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

20 aprile 2009 — Compito A

Cenni sulle soluzioni

Tempo a disposizione: un'ora e quarantacinque minuti. Rispondere in modo ordinato evidenziando bene la "bella copia" e, in essa, le risposte alle singole domande. Consegnare comunque tutti i fogli protocollo e il testo.

Domanda 1 (15%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Zona)`. Spiegare perché, per la ricerca di tutti i clienti di una certa zona, un indice secondario sull'attributo `Zona` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).
Discussione
Se l'attributo `Zona` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Zona` l'indice non serve, perché, dato un certo valore per `Zona` 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 di interesse, che sono il 10% del totale.

Domanda 2 (10%) Indicare quali fra le seguenti affermazioni sono vere per i data warehouse:

1. Le attività supportate sono quelle quotidiane (ad esempio la vendita e le attività di sportello) *falso*
2. Gli utenti sono prevalentemente di livello più alto rispetto a quelli che utilizzano i sistemi OLTP *vero*
3. Le proprietà "acide" sono rilevanti *falso*
4. I dati sono una sintesi sempre aggiornata in tempo reale dei dati operativi *falso*
5. Le strutture fisiche cercano di conciliare le esigenze degli aggiornamenti e quelle delle interrogazioni *falso*
6. Le operazioni sono complesse e non predefinite *vero*

Domanda 3 (15%) Si consideri uno schema dimensionale utilizzato per analizzare le vendite in una catena di supermercati che, fra le dimensioni, ne preveda una sui negozi, come la seguente:

KNegozio	Nome	Città	Provincia	Regione
101	Pane e pasta	Olbia	SS	Sardegna
102	Bontà	Olbia	SS	Sardegna
103	Pane e vino	San Teodoro	NU	Sardegna
104	Vino e pane	Nuoro	NU	Sardegna
105	Pasta e pane	Palermo	PA	Sicilia
...	...	...	...	...

Si supponga ora che si presentino le seguenti esigenze di modifica:

- i negozi cambiano nome nel tempo: per esempio, il negozio nella prima ennupla potrebbe ad un certo punto cambiare nome da "Pane e pasta" in "Pane e non solo"; interessano selezioni e aggregazioni relative alle vendite tanto con riferimento al nome del negozio (nel momento specifico) quanto alla sua identità (caratterizzata talvolta da un identificatore e talvolta dal nome più recente); le modifiche sono rare, ma è possibile che ci siano negozi con vari cambiamenti di nome;

- sia pure molto raramente, le province cambiano; specificamente, si supponga che interessi gestire la modifica delle province della Sardegna avvenuta recentemente (ad esempio, dal giugno 2005 il comune di Olbia e quello di San Teodoro appartengono alla provincia OT, Olbia-Tempio); in questo caso, si supponga che non interessi tanto la correlazione fra data della vendita e provincia nel momento della vendita, quanto la possibilità di fare analisi con riferimento alle due versioni del territorio, quella con le vecchie province e quella con le nuove (N.B. supporre che non interessino altre versioni oltre a queste due).

Modificare la dimensione (mostrando la nuova versione per la tabella in figura, con brevi commenti se necessario).
Discussione La nuova tabella dimensione

KNeg	IDNeg	NomeAlMomento	NomePiùRecente	Città	ProvVecchia	ProvNuova	Regione
101	1	Pane e pasta	Pane e non solo	Olbia	SS	OT	Sardegna
102	2	Bontà	Bontà	Olbia	SS	OT	Sardegna
103	3	Pane e vino	Pane e vino	San Teodoro	NU	OT	Sardegna
104	4	Vino e pane	Vino e pane	Nuoro	NU	NU	Sardegna
105	5	Pasta e pane	Pasta e pane	Palermo	PA	PA	Sicilia
...	...	...	...	...	...	...	...
991	1	Pane e non solo	Pane e non solo	Olbia	SS	OT	Sardegna

Domanda 4 (10%) Con riferimento allo schema dimensionale citato nella domanda precedente si supponga che la seguente sia la struttura della tabella dei fatti, con alcune delle ennuple:

KNegozio	KProdotto	KData	Quantità	Incasso
101	201	301	243	2.350,32
101	202	301	4	32,00
101	202	302	6	49,00
102	201	301	22	262,00
...	...	...	...	...

Si supponga ora che:

- per ogni negozio, interessi rappresentare anche il direttore, per svolgere analisi sulle vendite ascrivibili al direttore stesso; i direttori cambiano nel tempo e passano da un negozio all'altro (e possono anche dirigere due negozi nello stesso momento; ma ogni negozio ha, in un certo giorno, un solo direttore); è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, di interesse).

Modificare la tabella dei fatti (discutendo in particolare come si potrebbero aggiornare le sue ennuple e osservando se e di quanto varia la sua cardinalità; mostrare anche la nuova versione della tabella in figura).
Discussione
La nuova tabella fatti ha la stessa cardinalità di quella originaria e un attributo in più, che corrisponde alla dimensionde degenere Direttore, che dipende funzionalmente dalle altre, in particolare da Negozio e Data. I valori del nuovo attributo sono quindi disponibili, come scritto nel testo: "è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, al quale è relativo il data mart)". Ad esempio, il direttore con KDirettore 501 era direttore del negozio 101 nella data 301.

KNegozio	KProdotto	KData	KDirettore	Quantità	Incasso
101	201	301	501	243	2.350,32
101	202	301	501	4	32,00
101	202	302	502	6	49,00
102	201	301	503	22	262,00
...	...	...	...	...	...

