Basi di dati II, primo modulo — Tecnologia delle basi di dati

24 settembre 2010 — Compito A
Cenni sulle soluzioni

Rispondere alle prime tre domande su questo foglio e alla quarta sul foglio separato.
Tempo a disposizione: un'ora e trenta minuti.

Cognome _____ **Nome** _____ **Matricola** _____ **Ordin.** _____

Domanda 1 (20%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Regione)`. Spiegare perché, per la ricerca di tutti i clienti di una certa regione, un indice secondario sull'attributo `Regione` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).
Soluzione:
Se l'attributo `Regione` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Regione` l'indice non serve, perché, dato un certo valore per `Regione` quasi tutti i blocchi contengono record con tale valore, allora una ricerca sequenziale ha un costo forse inferiore. Se invece ci sono 400 valori diversi, con l'indice si potrà accedere ai soli record di interesse, che sono il 10% del totale.

Domanda 2 (20%) Si consideri una relazione *R(A B C D E)*, in cui gli attributi hanno tutti la stessa dimensione a (ad esempio, ma è irrilevante, 4Byte), molto più piccola della dimensione del blocco pari a P. Si supponga che la relazione sia molto grande (N ennuple) e che le operazioni più frequenti su di essa siano le seguenti:

o_1 `SELECT * FROM R ORDER BY A`, con frequenza f_1

o_2 `SELECT A, B, C FROM R ORDER BY A`, con frequenza $f_2 = 10 \times f_1$

Valutare le due seguenti alternative di memorizzazione, calcolando il costo complessivo (riportare la formula che indica il numero di accessi nell'unità di tempo, in base alle variabili sopra citate):

(i) memorizzazione della relazione *R(A B C D E)* ordinata su *A*

costo unitario di o_1: $c_1 = N/(P/(5a)) = 5aN/P$ costo unitario di o_2: $c_2 = N/(P/(5a)) = 5aN/P$

costo complessivo: $c_1 f_1 + c_2 f_2 = N/(P/(5a)) = 55 f_1 aN/P$

(ii) memorizzazione delle proiezioni *R1(A B C)* e *R2(A D E)*, entrambe ordinate su *A*

costo unitario di o_1: $c_1 = 2N/(P/(3a)) = 6aN/P$ costo unitario di o_2: $c_2 = N/(P/(3a)) = 3aN/P$

costo complessivo: $c_1 f_1 + c_2 f_2 = 36 f_1 aN/P$

Supporre che il join possa essere eseguito con il metodo merge-join (e che il costo del join stesso sia trascurabile rispetto alle due scansioni).

Domanda 3 (20%) Si consideri una forma di equivalenza fra schedule denominata *final-state*-equivalenza, secondo la quale due schedule S_1 e S_2 sono equivalenti se, per ogni istanza d della base di dati, essi la trasformano nello stesso modo (e quindi $S_1(d) = S_2(d)$, se con $S(d)$ indichiamo l'istanza della base di dati ottenuta applicando a d le operazioni dello schedule S).

1. Formulare una definizione per questa proprietà, variante della definizione di view-equivalenza.
2. Spiegare, almeno intuitivamente, il rapporto che esiste fra final-state-equivalenza e view-equivalenza (sono equivalenti, incomparabili, oppure una implica l'altra?).

Soluzione:
Due schedule sono view-equivalenti se trasformano la base di dati nello stesso modo e offrono all'esterno gli stessi valori. La final-state-equivalenza non tiene conto di questa seconda proprietà. Gli schedule $w_1(x), r_2(x), w_3(x)$ e $w_1(x), w_3(x), r_2(x)$ sono final-state-equivalenti ma non view-equivalenti (perché hanno diversa relazione "legge-da"). Una definizione formale di final-state-equivalenza potrebbe richiedere (i) stesse scritture finali; (ii) stessa relazione "legge-da" per le transazioni che scrivono dopo la lettura. Di conseguenza, la final-state-equivalenza è condizione necessaria, ma non sufficiente per la view-equivalenza.

Domanda 4 (40%) Una catena di negozi gestisce le attività utilizzando, in ciascun negozio, una base di dati con le seguenti relazioni (su cui sono definiti gli ovvi vincoli di integrità referenziale):

- Prodotti(CodiceProdotto,Descrizione,Prezzo,Categoria,Marca)
- Marca(Codice,Nome)
- Categorie(Codice,Descrizione,MacroCategoria)
- MacroCategorie(Codice,Descrizione)
- Vendite(NumeroScontrino,Ora)
- DettaglioVendite(NumeroScontrino,CodiceProdotto,Quantità)

Si noti che

- Le informazioni relative alle vendite vengono mantenute solo nel corso della giornata.
- Il prezzo di un prodotto può variare da un giorno all'altro.

Utilizzando tali informazioni, la catena vuole realizzare un data mart relativo alle vendite dei prodotti, avente come misure le quantità vendute e gli incassi, che permetta di effettuare analisi di tipo temporale (incluse, oltre ai giorni, anche le fasce orarie della giornata, ad esempio 9-10, 10-11 e così via, oppure mattina e pomeriggio) e su prodotti (singoli o per categoria e/o per marca) e sui negozi. Allo scopo, specificare un possibile dettaglio del data mart; in particolare

1. **specificare esplicitamente la grana scelta**, supponendo che la quantità delle vendite sia tale che si è deciso di non utilizzare il massimo livello di dettaglio, ma solo quello strettamente indispensabile (in altri termini, la grana non deve essere il singolo dettaglio di vendita, ma una opportuna aggregazione);
2. mostrare gli schemi delle tabelle (tabella dei fatti e tabelle delle dimensioni) indicare anche le dimensioni i cui dati non provengono dalla base di dati sopra mostrata

Basi di dati II, primo modulo — Tecnologia delle basi di dati

24 settembre 2010 — Compito B
Cenni sulle soluzioni

Rispondere alle prime tre domande su questo foglio e alla quarta sul foglio separato.
Tempo a disposizione: un'ora e trenta minuti.

