

**Basi di dati II, primo modulo — prova breve**  
**22 giugno 2010**

Cognome \_\_\_\_\_ Nome \_\_\_\_\_ Matricola \_\_\_\_\_

Rispondere in modo ordinato su un foglio protocollo diverso da quello utilizzato per le risposte alle domande relative al secondo modulo. Nell'ambito di questa prova, le domande sul primo modulo avranno un peso del 30% e quelle sul secondo del 70%. Qui (come nell'altro foglio con le tracce) sono indicati i pesi relativi.

**Domanda 1** (40%) (Rispondere sul foglio protocollo) Rispondere a questa domanda **dopo** avere risposto alla domanda 1 relativa al secondo modulo.

Illustrare brevemente una alternativa di memorizzazione che favorisca interrogazioni volte a conoscere solo codici fiscali (**cf**) dei clienti e costi dei noleggi. Fornire una breve spiegazione, con riferimento alle problematiche di gestione dei dati in memoria secondaria.

**Domanda 2** (60%) (Rispondere sul foglio protocollo) Illustrare brevemente, ma con chiarezza, gli aspetti salienti delle trasformazioni utilizzate per alimentare uno degli schemi dimensionali utilizzati nel progetto a partire dalla base di dati sorgente fornita.

**Basi di dati II, primo modulo — prova lunga**  
**22 giugno 2010 — Compito A**

Cognome \_\_\_\_\_ Nome \_\_\_\_\_ Matricola \_\_\_\_\_

Rispondere alle ultime due domande su un foglio protocollo diverso da quello utilizzato per le risposte alle domande relative al secondo modulo. Rispondere su questo foglio alle altre domande.

Nell'ambito di questa prova, le domande sul primo modulo avranno un peso del 55% e quelle sul secondo del 45%. Qui (come nell'altro foglio con le tracce) sono indicati i pesi relativi.

**Domanda 1** (25%) Sia data una relazione  $R(A, B, C)$  contenente circa  $L = 10.000.000$  ennuple di  $r = 20$  byte ciascuna, di cui  $a = 4$  per la chiave  $A$ , che contiene valori interi quasi consecutivi, da 1 a poco più di 10.000.000. Supporre che i blocchi abbiano dimensione  $B = 2\text{KB}$ , approssimabile come 2.000, che i puntatori ai record abbiano lunghezza  $p = 6$ ; e che i nodi intermedi degli indici possano essere contenuti nei buffer.

In ciascuno dei seguenti casi:

- (a) indice primario (sparso) su  $A$  realizzato con B+-tree;
- (b) indice secondario su  $A$  realizzato con B+-tree;
- (c) struttura primaria hash su  $A$ .

indicare il costo prevedibile per le seguenti operazioni

1. `SELECT * FROM R WHERE A >= 1000 AND A <=3000`
2. `SELECT COUNT(*) FROM R WHERE A >= 1000 AND A <=3000`
3. `SELECT * FROM R WHERE A = 2000`

Riportare le risposte nella tabella sottostante, indicando formula e valore numerico (con brevissimo commento, se necessario)

	Risposte
(a) 1	
(b) 1	
(c) 1	
(a) 2	
(b) 2	
(c) 2	
(a) 3	
(b) 3	
(c) 3	

**Domanda 2** (20%) Specificare, per ciascuno dei seguenti schedule, la sua appartenenza alle varie classi (scrivere “sì” o “no” nelle caselle della tabella; assumere che gli identificatori delle transazioni corrispondano ai timestamp).

	Seriale	VSR	CSR	TS	2PL
$r_2(x)w_2(x)r_1(x)w_1(x)$					
$r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$					
$r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$					
$r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$					

**Domanda 3** (15%) Spiegare perché, scrivendo un programma che accede a due basi di dati diverse utilizzando JDBC, non è possibile garantire il commit a due fasi. Indicare quali funzionalità aggiuntive sono necessarie per i server locali che vogliono realizzare tale servizio e come si può procedere con un sistema che non disponga di tali funzionalità.

**Domanda 4** (15%) (Rispondere sul foglio protocollo) Rispondere a questa domanda **dopo** avere risposto alla domanda 1 relativa al secondo modulo.

Illustrare brevemente una alternativa di memorizzazione che favorisca interrogazioni volte a conoscere solo codici fiscali (cf) dei clienti e costi dei noleggi. Fornire una breve spiegazione, con riferimento alle problematiche di gestione dei dati in memoria secondaria.