Domanda 5 (15%) Le seguenti situazioni corrispondono ad alcune delle note anomalie delle transazioni concorrenti. Indicare, per ciascuna di esse, a quale anomalia potrebbe corrispondere e quale livello di isolamento è opportuno per evitarla.

1. Un cliente consulta un calendario di concerti e ne individua uno che gli interessa. Quando chiede (con una seconda lettura) il biglietto, gli viene detto che il calendario è stato modificato e che quel concerto non esiste.
   Risposta: Lettura sporca oppure lettura inconsistente. Read committed o repeatable read
2. Un signore ha dieci mila Euro sul proprio conto corrente e firma due assegni da tre mila Euro ciascuno. I due beneficiari si presentano quasi contemporaneamente a due impiegati diversi della stessa banca, ciascuno dei quali verifica che i soldi sono disponibili (ci sono dieci mila Euro) e, pagato l'assegno, registra il nuovo saldo di sette mila Euro.
   Risposta: Perdita di aggiornamento. Repeatable read o serializable
3. Un appassionato lettore di gialli chiede quali libri di Agatha Christie siano disponibili. In risposta, riceve un elenco di tre libri. Chiede di ordinarli tutti e, quando apre il plico ricevuto, ne trova quattro. (Supporre qui che la transazione di richiesta faccia solo letture, in particolare, richieda due volte l'elenco dei libri di Agatha Christie).
   Risposta: Phantom. Serializable

Domanda 6 (15%) Specificare, con una breve giustificazione, a quali delle seguenti classi ciascuno degli schedule sotto mostrati appartiene: S (seriale), VSR (view-serializzabile), CSR (conflict-serializzabile), 2PL (generabile da uno scheduler basato sul lock a due fasi) and TS (generabile da uno scheduler che utilizzi il metodo dei timestamp; si assuma che gli identificatori delle transazioni corrispondano ai timestamp).

1. $r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$ *Risposta*: TS (e quindi CSR e VSR) e non 2PL
2. $r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$ *Risposta*: nessuna delle classi citate
3. $r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$ *Risposta*: TS e 2PL (e quindi CSR e VSR)
4. $r_2(x)w_2(x)r_1(x)w_1(x)$ *Risposta*: seriale (e quindi 2PL, CSR e VSR) ma non TS

Basi di dati II, primo modulo — Tecnologia delle basi di dati — 20 aprile 2009 — Compito A

Domanda 7 (20%) Un *quad tree*, nella sua versione più semplice, è un indice definito con riferimento a due diversi attributi di una relazione. Siano A e B gli attributi in questione. La struttura è costituita da un albero in cui ogni nodo intermedio ha due etichette, una per ciascuno degli attributi, e quattro sottoalberi. Se i valori di un nodo sono (a, b), allora

- il primo sottoalbero conterrà riferimenti ai record con valori di A minori o uguali ad a e valori di B minori o uguali a b
- il secondo ai record con valori di A minori o uguali ad a e valori di B maggiori di b
- il terzo ai record con valori di A maggiori di a e valori di B minori o uguali a b
- il quarto ai record con valori di A maggiori di a e valori di B maggiori di b

Intuitivamente, ogni nodo divide lo spazio in quattro parti e quindi una foglia di un quad tree contiene riferimenti corrispondenti ad un "quadratino" dello spazio bidimensionale, con valori di A compresi in un intervallo $(a_1, a_2]$ e quelli di B in un intervallo $(b_1, b_2]$; in contrasto, una foglia di un B+-tree definito su A, B (nell'ordine) contiene riferimenti "linearizzati," cioè ordinati per A e, a parità di A, per B.

Consideriamo ora una relazione di $N = 1.200.000$ di ennuple, in cui i due attributi, insieme, formino una chiave e in cui vi siano circa $v = 1.100$ valori diversi per ciascun attributo (abbastanza densi, ad esempio valori numerici compresi fra 1 e 1.200 e distibuiti omogeneamente). Supponiamo che le foglie contengano circa 100 riferimenti ciascuna (sia nel caso del quad tree sia in quello del B+-tree) e di avere un sistema che permetta di caricare nel buffer tutti i livelli intermedi (ma non le foglie) sia per i quad tree sia per i B+-tree. Calcolare in tale contesto il costo delle seguenti operazioni nel caso in cui sulla relazione è definito un quad tree e in quello in cui è definito un B+-tree su A, B:

1. conteggio delle ennuple con un dato valore di A
2. conteggio delle ennuple con un dato valore di B
3. ricerca della ennupla con valori dati per A e B
4. conteggio delle ennuple con valori di A e B compresi entrambi fra 100 e 115

Rispondere (sul foglio protocollo) riempiendo una tabella come la seguente, con qualche commento che spieghi la risposta:

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B				
B+-tree su A, B				

Possibile soluzione
Tutto dipende dall'organizzazione ordinata su A nel B+-tree vs "quadratino" nel quad tree. Data l'ipotesi sui buffer, è sufficiente contare le foglie coinvolte in ciascuna operazione

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B	circa 110	circa 110	2	circa 4
B+-tree su A, B	circa 11	12.000	2	circa 160

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

20 aprile 2009 — Compito B

Cenni sulle soluzioni

Tempo a disposizione: un'ora e quarantacinque minuti. Rispondere in modo ordinato evidenziando bene la "bella copia" e, in essa, le risposte alle singole domande. Consegnare comunque tutti i fogli protocollo e il testo.

