

Basi di dati II, primo modulo — 7 febbraio 2012

Rispondere su questo fascicolo.

Tempo a disposizione: **un'ora e trenta minuti**.

Cognome _____ Nome _____ Matricola _____

Domanda 1 (15%)

Considerare un disco con una velocità di rotazione di $v = 10.000$ giri al minuto e un tempo medio di posizionamento della testina (tempo di seek) $t_S = 5$ msec. Ogni traccia contiene $N = 400$ blocchi e ogni blocco contiene $B = 4$ KB. Rispondere alle seguenti domande mostrando formula e valore numerico (N.B. non servono calcolatrici, i risultati sono semplici, approssimare $1 \text{ MB} = 1000 \text{ KB}$).

1. Qual è tempo medio di latenza (attesa dovuta alla rotazione) t_L ?

2. Qual è la massima velocità di trasferimento (in MB al secondo)?

3. Qual è il tempo minimo t_B necessario per leggere un blocco?

4. Qual è il tempo medio necessario per leggere un blocco?

5. Qual è il tempo medio necessario per leggere una traccia?

Domanda 2 (20%) Una catena di supermercati ha una base di dati dei propri clienti che dispongono di una “tessera fedeltà,” con varie informazioni su ciascun cliente, fra cui (a) il totale dei punti acquisiti attraverso l’uso della tessera e (b) il negozio della catena cui fa riferimento (ad esempio, quello presso cui ha inizialmente richiesto la tessera). Si vuole eseguire su di essa l’interrogazione che calcola, per ciascun negozio, il numero dei clienti, la somma dei punti fedeltà dei clienti e la relativa media per cliente. Indicare quale livello di isolamento (READ UNCOMMITTED, READ COMMITTED, REPEATABLE READ o SERIALIZABLE) si potrebbe scegliere in ciascuno dei seguenti casi (si supponga che, in generale, sia stato rilevato che, nel corso degli inserimenti e delle modifiche, vengono inseriti valori sbagliati anche di vari ordini di grandezza, che sono poi corretti prima del commit):

1. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di tutti i clienti (a seguito di una ridefinizione dei criteri di assegnazione dei punti stessi), con la finalità di acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
2. L’operazione è eseguita mentre vengono inseriti molti nuovi clienti, con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
3. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di alcuni clienti (in ciascun negozio una piccola percentuale), con la finalità di acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
4. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di alcuni clienti (in ciascun negozio una piccola percentuale), con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
5. L’operazione è eseguita in un momento in cui non ci sono aggiornamenti di alcun genere, con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.

Risposte				
1.	2.	3.	4.	5.

Domanda 3 (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 $B = 2KB$, 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.

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`

in ciascuno dei seguenti casi:

- (a) indice primario (sparso) su A realizzato con B+-tree;
- (b) indice secondario (denso) su A realizzato con B+-tree.

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

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

Domanda 4 (25%)

Considerare i due seguenti scenari in ciascuno dei quali due client diversi inviano richieste ad un gestore del controllo di concorrenza. Ciascun client può inviare una richiesta solo dopo che è stata eseguita o rifiutata la precedente (se invece una richiesta viene bloccata da un lock, allora il client rimane inattivo fino alla concessione o allo scadere del timeout). Si supponga che, in caso di stallo, abortisca la transazione che ha avanzato la richiesta per prima. In caso di abort, si ignorino le successive richieste della transazione che ha abortito (senza rilanciarla).

scenario 1		scenario 2	
client 1	client 2	client 1	client 2
read(x)		read(x)	
x = x + 10	read(x)		read(x)
write(x)			x = x + 20
	x = x + 20		write(x)
	write(x)		commit
commit		read(x)	
	commit	commit	

Considerare uno scheduler che utilizzi il controllo di concorrenza basato su multiversioni e livelli di isolamento SERIALIZABLE e READ COMMITTED. Assumiamo che (come avviene di solito) multiversioni preveda

- **SERIALIZABLE:** le letture fanno riferimento allo stato della base di dati all'inizio della transazione e le scritture di una transazione T sono soggette ad un lock a due fasi stretto (solo per le scritture) e sono ammesse solo se il dato non è stato modificato, dopo l'inizio di T, da altre transazioni.
- **READ COMMITTED:** le letture fanno riferimento allo stato della base di dati all'inizio della specifica lettura e le scritture sono soggette ad un lock a due fasi stretto (solo per le scritture).

