Tecnologia delle basi di dati (ex Basi di dati, primo modulo)

20 luglio 2005 — Compito A

Tempo a disposizione: due ore e quindici minuti

Domanda 1 (15%)
Sulla base di quanto studiato nell'ambito di questo corso, commentare brevemente (massimo mezza pagina, non più di 150-200 parole) ma in modo organico

- l'osservazione di autorevoli studiosi (fra cui P. Salinger) secondo cui, visto il crescere del costo del lavoro, è necessario ridurre il costo di gestione dei DBMS

oppure

- l'osservazione di M. Stonebraker ("One size does not fit all") secondo cui non esiste un DBMS ottimale per tutti i tipi di applicazioni.

Domanda 2 (20%)

a. Concepire un algoritmo basato su tecniche hash (analoghe quindi a quella usata nell'hash-join) che permetta di eseguire in modo efficiente un'operazione di unione con eliminazione dei duplicati.
b. Provare a determinare quanto spazio di memoria (buffer) è approssimativamente necessario per far sì che l'algoritmo richieda un numero di accessi pari a circa tre o quattro volte (una prima lettura, una memorizzazione e una seconda lettura seguita eventualmente dalla memorizzazione del risultato) il numero di blocchi occupati complessivamente dai due operandi.
   Nota: si ricordi che, nell'hash join, per ciascun valore della funzione hash si usa un "bucket" costituito da un certo numero di n blocchi, lo stesso per tutti i valori della funzione (se poi n blocchi non bastano, si ha overflow); un algoritmo efficiente sceglie opportunamente il valore di n e la dimensione del buffer.

Domanda 3 (15%)
Considerare le seguenti due versioni del checkpoint e commentare brevemente le differenze in termini di prestazioni e di procedure di ripristino dopo i guasti

checkpoint, versione A

- si sospende l'accettazione di nuove transazioni e si aspetta che le transazioni attive raggiungano la conclusione
- si trasferiscono in memoria di massa (tramite force) tutte le pagine sporche relative a transazioni andate in commit
- si scrive sul log in modo sincrono (force) un record che indica il completamento del checkpoint
- si riprende l'accettazione delle operazioni

checkpoint, versione B

- si sospende l'accettazione di richieste di ogni tipo (scrittura, inserimenti, ..., commit, abort)
- si trasferiscono in memoria di massa (tramite force) tutte le pagine sporche relative a transazioni andate in commit
- si scrive sul log in modo sincrono (force) un record che indica il completamento del checkpoint, con gli identificatori delle transazioni in corso
- si riprende l'accettazione delle operazioni

Domanda 4 (25%)
Con riferimento ad un sistema che abbia blocchi di dimensione $B = 1000$ byte e puntatori ai blocchi di $p = 5$ byte, considerare una base di dati sulle seguenti relazioni, ognuna delle quali ha una struttura heap e un indice secondario sulla chiave

- R_1*(*$\underline{A}BC$*)*, con $N_1 = 1.000.000$ ennuple di lunghezza fissa pari a $l_1 = 50$ byte, di cui $l_A = 10$ per il campo chiave A, e vincolo di riferimento fra C e la chiave D di R_2
- R_2*(*$\underline{D}EF$*)*, con $N_2 = 400.000$ ennuple di lunghezza fissa pari a $l_2 = 100$ byte, di cui $l_D = 4$ per il campo chiave D
- R_3*(*$\underline{G}HL$*)* con $N_3 = 1.000.000$ ennuple di lunghezza fissa pari a $l_3 = 25$ byte, di cui $l_G = 5$ per il campo chiave G.

e con una vista definita come segue:

- `CREATE VIEW V AS SELECT * FROM (R1 JOIN R2 ON C=D) LEFT JOIN R3 ON F=G`

In tale contesto, considerare le seguenti interrogazioni

1. `SELECT A, L FROM V`
2. `SELECT A FROM V`

e, per ciascuna di esse

1. Mostrare un possibile piano di esecuzione (in termini di operatori dell'algebra relazionale e loro realizzazioni).
2. Stimare il costo, in termini di numero di accessi a memoria secondaria (ignorando la presenza di eventuali buffer).