Domanda 1 (15%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Classe)`. Spiegare perché, per la ricerca di tutti i clienti di una certa classe, un indice secondario sull'attributo `Classe` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).

Discussione

Se l'attributo `Classe` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Classe` l'indice non serve, perché, dato un certo valore per `Classe` 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 di interesse, che sono il 10% del totale.

Domanda 2 (10%) Indicare quali fra le seguenti affermazioni sono vere per i data warehouse:

1. Le proprietà "acide" sono rilevanti *falso*
2. I dati sono una sintesi sempre aggiornata in tempo reale dei dati operativi *falso*
3. Le strutture fisiche cercano di conciliare le esigenze degli aggiornamenti e quelle delle interrogazioni *falso*
4. Le operazioni sono complesse e non predefinite *vero*
5. Le attività supportate sono quelle quotidiane (ad esempio la vendita e le attività di sportello) *falso*
6. Gli utenti sono prevalentemente di livello più alto rispetto a quelli che utilizzano i sistemi OLTP *vero*

Domanda 3 (15%) Si consideri uno schema dimensionale utilizzato per analizzare le vendite in una catena di supermercati che, fra le dimensioni, ne preveda una sui negozi, come la seguente:

KNegozio	Nome	Città	Provincia	Regione
101	Pane e pasta	Tempio P.	SS	Sardegna
102	Bontà	Tempio P.	SS	Sardegna
103	Pane e vino	Budoni	NU	Sardegna
104	Vino e pane	Nuoro	NU	Sardegna
105	Pasta e pane	Palermo	PA	Sicilia
...	...	...	...	...

Si supponga ora che si presentino le seguenti esigenze di modifica:

- i negozi cambiano nome nel tempo: per esempio, il negozio nella prima ennupla potrebbe ad un certo punto cambiare nome da "Pane e pasta" in "Pane e non solo"; interessano selezioni e aggregazioni relative alle vendite tanto con riferimento al nome del negozio (nel momento specifico) quanto alla sua identità (caratterizzata talvolta da un identificatore e talvolta dal nome più recente); le modifiche sono rare, ma è possibile che ci siano negozi con vari cambiamenti di nome;

- sia pure molto raramente, le province cambiano; specificamente, si supponga che interessi gestire la modifica delle province della Sardegna avvenuta recentemente (ad esempio, dal giugno 2005 il comune di Tempio P. e quello di Budoni appartengono alla provincia OT, Olbia-Tempio); in questo caso, si supponga che non interessi tanto la correlazione fra data della vendita e provincia nel momento della vendita, quanto la possibilità di fare analisi con riferimento alle due versioni del territorio, quella con le vecchie province e quella con le nuove (N.B. supporre che non interessino altre versioni oltre a queste due).

Modificare la dimensione (mostrando la nuova versione per la tabella in figura, con brevi commenti se necessario).

Discussione La nuova tabella dimensione

KNeg	IDNeg	NomeAlMomento	NomePiùRecente	Città	ProvVecchia	ProvNuova	Regione
101	1	Pane e pasta	Pane e non solo	Tempio P.	SS	OT	Sardegna
102	2	Bontà	Bontà	Tempio P.	SS	OT	Sardegna
103	3	Pane e vino	Pane e vino	Budoni	NU	OT	Sardegna
104	4	Vino e pane	Vino e pane	Nuoro	NU	NU	Sardegna
105	5	Pasta e pane	Pasta e pane	Palermo	PA	PA	Sicilia
...	...	...	...	...	...	...	...
991	1	Pane e non solo	Pane e non solo	Tempio P.	SS	OT	Sardegna

Basi di dati II, primo modulo — Tecnologia delle basi di dati — 20 aprile 2009 — Compito B

Domanda 4 (10%) Con riferimento allo schema dimensionale citato nella domanda precedente si supponga che la seguente sia la struttura della tabella dei fatti, con alcune delle ennuple:

KNegozio	KProdotto	KData	Quantità	Incasso
101	201	301	243	1.250,32
101	202	301	4	32,00
101	202	302	6	49,00
102	201	301	22	262,00
...	...	...	...	...

Si supponga ora che:

- per ogni negozio, interessi rappresentare anche il direttore, per svolgere analisi sulle vendite ascrivibili al direttore stesso; i direttori cambiano nel tempo e passano da un negozio all'altro (e possono anche dirigere due negozi nello stesso momento; ma ogni negozio ha, in un certo giorno, un solo direttore); è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, di interesse).

Modificare la tabella dei fatti (discutendo in particolare come si potrebbero aggiornare le sue ennuple e osservando se e di quanto varia la sua cardinalità; mostrare anche la nuova versione della tabella in figura).
Discussione
La nuova tabella fatti ha la stessa cardinalità di quella originaria e un attributo in più, che corrisponde alla dimensionde degenere Direttore, che dipende funzionalmente dalle altre, in particolare da Negozio e Data. I valori del nuovo attributo sono quindi disponibili, come scritto nel testo: "è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, al quale è relativo il data mart)". Ad esempio, il direttore con KDirettore 501 era direttore del negozio 101 nella data 301.

KNegozio	KProdotto	KData	KDirettore	Quantità	Incasso
101	201	301	501	243	1.250,32
101	202	301	501	4	32,00
101	202	302	502	6	49,00
102	201	301	503	22	262,00
...	...	...	...	...	...