Mostrare il comportamento dello scheduler nei due casi seguenti, supponendo che il valore iniziale dell'oggetto x sia 100. Indicare le operazioni che vengono eseguite nell'ordine con, per ciascuna, il valore che viene letto o scritto. In conclusione, per ciascun caso, dire se si verificano o meno anomalie.

Scenario 1 SERIALIZABLE	Scenario 2 READ COMMITTED

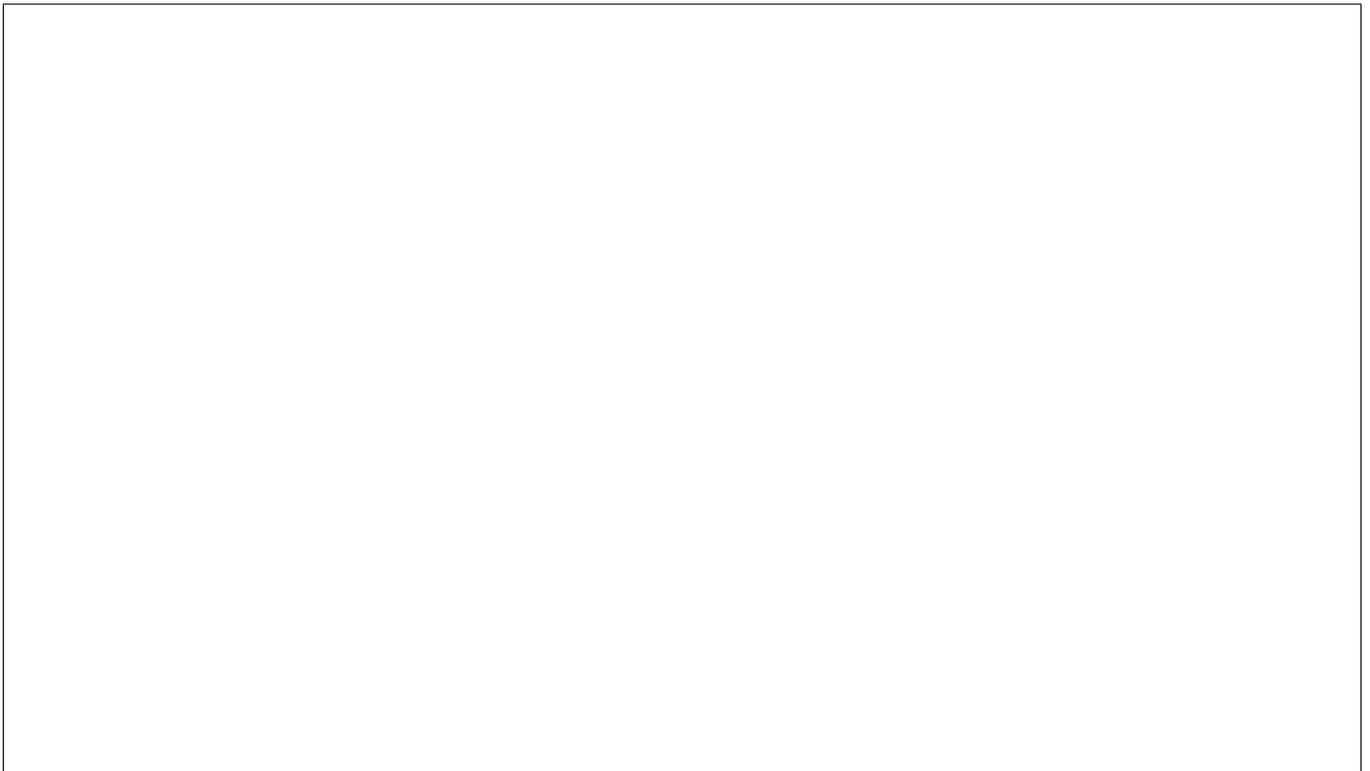
Domanda 5 (15%) Si consideri uno schema dimensionale utilizzato per analizzare le vendite in una catena di supermercati che, fra le dimensioni, ne preveda una sui negozi, come la seguente:

<u>K</u> Negozio	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).



Basi di dati II, primo modulo — 7 febbraio 2012
Cenni sulle soluzioni

Rispondere su questo fascicolo.

