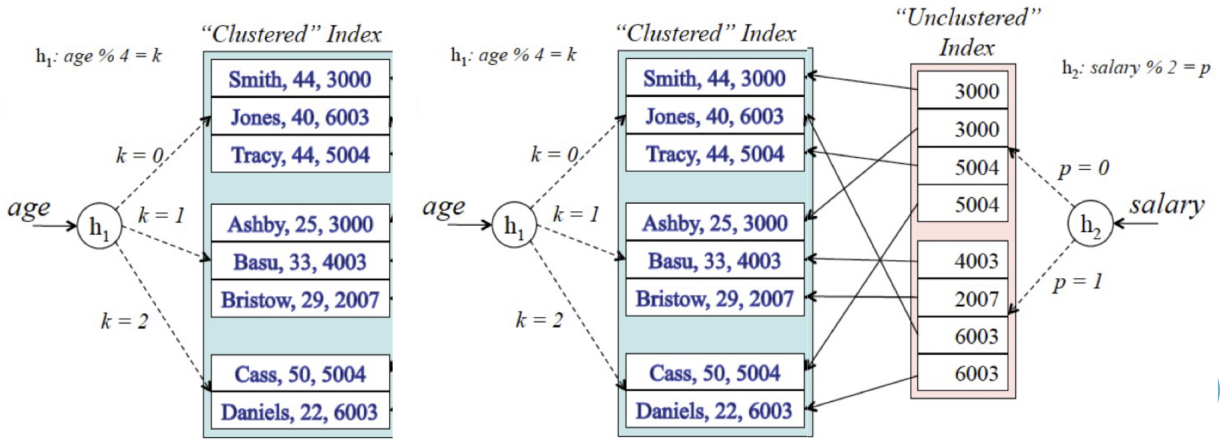
2 Creare indirizzi hash

- Specifica della funzione H.
- Specifica del metodo della gestione dei trabocchi.
- Specifica del fattore di caricamento d: si intende quanto pieni vogliamo i bucket, se l'istanza della base di dati cambia frequentemente è meglio avere un fattore di caricamento basso ($0 \leq d \leq 1$). Dipende anche da quanti blocchi il sistema decide di tenere in un bucket.
- L'amministratore del database può **al più** agire la funzione H, ma non in tutti i DBMS.

3 Indici hash clusterizzati e non clusterizzati

E' clusterizzato se i record con chiavi simili sono memorizzati nello stesso bucket. In caso contrario è non clusterizzato.
Finora abbiamo considerato solo gli indici clusterizzati.
Un file dei dati di tipo hash è sempre associato a un indice hash clusterizzato.
In presenza di un indice hash **clusterizzato** l'organizzazione primaria corrisponde ai record memorizzati nell'area primaria + i record memorizzati negli overflow.

4 Esempio



Gli indici non clusterizzati sono indici multilivello.
Per selezioni di uguaglianza sono preferibili gli indici hash, per le selezioni di range sono preferibili gli indici ad albero.