Domanda 5 (15%) Le seguenti situazioni corrispondono ad alcune delle note anomalie delle transazioni concorrenti. Indicare, per ciascuna di esse, a quale anomalia potrebbe corrispondere e quale livello di isolamento è opportuno per evitarla.

1. Un signore ha dieci mila Euro sul proprio conto corrente e firma due assegni da tre mila Euro ciascuno. I due beneficiari si presentano quasi contemporaneamente a due impiegati diversi della stessa banca, ciascuno dei quali verifica che i soldi sono disponibili (ci sono dieci mila Euro) e, pagato l'assegno, registra il nuovo saldo di sette mila Euro.
   Risposta: Perdita di aggiornamento. Repeatable read o serializable
2. Un appassionato lettore di gialli chiede quali libri di Agatha Christie siano disponibili. In risposta, riceve un elenco di tre libri. Chiede di ordinarli tutti e, quando apre il plico ricevuto, ne trova quattro. (Supporre qui che la transazione di richiesta faccia solo letture, in particolare, richieda due volte l'elenco dei libri di Agatha Christie).
   Risposta: Phantom. Serializable
3. Un cliente consulta un calendario di concerti e ne individua uno che gli interessa. Quando chiede (con una seconda lettura) il biglietto, gli viene detto che il calendario è stato modificato e che quel concerto non esiste.
   Risposta: Lettura sporca oppure lettura inconsistente. Read committed o repeatable read

Domanda 6 (15%) Specificare, con una breve giustificazione, a quali delle seguenti classi ciascuno degli schedule sotto mostrati appartiene: S (seriale), VSR (view-serializzabile), CSR (conflict-serializzabile), 2PL (generabile da uno scheduler basato sul lock a due fasi) and TS (generabile da uno scheduler che utilizzi il metodo dei timestamp; si assuma che gli identificatori delle transazioni corrispondano ai timestamp).

1. $r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$ *Risposta*: nessuna delle classi citate
2. $r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$ *Risposta*: TS e 2PL (e quindi CSR e VSR)
3. $r_2(x)w_2(x)r_1(x)w_1(x)$ *Risposta*: seriale (e quindi 2PL, CSR e VSR) ma non TS
4. $r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$ *Risposta*: TS (e quindi CSR e VSR) e non 2PL

Basi di dati II, primo modulo — Tecnologia delle basi di dati — 20 aprile 2009 — Compito B

Domanda 7 (20%) Un *quad tree*, nella sua versione più semplice, è un indice definito con riferimento a due diversi attributi di una relazione. Siano A e B gli attributi in questione. La struttura è costituita da un albero in cui ogni nodo intermedio ha due etichette, una per ciascuno degli attributi, e quattro sottoalberi. Se i valori di un nodo sono (a, b), allora

- il primo sottoalbero conterrà riferimenti ai record con valori di A minori o uguali ad a e valori di B minori o uguali a b
- il secondo ai record con valori di A minori o uguali ad a e valori di B maggiori di b
- il terzo ai record con valori di A maggiori di a e valori di B minori o uguali a b
- il quarto ai record con valori di A maggiori di a e valori di B maggiori di b

Intuitivamente, ogni nodo divide lo spazio in quattro parti e quindi una foglia di un quad tree contiene riferimenti corrispondenti ad un "quadratino" dello spazio bidimensionale, con valori di A compresi in un intervallo $(a_1, a_2]$ e quelli di B in un intervallo $(b_1, b_2]$; in contrasto, una foglia di un B+-tree definito su A, B (nell'ordine) contiene riferimenti "linearizzati," cioè ordinati per A e, a parità di A, per B.

Consideriamo ora una relazione di $N = 1.200.000$ di ennuple, in cui i due attributi, insieme, formino una chiave e in cui vi siano circa $v = 1.100$ valori diversi per ciascun attributo (abbastanza densi, ad esempio valori numerici compresi fra 1 e 1.200 e distibuiti omogeneamente). Supponiamo che le foglie contengano circa 100 riferimenti ciascuna (sia nel caso del quad tree sia in quello del B+-tree) e di avere un sistema che permetta di caricare nel buffer tutti i livelli intermedi (ma non le foglie) sia per i quad tree sia per i B+-tree. Calcolare in tale contesto il costo delle seguenti operazioni nel caso in cui sulla relazione è definito un quad tree e in quello in cui è definito un B+-tree su A, B:

1. conteggio delle ennuple con un dato valore di A
2. conteggio delle ennuple con un dato valore di B
3. ricerca della ennupla con valori dati per A e B
4. conteggio delle ennuple con valori di A e B compresi entrambi fra 100 e 115

Rispondere (sul foglio protocollo) riempiendo una tabella come la seguente, con qualche commento che spieghi la risposta:

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B				
B+-tree su A, B				

Possibile soluzione
Tutto dipende dall'organizzazione ordinata su A nel B+-tree vs "quadratino" nel quad tree. Data l'ipotesi sui buffer, è sufficiente contare le foglie coinvolte in ciascuna operazione

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B	circa 110	circa 110	2	circa 4
B+-tree su A, B	circa 11	12.000	2	circa 160

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

20 aprile 2009 — Compito C

Cenni sulle soluzioni

Tempo a disposizione: un'ora e quarantacinque minuti. Rispondere in modo ordinato evidenziando bene la "bella copia" e, in essa, le risposte alle singole domande. Consegnare comunque tutti i fogli protocollo e il testo.