Cognome _____ **Nome** _____ **Matricola** _____ **Ordin.** _____

Domanda 1 (20%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Città)`. Spiegare perché, per la ricerca di tutti i clienti di una certa città, un indice secondario sull'attributo `Città` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).
Soluzione:
Se l'attributo `Città` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Città` l'indice non serve, perché, dato un certo valore per `Città` quasi tutti i blocchi contengono record con tale valore, allora una ricerca sequenziale ha un costo forse inferiore. Se invece ci sono 400 valori diversi, con l'indice si potrà accedere ai soli record di interesse, che sono il 10% del totale.

Domanda 2 (20%) Si consideri una relazione *R(A B C D E)*, in cui gli attributi hanno tutti la stessa dimensione d (ad esempio, ma è irrilevante, 4Byte), molto più piccola della dimensione del blocco pari a P. Si supponga che la relazione sia molto grande (L ennuple) e che le operazioni più frequenti su di essa siano le seguenti:

o_1 `SELECT * FROM R ORDER BY A`, con frequenza f_1

o_2 `SELECT A, B, C FROM R ORDER BY A`, con frequenza $f_2 = 10 \times f_1$

Valutare le due seguenti alternative di memorizzazione, calcolando il costo complessivo (riportare la formula che indica il numero di accessi nell'unità di tempo, in base alle variabili sopra citate):

(i) memorizzazione della relazione *R(A B C D E)* ordinata su A

costo unitario di o_1: $c_1 = L/(P/(5d)) = 5dL/P$ costo unitario di o_2: $c_2 = L/(P/(5d)) = 5dL/P$

costo complessivo: $c_1 f_1 + c_2 f_2 = L/(P/(5d)) = 55 f_1 a L/P$

(ii) memorizzazione delle proiezioni *R1(A B C)* e *R2(A D E)*, entrambe ordinate su A

costo unitario di o_1: $c_1 = 2L/(P/(3d)) = 6dL/P$ costo unitario di o_2: $c_2 = L/(P/(3d)) = 3dL/P$

costo complessivo: $c_1 f_1 + c_2 f_2 = 36 f_1 a L/P$

Supporre che il join possa essere eseguito con il metodo merge-join (e che il costo del join stesso sia trascurabile rispetto alle due scansioni).

Domanda 3 (20%) Si consideri una forma di equivalenza fra schedule denominata *final-state*-equivalenza, secondo la quale due schedule S_1 e S_2 sono equivalenti se, per ogni istanza i della base di dati, essi la trasformano nello stesso modo (e quindi $S_1(i) = S_2(i)$, se con $S(i)$ indichiamo l'istanza della base di dati ottenuta applicando a i le operazioni dello schedule S).

1. Formulare una definizione per questa proprietà, variante della definizione di view-equivalenza.
2. Spiegare, almeno intuitivamente, il rapporto che esiste fra final-state-equivalenza e view-equivalenza (sono equivalenti, incomparabili, oppure una implica l'altra?).

Soluzione:
Due schedule sono view-equivalenti se trasformano la base di dati nello stesso modo e offrono all'esterno gli stessi valori. La final-state-equivalenza non tiene conto di questa seconda proprietà. Gli schedule $w_1(x), r_2(x), w_3(x)$ e $w_1(x), w_3(x), r_2(x)$ sono final-state-equivalenti ma non view-equivalenti (perché hanno diversa relazione "legge-da"). Una definizione formale di final-state-equivalenza potrebbe richiedere (i) stesse scritture finali; (ii) stessa relazione "legge-da" per le transazioni che scrivono dopo la lettura. Di conseguenza, la final-state-equivalenza è condizione necessaria, ma non sufficiente per la view-equivalenza.

Domanda 4 (40%) Una catena di negozi gestisce le attività utilizzando, in ciascun negozio, una base di dati con le seguenti relazioni (su cui sono definiti gli ovvi vincoli di integrità referenziale):

- Prodotti(CodiceProdotto,Descrizione,Prezzo,Categoria,Produttore)
- Produttore(Codice,Nome)
- Categorie(Codice,Descrizione,MacroCategoria)
- MacroCategorie(Codice,Descrizione)
- Vendite(NumeroScontrino,Ora)
- DettaglioVendite(NumeroScontrino,CodiceProdotto,Quantità)

Si noti che

- Le informazioni relative alle vendite vengono mantenute solo nel corso della giornata.
- Il prezzo di un prodotto può variare da un giorno all'altro.

Utilizzando tali informazioni, la catena vuole realizzare un data mart relativo alle vendite dei prodotti, avente come misure le quantità vendute e gli incassi, che permetta di effettuare analisi di tipo temporale (incluse, oltre ai giorni, anche le fasce orarie della giornata, ad esempio 9-10, 10-11 e così via, oppure mattina e pomeriggio) e su prodotti (singoli o per categoria e/o per produttore) e sui negozi. Allo scopo, specificare un possibile dettaglio del data mart; in particolare

1. **specificare esplicitamente la grana scelta**, supponendo che la quantità delle vendite sia tale che si è deciso di non utilizzare il massimo livello di dettaglio, ma solo quello strettamente indispensabile (in altri termini, la grana non deve essere il singolo dettaglio di vendita, ma una opportuna aggregazione);
2. mostrare gli schemi delle tabelle (tabella dei fatti e tabelle delle dimensioni) indicare anche le dimensioni i cui dati non provengono dalla base di dati sopra mostrata