Tempo a disposizione: **un'ora e trenta minuti**.

Cognome _____ Nome _____ Matricola _____

Domanda 1 (15%)

Considerare un disco con una velocità di rotazione di $v = 10.000$ giri al minuto e un tempo medio di posizionamento della testina (tempo di seek) $t_S = 5$ msec. Ogni traccia contiene $N = 400$ blocchi e ogni blocco contiene $B = 4$ KB. Rispondere alle seguenti domande mostrando formula e valore numerico (N.B. non servono calcolatrici, i risultati sono semplici, approssimare $1 \text{ MB} = 1000 \text{ KB}$).

1. Qual è tempo medio di latenza (attesa dovuta alla rotazione) t_L ?

Il tempo necessario per mezzo giro:

$$t_L = \frac{1}{2} \times \frac{1}{v} = \frac{1}{2} \times \frac{1}{10.000} \text{ min} = \frac{1}{2} \times \frac{1}{10.000} \times 60 \text{ sec} = 0,003 \text{ sec} = 3 \text{ msec}$$

2. Qual è la massima velocità di trasferimento (in MB al secondo)?

Velocità di rotazione moltiplicato capacità di una traccia

$$v \times N \times B = 10.000 \times 400 \times 4 \text{ KB/min} = 10.000 \times 400 \times 4 \times \frac{1}{60} \text{ KB/sec} = 266.000 \text{ KB/sec} = 266 \text{ MB/sec}$$

3. Qual è il tempo minimo t_B necessario per leggere un blocco?

Il tempo necessario per fare un giro diviso il numero di blocchi di una traccia

$$t_B = \frac{1}{v} \times 60 \times \frac{1}{N} = \frac{1}{10.000} \times \frac{1}{400} \text{ min} = \frac{1}{10.000} \times 60 \times \frac{1}{400} \text{ sec} = 15 \mu\text{sec}$$

4. Qual è il tempo medio necessario per leggere un blocco?

La somma del tempo medio di seek, del tempo medio di latenza e del tempo minimo di lettura di un blocco

$$t_S + t_L + t_B = 5 + 3 + 0,015 \text{ msec} = \text{ca } 8 \text{ msec}$$

5. Qual è il tempo medio necessario per leggere una traccia?

Il tempo medio di seek più il tempo necessario per un giro

$$t_S + \frac{1}{v} = 5 \text{ msec} + \frac{1}{10.000} \text{ min} = 5 \text{ msec} + \frac{1}{10.000} \times 60 \text{ sec} = 11 \text{ msec}$$

Domanda 2 (20%) Una catena di supermercati ha una base di dati dei propri clienti che dispongono di una “tessera fedeltà,” con varie informazioni su ciascun cliente, fra cui (a) il totale dei punti acquisiti attraverso l’uso della tessera e (b) il negozio della catena cui fa riferimento (ad esempio, quello presso cui ha inizialmente richiesto la tessera). Si vuole eseguire su di essa l’interrogazione che calcola, per ciascun negozio, il numero dei clienti, la somma dei punti fedeltà dei clienti e la relativa media per cliente. Indicare quale livello di isolamento (READ UNCOMMITTED, READ COMMITTED, REPEATABLE READ o SERIALIZABLE) si potrebbe scegliere in ciascuno dei seguenti casi (si supponga che, in generale, sia stato rilevato che, nel corso degli inserimenti e delle modifiche, vengono inseriti valori sbagliati anche di vari ordini di grandezza, che sono poi corretti prima del commit):

1. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di tutti i clienti (a seguito di una ridefinizione dei criteri di assegnazione dei punti stessi), con la finalità di acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
2. L’operazione è eseguita mentre vengono inseriti molti nuovi clienti, con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
3. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di alcuni clienti (in ciascun negozio una piccola percentuale), con la finalità di acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
4. L’operazione è eseguita mentre vengono modificati i valori dei punti fedeltà di alcuni clienti (in ciascun negozio una piccola percentuale), con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
5. L’operazione è eseguita in un momento in cui non ci sono aggiornamenti di alcun genere, con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.

Risposte				
1.	2.	3.	4.	5.
RR	S	RC	RR	RU

Domanda 3 (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 $B = 2KB$, 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.