Domanda 1 (15%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Categoria)`. Spiegare perché, per la ricerca di tutti i clienti di una certa categoria, un indice secondario sull'attributo `Categoria` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).

Discussione

Se l'attributo `Categoria` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Categoria` l'indice non serve, perché, dato un certo valore per `Categoria` 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 di interesse, che sono il 10% del totale.

Domanda 2 (10%) Indicare quali fra le seguenti affermazioni sono vere per i data warehouse:

1. I dati sono una sintesi sempre aggiornata in tempo reale dei dati operativi *falso*
2. Le strutture fisiche cercano di conciliare le esigenze degli aggiornamenti e quelle delle interrogazioni *falso*
3. Le operazioni sono complesse e non predefinite *vero*
4. Le attività supportate sono quelle quotidiane (ad esempio la vendita e le attività di sportello) *falso*
5. Gli utenti sono prevalentemente di livello più alto rispetto a quelli che utilizzano i sistemi OLTP *vero*
6. Le proprietà "acide" sono rilevanti *falso*

Domanda 3 (15%) Si consideri uno schema dimensionale utilizzato per analizzare le vendite in una catena di supermercati che, fra le dimensioni, ne preveda una sui negozi, come la seguente:

KNegozio	Nome	Città	Provincia	Regione
101	Pane e pasta	Olbia	SS	Sardegna
102	Bontà	Olbia	SS	Sardegna
103	Pane e vino	Budoni	NU	Sardegna
104	Vino e pane	Nuoro	NU	Sardegna
105	Pasta e pane	Palermo	PA	Sicilia
...	...	...	...	...

Si supponga ora che si presentino le seguenti esigenze di modifica:

- i negozi cambiano nome nel tempo: per esempio, il negozio nella prima ennupla potrebbe ad un certo punto cambiare nome da "Pane e pasta" in "Pane e non solo"; interessano selezioni e aggregazioni relative alle vendite tanto con riferimento al nome del negozio (nel momento specifico) quanto alla sua identità (caratterizzata talvolta da un identificatore e talvolta dal nome più recente); le modifiche sono rare, ma è possibile che ci siano negozi con vari cambiamenti di nome;

- sia pure molto raramente, le province cambiano; specificamente, si supponga che interessi gestire la modifica delle province della Sardegna avvenuta recentemente (ad esempio, dal giugno 2005 il comune di Olbia e quello di Budoni appartengono alla provincia OT, Olbia-Tempio); in questo caso, si supponga che non interessi tanto la correlazione fra data della vendita e provincia nel momento della vendita, quanto la possibilità di fare analisi con riferimento alle due versioni del territorio, quella con le vecchie province e quella con le nuove (N.B. supporre che non interessino altre versioni oltre a queste due).

Modificare la dimensione (mostrando la nuova versione per la tabella in figura, con brevi commenti se necessario).

Discussione La nuova tabella dimensione

KNeg	IDNeg	NomeAlMomento	NomePiùRecente	Città	ProvVecchia	ProvNuova	Regione
101	1	Pane e pasta	Pane e non solo	Olbia	SS	OT	Sardegna
102	2	Bontà	Bontà	Olbia	SS	OT	Sardegna
103	3	Pane e vino	Pane e vino	Budoni	NU	OT	Sardegna
104	4	Vino e pane	Vino e pane	Nuoro	NU	NU	Sardegna
105	5	Pasta e pane	Pasta e pane	Palermo	PA	PA	Sicilia
...	...	...	...	...	...	...	...
991	1	Pane e non solo	Pane e non solo	Olbia	SS	OT	Sardegna

Domanda 4 (10%) Con riferimento allo schema dimensionale citato nella domanda precedente si supponga che la seguente sia la struttura della tabella dei fatti, con alcune delle ennuple:

KNegozio	KProdotto	KData	Quantità	Incasso
101	201	301	243	2.350,32
101	202	301	4	32,00
101	202	302	6	49,00
102	201	301	22	262,00
...	...	...	...	...

Si supponga ora che:

- per ogni negozio, interessi rappresentare anche il direttore, per svolgere analisi sulle vendite ascrivibili al direttore stesso; i direttori cambiano nel tempo e passano da un negozio all'altro (e possono anche dirigere due negozi nello stesso momento; ma ogni negozio ha, in un certo giorno, un solo direttore); è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, di interesse).

Modificare la tabella dei fatti (discutendo in particolare come si potrebbero aggiornare le sue ennuple e osservando se e di quanto varia la sua cardinalità; mostrare anche la nuova versione della tabella in figura).
Discussione
La nuova tabella fatti ha la stessa cardinalità di quella originaria e un attributo in più, che corrisponde alla dimensionde degenere Direttore, che dipende funzionalmente dalle altre, in particolare da Negozio e Data. I valori del nuovo attributo sono quindi disponibili, come scritto nel testo: "è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, al quale è relativo il data mart)". Ad esempio, il direttore con KDirettore 501 era direttore del negozio 101 nella data 301.

KNegozio	KProdotto	KData	KDirettore	Quantità	Incasso
101	201	301	501	243	2.350,32
101	202	301	501	4	32,00
101	202	302	502	6	49,00
102	201	301	503	22	262,00
...	...	...	...	...	...

Domanda 5 (15%) Le seguenti situazioni corrispondono ad alcune delle note anomalie delle transazioni concorrenti. Indicare, per ciascuna di esse, a quale anomalia potrebbe corrispondere e quale livello di isolamento è opportuno per evitarla.