Domanda 5 (25%) Si consideri la base di dati seguente, relativa alla segreteria studenti di una università:

- Studenti(Matricola, Cognome, Nome, DataNascita, CorsoDiLaurea, AnnoDiCorso, AnnoDiImmatricolazione, Residenza);
- Esami(Studente, Corso, Data, Voto), con vincoli di riferimento verso Studenti e verso Corsi
- Corso(Codice, Nome, AnnoDiCorso, Crediti)

Con riferimento a tale base di dati, progettare uno schema dimensionale che permetta di rispondere facilmente ad interrogazioni quali ad esempio (la lista non ha pretesa di essere esaustiva):

- calcolare il numero di studenti che hanno superato l'esame di un certo corso in un certo intervallo di tempo (specificato con giorno iniziale e giorno finale) e il relativo voto medio;
- calcolare l'età media degli studenti che hanno superato l'esame di un certo corso;
- calcolare il numero medio di crediti acquisiti dagli studenti nel momento in cui hanno superato l'esame di un certo corso;
- calcolare il numero di esami superati e la media complessiva dei voti per provincia di residenza (ad esempio, gli studenti provenienti da Frosionone hanno superato 4 esami con voto medio 23 e quelli di Latina 3 con voto medio 24) per gli studenti immatricolati in un certo anno;
- calcolare, per ciascun corso, il numero di studenti di ciascun anno di corso che hanno superato l'esame, con la relativa media.

Assumere i seguenti vincoli e ipotesi:

- per (presunte) ragioni di privatezza, non si possono rappresentare i singoli esami come fatti, ma solo loro aggregazioni;
- l'unico attributo di studente che cambia nel tempo è l'anno di corso;
- l'anno e il numero di crediti attribuiti ad un corso non cambiano nel tempo.

Allo scopo:

1. specificare fatti, misure e dimensioni dello schema dimensionale utilizzato (uno solo);
2. supponendo che l'alimentazione del data mart sia quotidiana, specificare le operazioni (ad esempio in termini di istruzioni SQL o di espressioni dell'algebra relazionale) da svolgere per popolare la tabella dei fatti.

Tecnologia delle basi di dati (ex Basi di dati, primo modulo)

20 luglio 2005 — Compito B

Tempo a disposizione: due ore e quindici minuti

Domanda 1 (15%)
Sulla base di quanto studiato nell'ambito di questo corso, commentare brevemente (massimo mezza pagina, non più di 150-200 parole) ma in modo organico

- l'osservazione di autorevoli studiosi (fra cui P. Salinger) secondo cui, visto il crescere del costo del lavoro, è necessario ridurre il costo di gestione dei DBMS

oppure

- l'osservazione di M. Stonebraker ("One size does not fit all") secondo cui non esiste un DBMS ottimale per tutti i tipi di applicazioni.

Domanda 2 (20%)

a. Concepire un algoritmo basato su tecniche hash (analoghe quindi a quella usata nell'hash-join) che permetta di eseguire in modo efficiente un'operazione di proiezione con eliminazione dei duplicati.
b. Provare a determinare quanto spazio di memoria (buffer) è approssimativamente necessario per far sì che l'algoritmo richieda un numero di accessi pari a circa tre o quattro volte (una prima lettura, una memorizzazione e una seconda lettura seguita eventualmente dalla memorizzazione del risultato) il numero di blocchi occupati dall'operando.
   Nota: si ricordi che, nell'hash join, per ciascun valore della funzione hash si usa un "bucket" costituito da un certo numero di n blocchi, lo stesso per tutti i valori della funzione (se poi n blocchi non bastano, si ha overflow); un algoritmo efficiente sceglie opportunamente il valore di n e la dimensione del buffer.

