Basi di dati II

Esame — 17 settembre 2021

Tempo a disposizione: due ore.

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

Basi di dati II — 17 settembre 2021

Domanda 1 (25%)
Considerare la relazione sotto schematizzata, definita su vari attributi, uno dei quali è la chiave, i cui valori sono mostrati. Supporre che la relazione abbia un fattore di blocco pari a 2 (e quindi occupi 15 blocchi) e che siano disponibili 9 buffer. Considerare l'esecuzione di un mergesort a più vie (e due passate) sulla relazione e mostrare lo stato delle strutture in memoria centrale e secondaria dopo l'esecuzione di sei chiamate al metodo `next()` sullo scan che implementa il mergesort. In particolare, mostrare i "run" (cioè le porzioni di file ordinate durante prima passata) memorizzati su disco e i buffer in memoria centrale, evidenziando per ciascun buffer il record corrente. Mostrare anche i record prodotti dalle prime sei chiamate di `next()`.

Run su disco

Buffer

Record prodotti dalle prime 5 `next()`

622	...
211	...
521	...
322	...
111	...
421	...
742	...
871	...
783	...
144	...
256	...
585	...
325	...
435	...
686	...
885	...
735	...
386	...
539	...
178	...
487	...
839	...
267	...
647	...
535	...
171	...
484	...
838	...
262	...
646	...

Considerare ancora quanto illustrato nella domanda precedente, con l'unica differenza nel numero di buffer disponibili: supporre che siano 3. Illustrare schematicamente i vari passi e i risultati intermedi (mostrando una quantità di dati che permetta di comprendere l'esecuzione dell'algoritmo).

622	...
211	...
521	...
322	...
111	...
421	...
742	...
871	...
783	...
144	...
256	...
585	...
325	...
435	...
686	...
885	...
735	...
386	...
539	...
178	...
487	...
839	...
267	...
647	...
535	...
171	...
484	...
838	...
262	...
646	...

Basi di dati II — 17 settembre 2021

Domanda 2 (25%)
Considerare uno schema dimensionale relativo ai risultati delle elezioni politiche nelle varie regioni d'Italia che utilizzi, come tabella dei fatti e come una delle dimensioni, relazioni come le seguenti (KE denota la chiave della dimensione relativa alle elezioni, KP di quella relativa al partito e KR di quella relativa alla regione; i voti sono in migliaia):

KE	KR	KP	Voti	Percent	...
501	101	1001	805	28,08	...
501	101	1002	106	3,51	...
501	101	1003	708	23,33	...
501	102	1001	1454	27,31	...
502	101	1003	75	7,8	...
...	...	...	...	...	...

KP	Sigla	Nome	...
1001	PB	Partito XXX	...
1003	PF	Partito AAA	...
...	...	...	...

Con riferimento a questo contesto, considerare le esigenze di modifica seguenti. Per ciascuna, proporre una modifica allo schema e rispondere alle eventuali domande.

- i partiti cambiano nome nel tempo: per esempio, il "Partito XXX" potrebbe ad un certo punto diventare il "Partito XYZ"; interessano selezioni e aggregazioni relative ai voti tanto con riferimento al nome del partito (al momento delle elezioni) 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 partito); le modifiche sono rare, ma è possibile che ci siano partiti con vari cambiamenti di nome; mostrare modifiche alle relazioni (una o entrambe) 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).

- per ogni partito, interessa rappresentare anche il leader al momento dell'elezione, per supportare analisi sui risultati di ciascun leader; i leader cambiano nel tempo (e possono passare da un partito all'altro, sia pure raramente ...); è disponibile l'informazione relativa ai leader dei partiti nel tempo (per tutto il periodo, anche passato, di interesse).

Basi di dati II — 17 settembre 2021

- Nelle ultime tornate elettorali, è diventato rilevante il concetto di "coalizione": ogni partito partecipa ad una coalizione (supponiamo per semplicità che un partito non coalizzato costituisca una coalizione da solo). Le coalizioni possono variare nel tempo. Proporre una ristrutturazione dello schema dimensionale e indicare se e come è possibile realizzare la modifica proposta anche con riferimento ad elezioni precedenti (indicare cioè quali dati debbono essere disponibili allo scopo nella staging area).

Basi di dati II — 17 settembre 2021

Domanda 3 (25%)
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)		
	begin read(x) x = x + 200 write(x)	
x = x + 100 write(x)		
	commit	
commit		
		(dopo molto tempo) 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 **10.000**. 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 — 17 settembre 2021

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

client 1	client 2	client 3
		begin
		read(x)
begin		
read(x)		
	begin	
	read(x)	
	x = x + 200	
	write(x)	
x = x + 100		
write(x)		
	commit	
commit		
		(dopo molto tempo)
		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 **10.000**.

client 1	client 2	client 3

Si verificano anomalie?

Basi di dati II — 17 settembre 2021

Domanda 4 (25%) Si considerino un sistema con blocchi di dimensione $B = 4000$ byte e una relazione *R(ID, CodiceFiscale, Cognome, ...)* di cardinalità pari circa a $L = 400.000$, con ennuple di $e = 80$ byte, con due chiavi, *ID* e *CodiceFiscale* (cioè il valore di ciascuna di esse, da solo, identifica univocamente una ennupla). Supporre che il sistema offra

- strutture primarie disordinate, ordinate o hash
- indici di tipo B-tree secondari o anche primari

Considerare un carico applicativo che preveda le seguenti operazioni **tutte di lettura**

1. ricerca di una ennupla sulla base del valore completo di *ID*, frequenza oraria $f_1 = 10.000$
2. ricerca di ennuple sulla base del *CodiceFiscale*, eventualmente parziale, con frequenza oraria $f_2 = 10$; supporre che il valore parziale sia molto selettivo e porti alla identificazione, in media, di $s = 2$ ennuple;
3. ricerca di una ennupla sulla base del valore parziale (una sottostringa iniziale) dell'attributo *Cognome*, con frequenza oraria $f_3 = 100$; supporre che il valore parziale porti alla identificazione, in media, di $s = 10$ ennuple.

Progettare l'organizzazione fisica della relazione, individuando la struttura primaria (hash, disordinata o ordinata) e gli eventuali indici (da nessuno a tre). Ragionare in termini di numero di accessi a memoria secondaria, assumendo che: (i) gli indici abbiano profondità $p = 4$, (ii) il buffer disponibile permetta di mantenere stabilmente in memoria due livelli di indice. Proporre almeno due alternative (quelle che intuitivamente si ritengono migliori) e valutarne il costo. Rispondere negli spazi sottostanti, in forma sia simbolica sia numerica.

	Alternativa 1	Alternativa 2	Alternativa 3 (eventuale)
struttura primaria indici utilizzati			
Costo Op. 1			
Costo Op. 2			
Costo Op. 3			
Costo tot			