1. Un signore ha dieci mila Euro sul proprio conto corrente e firma due assegni da tre mila Euro ciascuno. I due beneficiari si presentano quasi contemporaneamente a due impiegati diversi della stessa banca, ciascuno dei quali verifica che i soldi sono disponibili (ci sono dieci mila Euro) e, pagato l'assegno, registra il nuovo saldo di sette mila Euro.
   Risposta: Perdita di aggiornamento. Repeatable read o serializable
2. Un cliente consulta un calendario di concerti e ne individua uno che gli interessa. Quando chiede (con una seconda lettura) il biglietto, gli viene detto che il calendario è stato modificato e che quel concerto non esiste.
   Risposta: Lettura sporca oppure lettura inconsistente. Read committed o repeatable read
3. Un appassionato lettore di gialli chiede quali libri di Agatha Christie siano disponibili. In risposta, riceve un elenco di tre libri. Chiede di ordinarli tutti e, quando apre il plico ricevuto, ne trova quattro. (Supporre qui che la transazione di richiesta faccia solo letture, in particolare, richieda due volte l'elenco dei libri di Agatha Christie).
   Risposta: Phantom. Serializable

Domanda 6 (15%) Specificare, con una breve giustificazione, a quali delle seguenti classi ciascuno degli schedule sotto mostrati appartiene: S (seriale), VSR (view-serializzabile), CSR (conflict-serializzabile), 2PL (generabile da uno scheduler basato sul lock a due fasi) and TS (generabile da uno scheduler che utilizzi il metodo dei timestamp; si assuma che gli identificatori delle transazioni corrispondano ai timestamp).

1. $r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$ *Risposta*: TS e 2PL (e quindi CSR e VSR)
2. $r_2(x)w_2(x)r_1(x)w_1(x)$ *Risposta*: seriale (e quindi 2PL, CSR e VSR) ma non TS
3. $r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$ *Risposta*: TS (e quindi CSR e VSR) e non 2PL
4. $r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$ *Risposta*: nessuna delle classi citate

Basi di dati II, primo modulo — Tecnologia delle basi di dati — 20 aprile 2009 — Compito C

Domanda 7 (20%) Un *quad tree*, nella sua versione più semplice, è un indice definito con riferimento a due diversi attributi di una relazione. Siano A e B gli attributi in questione. La struttura è costituita da un albero in cui ogni nodo intermedio ha due etichette, una per ciascuno degli attributi, e quattro sottoalberi. Se i valori di un nodo sono (a, b), allora

- il primo sottoalbero conterrà riferimenti ai record con valori di A minori o uguali ad a e valori di B minori o uguali a b
- il secondo ai record con valori di A minori o uguali ad a e valori di B maggiori di b
- il terzo ai record con valori di A maggiori di a e valori di B minori o uguali a b
- il quarto ai record con valori di A maggiori di a e valori di B maggiori di b

Intuitivamente, ogni nodo divide lo spazio in quattro parti e quindi una foglia di un quad tree contiene riferimenti corrispondenti ad un "quadratino" dello spazio bidimensionale, con valori di A compresi in un intervallo $(a_1, a_2]$ e quelli di B in un intervallo $(b_1, b_2]$; in contrasto, una foglia di un B+-tree definito su A, B (nell'ordine) contiene riferimenti "linearizzati," cioè ordinati per A e, a parità di A, per B.

Consideriamo ora una relazione di $N = 1.200.000$ di ennuple, in cui i due attributi, insieme, formino una chiave e in cui vi siano circa $v = 1.100$ valori diversi per ciascun attributo (abbastanza densi, ad esempio valori numerici compresi fra 1 e 1.200 e distibuiti omogeneamente). Supponiamo che le foglie contengano circa 100 riferimenti ciascuna (sia nel caso del quad tree sia in quello del B+-tree) e di avere un sistema che permetta di caricare nel buffer tutti i livelli intermedi (ma non le foglie) sia per i quad tree sia per i B+-tree. Calcolare in tale contesto il costo delle seguenti operazioni nel caso in cui sulla relazione è definito un quad tree e in quello in cui è definito un B+-tree su A, B:

1. conteggio delle ennuple con un dato valore di A
2. conteggio delle ennuple con un dato valore di B
3. ricerca della ennupla con valori dati per A e B
4. conteggio delle ennuple con valori di A e B compresi entrambi fra 100 e 115

Rispondere (sul foglio protocollo) riempiendo una tabella come la seguente, con qualche commento che spieghi la risposta:

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B				
B+-tree su A, B				

Possibile soluzione
Tutto dipende dall'organizzazione ordinata su A nel B+-tree vs "quadratino" nel quad tree. Data l'ipotesi sui buffer, è sufficiente contare le foglie coinvolte in ciascuna operazione

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B	circa 110	circa 110	2	circa 4
B+-tree su A, B	circa 11	12.000	2	circa 160

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

20 aprile 2009 — Compito D

Cenni sulle soluzioni

Tempo a disposizione: un'ora e quarantacinque minuti. Rispondere in modo ordinato evidenziando bene la "bella copia" e, in essa, le risposte alle singole domande. Consegnare comunque tutti i fogli protocollo e il testo.