Domanda 3 (15%)
Considerare le seguenti due versioni del checkpoint e commentare brevemente le differenze in termini di prestazioni e di procedure di ripristino dopo i guasti

checkpoint, versione A

- si sospende l'accettazione di richieste di ogni tipo (scrittura, inserimenti, ..., commit, abort)
- si trasferiscono in memoria di massa (tramite force) tutte le pagine sporche relative a transazioni andate in commit
- si scrive sul log in modo sincrono (force) un record che indica il completamento del checkpoint, con gli identificatori delle transazioni in corso
- si riprende l'accettazione delle operazioni

checkpoint, versione B

- si sospende l'accettazione di nuove transazioni e si aspetta che le transazioni attive raggiungano la conclusione
- si trasferiscono in memoria di massa (tramite force) tutte le pagine sporche relative a transazioni andate in commit
- si scrive sul log in modo sincrono (force) un record che indica il completamento del checkpoint
- si riprende l'accettazione delle operazioni

Domanda 4 (25%)
Con riferimento ad un sistema che abbia blocchi di dimensione $B = 1000$ byte e puntatori ai blocchi di $p = 5$ byte, considerare una base di dati sulle seguenti relazioni, ognuna delle quali ha una struttura heap e un indice secondario sulla chiave

- R_1*(*$\underline{A}BC$*)*, con $N_1 = 1.000.000$ ennuple di lunghezza fissa pari a $l_1 = 50$ byte, di cui $l_A = 10$ per il campo chiave A, e vincolo di riferimento fra C e la chiave D di R_2
- R_2*(*$\underline{D}EF$*)*, con $N_2 = 400.000$ ennuple di lunghezza fissa pari a $l_2 = 100$ byte, di cui $l_D = 4$ per il campo chiave D
- R_3*(*$\underline{G}HL$*)* con $N_3 = 1.000.000$ ennuple di lunghezza fissa pari a $l_3 = 25$ byte, di cui $l_G = 5$ per il campo chiave G.

e con una vista definita come segue:

- `CREATE VIEW V AS SELECT * FROM (R1 JOIN R2 ON C=D) LEFT JOIN R3 ON F=G`

In tale contesto, considerare le seguenti interrogazioni

1. `SELECT A, L FROM V`
2. `SELECT A FROM V`

e, per ciascuna di esse

1. Mostrare un possibile piano di esecuzione (in termini di operatori dell'algebra relazionale e loro realizzazioni).
2. Stimare il costo, in termini di numero di accessi a memoria secondaria (ignorando la presenza di eventuali buffer).

Domanda 5 (25%) Si consideri la base di dati seguente, relativa alla segreteria studenti di una università:

- Studenti(Matricola, Cognome, Nome, DataNascita, CorsoDiLaurea, AnnoDiCorso, AnnoDiImmatricolazione, TipoScuolaSuperiore);
- Esami(Studente, Corso, Data, Voto), con vincoli di riferimento verso Studenti e verso Corsi
- Corso(Codice, Nome, AnnoDiCorso, Crediti)

Con riferimento a tale base di dati, progettare uno schema dimensionale che permetta di rispondere facilmente ad interrogazioni quali ad esempio (la lista non ha pretesa di essere esaustiva):

- calcolare il numero di studenti che hanno superato l'esame di un certo corso in un certo intervallo di tempo (specificato con giorno iniziale e giorno finale) e il relativo voto medio;
- calcolare l'età media degli studenti che hanno superato l'esame di un certo corso;
- calcolare il numero medio di crediti acquisiti dagli studenti nel momento in cui hanno superato l'esame di un certo corso;
- calcolare il numero di esami superati e la media complessiva dei voti per tipo di scuola di provenienza (ad esempio, gli studenti provenienti dal liceo scientifico hanno superato 4 esami con voto medio 23 e quelli dell'istituto tecnico 3 con voto medio 24) per gli studenti immatricolati in un certo anno;
- calcolare, per ciascun corso, il numero di studenti di ciascun anno di corso che hanno superato l'esame, con la relativa media.

Assumere i seguenti vincoli e ipotesi:

- per (presunte) ragioni di privatezza, non si possono rappresentare i singoli esami come fatti, ma solo loro aggregazioni;
- l'unico attributo di studente che cambia nel tempo è l'anno di corso;
- l'anno e il numero di crediti attribuiti ad un corso non cambiano nel tempo.

Allo scopo:

1. specificare fatti, misure e dimensioni dello schema dimensionale utilizzato (uno solo);
2. supponendo che l'alimentazione del data mart sia quotidiana, specificare le operazioni (ad esempio in termini di istruzioni SQL o di espressioni dell'algebra relazionale) da svolgere per popolare la tabella dei fatti.