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`

in ciascuno dei seguenti casi:

- (a) indice primario (sparso) su A realizzato con B+-tree;
- (b) indice secondario (denso) su A realizzato con B+-tree.

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

Possibile soluzione

Indichiamo con

- $F = B/r = 100$ il fattore di blocco del file
- $F_I = B/(a+p) = 200$ il fattore di blocco massimo dell’indice; trattandosi di B+-tree, quello reale F'_I sarà minore (assumiamo del 30%) $F'_I = 150$ circa
- $N = 2000$ (o poco meno) il numero di record con valore di A compreso fra 1000 e 3000

- (a)1 : poiché il file è ordinato, gli N record da trovare si trovano in $N/F = 20$ blocchi, che sono quindi accessibili (visto che l’indice è sparso), attraverso $\lceil (N/F)/F'_I \rceil = 1$ blocchi dell’indice; il costo è pari all’accesso alla foglia dell’indice (i livelli più alti sono nel buffer, quindi non costano niente) più a quelli dei record: $\lceil (N/F)/F'_I \rceil + N/F =$ circa 20
- (b)1 : il file non è ordinato, i record sono sparpagliati, vanno acceduti a uno a uno e quindi servono N accessi più le foglie dell’indice (che è denso) $\lceil N/F'_I \rceil =$ circa 15 quindi trascurabile; costo totale circa $N = 2000$
- (a)2 : come a(1) (l’indice è sparso e quindi per contare è necessario accedere ai record)
- (b)2 : l’indice è denso, bastano le foglie dell’indice: $\lceil N/F'_I \rceil =$ circa 15
- (a)3 : accesso diretto: una foglia dell’indice più un blocco del file, totale 2
- (b)3 : come (a)3

Domanda 4 (25%)

Considerare i due seguenti scenari in ciascuno dei quali due client diversi inviano richieste ad un gestore del controllo di concorrenza. Ciascun client può inviare una richiesta solo dopo che è stata eseguita o rifiutata la precedente (se invece una richiesta viene bloccata da un lock, allora il client rimane inattivo fino alla concessione o allo scadere del timeout). Si supponga che, in caso di stallo, abortisca la transazione che ha avanzato la richiesta per prima. In caso di abort, si ignorino le successive richieste della transazione che ha abortito (senza rilanciarla).

scenario 1		scenario 2	
client 1	client 2	client 1	client 2
read(x)		read(x)	
x = x + 10	read(x)		read(x)
write(x)			x = x + 20
	x = x + 20		write(x)
	write(x)		commit
commit		read(x)	
	commit	commit	

Considerare uno scheduler che utilizzi il controllo di concorrenza basato su multiversioni e livelli di isolamento SERIALIZABLE e READ COMMITTED. Assumiamo che (come avviene di solito) multiversioni preveda

- **SERIALIZABLE**: le letture fanno riferimento allo stato della base di dati all’inizio della transazione e le scritture di una transazione T sono soggette ad un lock a due fasi stretto (solo per le scritture) e sono ammesse solo se il dato non è stato modificato, dopo l’inizio di T, da altre transazioni.
- **READ COMMITTED**: le letture fanno riferimento allo stato della base di dati all’inizio della specifica lettura e le scritture sono soggette ad un lock a due fasi stretto (solo per le scritture).

Mostrare il comportamento dello scheduler nei due casi seguenti, supponendo che il valore iniziale dell’oggetto x sia 100. Indicare le operazioni che vengono eseguite nell’ordine con, per ciascuna, il valore che viene letto o scritto. In conclusione, per ciascun caso, dire se si verificano o meno anomalie.

Scenario 1 SERIALIZABLE				Scenario 2 READ COMMITTED			
client 1		client 2		client 1		client 2	
read(x)	legge 100	read(x)	legge 100	read(x)	legge 100	read(x)	legge 100
x = x + 10	x vale 110					x = x + 20	x vale 120
write(x)	scrive 110	x = x + 20	x vale 120	write(x)	scrive 120	write(x)	scrive 120
		xlock(x)	bloccata	commit		commit	
commit		abort					
non c'è anomalia; l'abort è dovuto al fatto che x è stato modificato dopo l'avvio della transazione che sta scrivendo				anomalia: lettura inconsistente; la seconda lettura del client 1 legge il valore corrente di x			

Domanda 5 (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:

<u>KNegozio</u>	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 enupla 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

<u>KNeg</u>	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