Domanda 1 (15%) Si consideri una relazione `R(CodiceCliente,Cognome,Nome,Categoria)`. Spiegare perché, per la ricerca di tutti i clienti di una certa categoria, un indice secondario sull'attributo `Categoria` potrebbe in alcuni casi risultare conveniente e in altri no (mostrare anche esempi che illustrino le due situazioni da un punto di vista quantitativo).

Discussione

Se l'attributo `Categoria` è selettivo, allora l'indice conviene altrimenti no. Ad esempio, se il fattore di blocco è 40 e ci sono 20 valori diversi per `Categoria` l'indice non serve, perché, dato un certo valore per `Categoria` 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 di interesse, che sono il 10% del totale.

Domanda 2 (10%) Indicare quali fra le seguenti affermazioni sono vere per i data warehouse:

1. Le operazioni sono complesse e non predefinite *vero*
2. Le attività supportate sono quelle quotidiane (ad esempio la vendita e le attività di sportello) *falso*
3. Gli utenti sono prevalentemente di livello più alto rispetto a quelli che utilizzano i sistemi OLTP *vero*
4. Le proprietà "acide" sono rilevanti *falso*
5. I dati sono una sintesi sempre aggiornata in tempo reale dei dati operativi *falso*
6. Le strutture fisiche cercano di conciliare le esigenze degli aggiornamenti e quelle delle interrogazioni *falso*

Domanda 3 (15%) Si consideri uno schema dimensionale utilizzato per analizzare le vendite in una catena di supermercati che, fra le dimensioni, ne preveda una sui negozi, come la seguente:

KNegozio	Nome	Città	Provincia	Regione
101	Pane e pasta	Tempio P.	SS	Sardegna
102	Bontà	Tempio P.	SS	Sardegna
103	Pane e vino	San Teodoro	NU	Sardegna
104	Vino e pane	Nuoro	NU	Sardegna
105	Pasta e pane	Palermo	PA	Sicilia
...	...	...	...	...

Si supponga ora che si presentino le seguenti esigenze di modifica:

- i negozi cambiano nome nel tempo: per esempio, il negozio nella prima ennupla potrebbe ad un certo punto cambiare nome da "Pane e pasta" in "Pane e non solo"; interessano selezioni e aggregazioni relative alle vendite tanto con riferimento al nome del negozio (nel momento specifico) quanto alla sua identità (caratterizzata talvolta da un identificatore e talvolta dal nome più recente); le modifiche sono rare, ma è possibile che ci siano negozi con vari cambiamenti di nome;

- sia pure molto raramente, le province cambiano; specificamente, si supponga che interessi gestire la modifica delle province della Sardegna avvenuta recentemente (ad esempio, dal giugno 2005 il comune di Tempio P. e quello di San Teodoro appartengono alla provincia OT, Olbia-Tempio); in questo caso, si supponga che non interessi tanto la correlazione fra data della vendita e provincia nel momento della vendita, quanto la possibilità di fare analisi con riferimento alle due versioni del territorio, quella con le vecchie province e quella con le nuove (N.B. supporre che non interessino altre versioni oltre a queste due).

Modificare la dimensione (mostrando la nuova versione per la tabella in figura, con brevi commenti se necessario).

Discussione La nuova tabella dimensione

KNeg	IDNeg	NomeAlMomento	NomePiùRecente	Città	ProvVecchia	ProvNuova	Regione
101	1	Pane e pasta	Pane e non solo	Tempio P.	SS	OT	Sardegna
102	2	Bontà	Bontà	Tempio P.	SS	OT	Sardegna
103	3	Pane e vino	Pane e vino	San Teodoro	NU	OT	Sardegna
104	4	Vino e pane	Vino e pane	Nuoro	NU	NU	Sardegna
105	5	Pasta e pane	Pasta e pane	Palermo	PA	PA	Sicilia
...	...	...	...	...	...	...	...
991	1	Pane e non solo	Pane e non solo	Tempio P.	SS	OT	Sardegna

Domanda 4 (10%) Con riferimento allo schema dimensionale citato nella domanda precedente si supponga che la seguente sia la struttura della tabella dei fatti, con alcune delle ennuple:

KNegozio	KProdotto	KData	Quantità	Incasso
101	201	301	243	1.250,32
101	202	301	4	32,00
101	202	302	6	49,00
102	201	301	22	262,00
...	...	...	...	...

Si supponga ora che:

- per ogni negozio, interessi rappresentare anche il direttore, per svolgere analisi sulle vendite ascrivibili al direttore stesso; i direttori cambiano nel tempo e passano da un negozio all'altro (e possono anche dirigere due negozi nello stesso momento; ma ogni negozio ha, in un certo giorno, un solo direttore); è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, di interesse).

Modificare la tabella dei fatti (discutendo in particolare come si potrebbero aggiornare le sue ennuple e osservando se e di quanto varia la sua cardinalità; mostrare anche la nuova versione della tabella in figura).
Discussione
La nuova tabella fatti ha la stessa cardinalità di quella originaria e un attributo in più, che corrisponde alla dimensionde degenere Direttore, che dipende funzionalmente dalle altre, in particolare da Negozio e Data. I valori del nuovo attributo sono quindi disponibili, come scritto nel testo: "è disponibile l'informazione relativa ai direttori dei negozi nel tempo (per tutto il periodo, anche passato, al quale è relativo il data mart)". Ad esempio, il direttore con KDirettore 501 era direttore del negozio 101 nella data 301.

