Basi di dati II — Esame — 21 febbraio 2019

Tempo a disposizione: due ore.

Cognome _____ **Nome** _____ **Matricola** _____

Basi di dati II — 21 febbraio 2019

Domanda 1 (20%)
Considerare il seguente scenario in cui tre 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 supponga che il client rilanci la stessa transazione (subito dopo l'esecuzione delle altre azioni in attesa sullo stesso dato).

client 1	client 2	client 3
begin		
read(x)		
	begin	
	read(x)	
x = x + 10		
write(x)		
		begin
		read(x)
	x = x + 20	
	write(x)	
	commit	
commit		
		read(x)
		commit

Considerare uno scheduler con controllo di concorrenza basato su **Multversioni** (come in Postgres) e livello di isolamento **SERIALIZABLE** sulle prime due transazioni e **READ COMMITTED** sulla terza. Mostrare il comportamento dello scheduler, supponendo che il valore iniziale dell'oggetto x sia **400**. Indicare, nell'ordine, le operazioni che vengono eseguite da ciascun client, specificando, per ciascuna, il valore che viene letto o scritto. In conclusione, dire se si verificano o meno anomalie.

client 1	client 2	client 3

Si verificano anomalie?

Basi di dati II — 21 febbraio 2019

Considerare nuovamente lo scenario della pagina precedente, ripetuto qui sotto per comodità.

client 1	client 2	client 3
begin read(x)		
	begin read(x)	
x = x + 10 write(x)		
		begin read(x)
	x = x + 20 write(x) commit	
commit		
		read(x) commit

Considerare uno scheduler con controllo di concorrenza ancora basato su **Multversioni** (come in Postgres) ma con livello di isolamento **READ COMMITTED** sulle prime due transazioni e **SERIALIZABLE** sulla terza. Mostrare il comportamento dello scheduler, supponendo che il valore iniziale dell'oggetto x sia ancora **4**.

client 1	client 2	client 3

Si verificano anomalie?

Basi di dati II — 21 febbraio 2019

Domanda 2 (20%) Si consideri la seguente base di dati, relativa alle ricette acquisite da un insieme di farmacie:

- Ricette(Numero, CodFarmacia, CFPaziente, Data)
- Farmacie(CodFarmacia, Nome, CodIndirizzo)
- ElementiRicetta(NumeroRicetta, CodFarmaco, Quantità)
- Farmaci(Codice, Descrizione, CodMolecola, CodCasa, Prezzo, Fascia)
- Molecole(CodMolecola, Descrizione)
- Pazienti(CF, Cognome, Nome, DataNascita, CodIndirizzo)
- CaseFarmaceutiche(CodCasa, Nome)
- ASL(Codice, Nome)
- Territorio(CodIndirizzo,Via, NumeroCivico, Comune, ASL)

Ci sono dati che cambiano nel tempo fra cui prezzi e fasce ('A', 'B' o 'C') dei farmaci e indirizzi dei pazienti. Costruire, in tale contesto, uno schema a stella che permetta di analizzare le prescrizioni (quantità e prezzi complessivi) rispetto a

- data (dimensione standard i cui dettagli possono essere omessi);
- farmaci, con le loro proprietà (molecola e casa farmaceutica);
- prescrizione di farmaci nella stessa ricetta
- ASL di residenza e fascia d'età (ad esempio, 0-3,4-17, 18-30, ...; ma potrebbero variare) dei pazienti;
- ASL della farmacia

Supporre che, per ovvie ragioni di privacy, non possano essere riportati dati che permettano di risalire alle identità dei pazienti (CF, cognome, nome, data di nascita e indirizzo) **Indicare esplicitamente la grana dei fatti.**

Grana dei fatti:

Schema dimensionale:

Basi di dati II — 21 febbraio 2019

Domanda 3 (25%) Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco. Quindi R1 occupa $B_1 = 6$ blocchi e R2 ne occupa $B_2 = 8$.