**Domanda 5** (25%) (Rispondere sul foglio protocollo) Illustrare brevemente, ma con chiarezza, gli aspetti salienti delle trasformazioni utilizzate per alimentare uno degli schemi dimensionali utilizzati nel progetto a partire dalla base di dati sorgente fornita.

**Basi di dati II, primo modulo — prova lunga**  
**22 giugno 2010 — Compito B**

Cognome \_\_\_\_\_ Nome \_\_\_\_\_ Matricola \_\_\_\_\_

Rispondere alle ultime due domande su un foglio protocollo diverso da quello utilizzato per le risposte alle domande relative al secondo modulo. Rispondere su questo foglio alle altre domande.

Nell'ambito di questa prova, le domande sul primo modulo avranno un peso del 55% e quelle sul secondo del 45%. Qui (come nell'altro foglio con le tracce) sono indicati i pesi relativi.

**Domanda 1** (25%) Sia data una relazione  $R(A, B, C)$  contenente circa  $L = 10.000.000$  ennuple di  $r = 20$  byte ciascuna, di cui  $a = 4$  per la chiave  $A$ , che contiene valori interi quasi consecutivi, da 1 a poco più di 10.000.000. Supporre che i blocchi abbiano dimensione  $B = 2\text{KB}$ , approssimabile come 2.000, che i puntatori ai record abbiano lunghezza  $p = 6$ ; e che i nodi intermedi degli indici possano essere contenuti nei buffer.

In ciascuno dei seguenti casi:

- (a) indice primario (sparso) su  $A$  realizzato con B+-tree;
- (b) indice secondario su  $A$  realizzato con B+-tree;
- (c) struttura primaria hash su  $A$ .

indicare il costo prevedibile per le seguenti operazioni

1. `SELECT * FROM R WHERE A >= 1000 AND A <=3000`
2. `SELECT COUNT(*) FROM R WHERE A >= 1000 AND A <=3000`
3. `SELECT * FROM R WHERE A = 2000`

Riportare le risposte nella tabella sottostante, indicando formula e valore numerico (con brevissimo commento, se necessario)

	Risposte
(a) 1	
(b) 1	
(c) 1	
(a) 2	
(b) 2	
(c) 2	
(a) 3	
(b) 3	
(c) 3	

**Domanda 2** (20%) Specificare, per ciascuno dei seguenti schedule, la sua appartenenza alle varie classi (scrivere “sì” o “no” nelle caselle della tabella; assumere che gli identificatori delle transazioni corrispondano ai timestamp).

	Seriale	VSR	CSR	TS	2PL
$r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$					
$r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$					
$r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$					
$r_2(x)w_2(x)r_1(x)w_1(x)$					

**Domanda 3** (15%) Spiegare perché, scrivendo un programma che accede a due basi di dati diverse utilizzando JDBC, non è possibile garantire il commit a due fasi. Indicare quali funzionalità aggiuntive sono necessarie per i server locali che vogliono realizzare tale servizio e come si può procedere con un sistema che non disponga di tali funzionalità.

**Domanda 4** (15%) (Rispondere sul foglio protocollo) Rispondere a questa domanda **dopo** avere risposto alla domanda 1 relativa al secondo modulo.

Illustrare brevemente una alternativa di memorizzazione che favorisca interrogazioni volte a conoscere solo codici fiscali (cf) dei clienti e costi dei noleggi. Fornire una breve spiegazione, con riferimento alle problematiche di gestione dei dati in memoria secondaria.

**Domanda 5** (25%) (Rispondere sul foglio protocollo) Illustrare brevemente, ma con chiarezza, gli aspetti salienti delle trasformazioni utilizzate per alimentare uno degli schemi dimensionali utilizzati nel progetto a partire dalla base di dati sorgente fornita.

**Tecnologia delle basi di dati**  
**22 giugno 2010**

Cognome \_\_\_\_\_ Nome \_\_\_\_\_ Matricola \_\_\_\_\_

**Domanda 1** (25%) Sia data una relazione  $R(A, B, C)$  contenente circa  $L = 10.000.000$  ennuple di  $r = 20$  byte ciascuna, di cui  $a = 4$  per la chiave  $A$ , che contiene valori interi quasi consecutivi, da 1 a poco più di 10.000.000. Supporre che i blocchi abbiano dimensione  $B = 2\text{KB}$ , approssimabile come 2.000, che i puntatori ai record abbiano lunghezza  $p = 6$ ; e che i nodi intermedi degli indici possano essere contenuti nei buffer.