KNegozio	KProdotto	KData	KDirettore	Quantità	Incasso
101	201	301	501	243	1.250,32
101	202	301	501	4	32,00
101	202	302	502	6	49,00
102	201	301	503	22	262,00
...	...	...	...	...	...

Domanda 5 (15%) Le seguenti situazioni corrispondono ad alcune delle note anomalie delle transazioni concorrenti. Indicare, per ciascuna di esse, a quale anomalia potrebbe corrispondere e quale livello di isolamento è opportuno per evitarla.

1. Un appassionato lettore di gialli chiede quali libri di Agatha Christie siano disponibili. In risposta, riceve un elenco di tre libri. Chiede di ordinarli tutti e, quando apre il plico ricevuto, ne trova quattro. (Supporre qui che la transazione di richiesta faccia solo letture, in particolare, richieda due volte l'elenco dei libri di Agatha Christie).
   Risposta: Phantom. Serializable
2. Un signore ha dieci mila Euro sul proprio conto corrente e firma due assegni da tre mila Euro ciascuno. I due beneficiari si presentano quasi contemporaneamente a due impiegati diversi della stessa banca, ciascuno dei quali verifica che i soldi sono disponibili (ci sono dieci mila Euro) e, pagato l'assegno, registra il nuovo saldo di sette mila Euro.
   Risposta: Perdita di aggiornamento. Repeatable read o serializable
3. Un cliente consulta un calendario di concerti e ne individua uno che gli interessa. Quando chiede (con una seconda lettura) il biglietto, gli viene detto che il calendario è stato modificato e che quel concerto non esiste.
   Risposta: Lettura sporca oppure lettura inconsistente. Read committed o repeatable read

Domanda 6 (15%) Specificare, con una breve giustificazione, a quali delle seguenti classi ciascuno degli schedule sotto mostrati appartiene: S (seriale), VSR (view-serializzabile), CSR (conflict-serializzabile), 2PL (generabile da uno scheduler basato sul lock a due fasi) and TS (generabile da uno scheduler che utilizzi il metodo dei timestamp; si assuma che gli identificatori delle transazioni corrispondano ai timestamp).

1. $r_2(x)w_2(x)r_1(x)w_1(x)$ *Risposta*: seriale (e quindi 2PL, CSR e VSR) ma non TS
2. $r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$ *Risposta*: TS (e quindi CSR e VSR) e non 2PL
3. $r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$ *Risposta*: nessuna delle classi citate
4. $r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$ *Risposta*: TS e 2PL (e quindi CSR e VSR)

Basi di dati II, primo modulo — Tecnologia delle basi di dati — 20 aprile 2009 — Compito D

Domanda 7 (20%) Un *quad tree*, nella sua versione più semplice, è un indice definito con riferimento a due diversi attributi di una relazione. Siano A e B gli attributi in questione. La struttura è costituita da un albero in cui ogni nodo intermedio ha due etichette, una per ciascuno degli attributi, e quattro sottoalberi. Se i valori di un nodo sono (a, b), allora

- il primo sottoalbero conterrà riferimenti ai record con valori di A minori o uguali ad a e valori di B minori o uguali a b
- il secondo ai record con valori di A minori o uguali ad a e valori di B maggiori di b
- il terzo ai record con valori di A maggiori di a e valori di B minori o uguali a b
- il quarto ai record con valori di A maggiori di a e valori di B maggiori di b

Intuitivamente, ogni nodo divide lo spazio in quattro parti e quindi una foglia di un quad tree contiene riferimenti corrispondenti ad un "quadratino" dello spazio bidimensionale, con valori di A compresi in un intervallo $(a_1, a_2]$ e quelli di B in un intervallo $(b_1, b_2]$; in contrasto, una foglia di un B+-tree definito su A, B (nell'ordine) contiene riferimenti "linearizzati," cioè ordinati per A e, a parità di A, per B.

Consideriamo ora una relazione di $N = 1.200.000$ di ennuple, in cui i due attributi, insieme, formino una chiave e in cui vi siano circa $v = 1.100$ valori diversi per ciascun attributo (abbastanza densi, ad esempio valori numerici compresi fra 1 e 1.200 e distibuiti omogeneamente). Supponiamo che le foglie contengano circa 100 riferimenti ciascuna (sia nel caso del quad tree sia in quello del B+-tree) e di avere un sistema che permetta di caricare nel buffer tutti i livelli intermedi (ma non le foglie) sia per i quad tree sia per i B+-tree. Calcolare in tale contesto il costo delle seguenti operazioni nel caso in cui sulla relazione è definito un quad tree e in quello in cui è definito un B+-tree su A, B:

1. conteggio delle ennuple con un dato valore di A
2. conteggio delle ennuple con un dato valore di B
3. ricerca della ennupla con valori dati per A e B
4. conteggio delle ennuple con valori di A e B compresi entrambi fra 100 e 115

Rispondere (sul foglio protocollo) riempiendo una tabella come la seguente, con qualche commento che spieghi la risposta:

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B				
B+-tree su A, B				

Possibile soluzione
Tutto dipende dall'organizzazione ordinata su A nel B+-tree vs "quadratino" nel quad tree. Data l'ipotesi sui buffer, è sufficiente contare le foglie coinvolte in ciascuna operazione

	conteggio puntuale su A	conteggio puntuale su B	accesso puntuale su A, B	conteggio in intervallo su A, B
quad tree su A, B	circa 110	circa 110	2	circa 4
B+-tree su A, B	circa 11	12.000	2	circa 160