Relazione R1

30	X01	AA	31	Y01	DA	32	Z03	AB	33	K03	AB	34	Z03	AB	35	Z03	AB
	Y42	CA		X42	CC		W05	EF		W07	EF		W08	EF		W09	EF
	W73	CC		W93	CB		X52	HA		X59	HA		X50	HA		X56	HA
	Z55	GC		W54	LB		Y55	EA		Y54	EA		Y51	EA		Y57	EA

Relazione R2

40	AA	3	41	BC	4	42	LB	7	43	AA	8	44	AC	3	45	EA	7	46	BA	5	47	EF	6
	DA	7		GB	7		HB	3		EC	2		CB	5		LB	8		BB	4		GA	8

Si supponga di disporre di un buffer di $p = 4$ pagine.
Considerare l'esecuzione del join di R1 ed R2, sulla base dei valori del secondo attributo di R1 e del primo di R2, con il metodo nested loop senza utilizzo di indici. Supporre che non serva memorizzare il risultato e che quindi esso possa essere prodotto una ennupla alla volta (approccio "pipelining")

Indicare, una notazione del tipo 'pin(30)' e 'unpin(30),' tutte le operazioni di pin (o fix) e unpin necessarie per eseguire l'intero join.

Indicare il numero complessivo di accessi a memoria secondaria necessari per eseguire il join (indicare formula e numero)

Indicare, nell'ordine, le prime quattro ennuple che vengono prodotte

Indicare gli indirizzi dei blocchi che si trovano nel buffer dopo che sono state prodotte le prime quattro ennuple.

Basi di dati II — 21 febbraio 2019

Domanda 4 (20%)
Si considerino due relazioni $R_1(\underline{A}, C)$, $R_2(\underline{A}, D, E, F)$, in cui gli attributi hanno tutti la stessa dimensione $a = 2$Byte, molto più piccola della dimensione del blocco pari a $B = 8000$ Byte. Si supponga che le relazioni abbiano entrambe $L = 4.000.000$ ennuple, con gli stessi valori su A, e che le operazioni più frequenti su di essa siano le seguenti:

- o_1 selezione di una ennupla del join di R_1 e R_2 (sulla base del valore di A), con frequenza $f_1 = 10.000$;
- o_2 scansione dell'intera relazione R_1, con frequenza $f_2 = 1$

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 separata delle due relazioni, entrambe ordinate su A e con indice primario su A con profondità $p = 4$, con 2 livelli mediamente disponibili nel buffer.

(ii) memorizzazione in un cluster delle due relazioni pure entrambe ordinate su A e con indice primario su A con profondità sempre $p = 4$, con 2 livelli mediamente disponibili nel buffer.

In conclusione, conviene quindi la memorizzazione in un cluster? (Sì o No)

Ripetere la valutazione effettuata alla pagina precedente nel caso in cui le due frequenze sono invece $f_1 = 1000$ e $f_2 = 10$

(i) memorizzazione separata delle due relazioni ...

(ii) memorizzazione in un cluster delle due relazioni ...

In conclusione, conviene quindi la memorizzazione in un cluster? (Sì o No)

Basi di dati II — 21 febbraio 2019

Domanda 5 (15%)
Considerare un sistema che utilizzi blocchi di lunghezza $D = 4$ KB (approssimabili a 4000 byte) e una tabella R con una struttura fisica heap con record a lunghezza fissa che occupano $L = 20$ byte ciascuno, in cui vengono inserite $N = 25.000$ ennuple, con valori della chiave tutti diversi fra loro e da quelli già nella relazione (quindi il sistema verifica il soddisfacimento del vincolo di chiave e ammette tutte le operazioni).
Rispondere alle domande seguenti, indicando formule e valori numerici:

Indicare il numero di scritture in memoria secondaria necessarie per realizzare i 25.000 inserimenti, supponendo che i record di log abbiano una lunghezza pari a circa il triplo di quella dei record del file, con riferimento ad un programma che utilizzi una transazione separata per ciascun inserimento

- numero di scritture di pagine di log:

- numero di scritture di pagine della relazione, nei tre casi seguenti:
  - strategia undo-redo senza vincoli particolari
  - strategia redo-only (no-undo)
  - strategia undo-only (no-redo)

Come nel caso precedente, ma con riferimento ad un programma che, per realizzare i 25.000 inserimenti, utilizzi complessivamente $k = 5$ transazioni, ognuna con 5000 inserimenti (e supponendo che non vi siano altre transazioni attive)

- numero di scritture di pagine di log:

- numero di scritture di pagine della relazione, nei tre casi seguenti:
  - strategia undo-redo senza vincoli particolari
  - strategia redo-only (no-undo)
  - strategia undo-only (no-redo)