In ciascuno dei seguenti casi:

- (a) indice primario (sparso) su  $A$  realizzato con B+-tree;
- (b) indice secondario su  $A$  realizzato con B+-tree;
- (c) struttura primaria hash su  $A$ .

indicare il costo prevedibile per le seguenti operazioni

1. `SELECT * FROM R WHERE A >= 1000 AND A <=3000`
2. `SELECT COUNT(*) FROM R WHERE A >= 1000 AND A <=3000`
3. `SELECT * FROM R WHERE A = 2000`

Riportare le risposte nella tabella sottostante, indicando formula e valore numerico (con brevissimo commento, se necessario)

	Risposte
(a) 1	
(b) 1	
(c) 1	
(a) 2	
(b) 2	
(c) 2	
(a) 3	
(b) 3	
(c) 3	

**Domanda 2** (20%) Specificare, per ciascuno dei seguenti schedule, la sua appartenenza alle varie classi (scrivere “sì” o “no” nelle caselle della tabella; assumere che gli identificatori delle transazioni corrispondano ai timestamp).

	Seriale	VSR	CSR	TS	2PL
$r_1(y)r_2(z)r_2(y)w_2(y)w_2(z)r_1(z)$					
$r_1(x)w_1(x)r_2(x)w_2(x)r_0(y)w_1(y)$					
$r_1(x)r_2(z)w_2(z)w_1(x)r_2(x)w_2(x)$					
$r_2(x)w_2(x)r_1(x)w_1(x)$					

**Domanda 3** (15%) Spiegare perché, scrivendo un programma che accede a due basi di dati diverse utilizzando JDBC, non è possibile garantire il commit a due fasi. Indicare quali funzionalità aggiuntive sono necessarie per i server locali che vogliono realizzare tale servizio e come si può procedere con un sistema che non disponga di tali funzionalità.

**Domanda 4** (15%) Commentare brevemente la seguente affermazione: “le tecniche per il controllo di concorrenza basate su 2PL e su timestamp si basano su condizioni sufficienti ma non necessarie per la view-serializzabilità.” In particolare, chiarire se essa è vera o falsa, spiegare perché e motivare le ragioni pratiche per le quali si utilizzano tali tecniche anziché la view-serializzabilità stessa.

**Domanda 5** (25%) Considerare uno schema dimensionale relativo agli esami, che utilizzi, come tabella dei fatti e come una delle dimensioni, le relazioni come quelle qui schematizzate:

<u>KStudente</u>	<u>KCorso</u>	<u>KData</u>	Voto	...
301	201	405	25	...
301	202	406	28	...
302	201	407	30	...
302	203	407	22	...
...	...	...	...	...

<u>KCorso</u>	Titolo	Crediti	...
201	Fisica I	6	...
202	Chimica	9	...
203	Geometria	6	...
...	...	...	...

Supporre che si presentino le seguenti esigenze di modifica:

- i corsi cambiano nome nel tempo: per esempio, il corso nella prima enupla potrebbe ad un certo punto cambiare nome da “Fisica I” in “Meccanica”; interessano selezioni e aggregazioni relative agli esami tanto con riferimento al nome del corso (al momento dell’esame) quanto alla sua identità (un codice che viene introdotto allo scopo, ma non sempre viene utilizzato, perché alcuni analisti preferiscono fare riferimento al nome corrente del corso); le modifiche sono rare, ma è possibile che ci siano corsi con vari cambiamenti di nome;
- per ogni corso, interessa rappresentare anche il docente, per supportare analisi sugli esami svolti da ciascun docente; i docenti cambiano nel tempo e passano da un corso all’altro (e possono anche tenere più corsi nello stesso momento, ma ogni corso ha, in un certo giorno, un solo docente); è disponibile l’informazione relativa ai docenti dei corsi nel tempo (per tutto il periodo, anche passato, di interesse).

Mostrare nuove versioni delle due tabelle che permettano di soddisfare le esigenze sopra citate (mostrare anche i dati, con riferimento a quelli presenti negli esempi sopra, aggiungendo nuovi dati ragionevoli, che permettano di comprendere le modifiche).