Basi di dati II

Esame — 9 febbraio 2022

Tempo a disposizione: due ore.

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

Domanda 1 (20%)
Considerare un sistema distribuito con tre nodi X, Y e Z, che eseguono due transazioni T_1 e T_2 che coinvolgono i tre nodi, in modo diverso. Per la prima transazione X è il coordinatore, mentre per la seconda il coordinatore è Y. I due coordinatori inviano, come riportato nello schema sottostante, le richieste di `prepare`. Il nodo Z va in crash subito dopo aver risposto alla prima richiesta (senza avere il tempo di ricevere il messaggio di `commit`) e prima di ricevere la seconda. Poi va in crash anche il nodo Y. Indicare, nello schema sottostante, una possibile sequenza di scritture sui log e invio di messaggi (che includa anche i passi sopra illustrati), supponendo che entrambi i nodi siano ripristinati abbastanza presto (ma che vengano persi alcuni messaggi di risposta, ad esempio inviati a seguito di una decisione). Per i messaggi si usi la notazione *tipo(transaz)→destinatari* (come nell'esempio: `prepare(`T_1`)→Y,Z`). Supporre che nel log del coordinatore si scrivano solo i record di `prepare`, `commit` e `complete`, con i messaggi gestiti invece in memoria. Indicare ragionevoli istanti per i timeout, che permettano di concludere il protocollo per entrambe le transazioni.

Nodo X		Nodo Y		Nodo Z	
Log	Messaggi	Log	Messaggi	Log	Messaggi
`prep(`T_1`,Y,Z)`	`prep(`T_1`)→Y,Z`				
				crash	
		`prep(`T_2`,X,Z)`			
			`prep(`T_2`)→X,Z`		
		crash			
		restart			
				restart	

Basi di dati II — 9 febbraio 2022

Domanda 2 (15%) 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
- indici di tipo B-tree

Considerare un carico applicativo che preveda le seguenti operazioni

1. inserimento di una ennupla, con verifica dei due vincoli di chiave (su *CodiceFiscale* e su *ID*) con frequenza oraria $f_1 = 10.000$;
2. ricerca di una ennupla sulla base del valore completo di *ID*, frequenza oraria $f_2 = 10.000$
3. ricerca di ennuple sulla base del *CodiceFiscale*, eventualmente parziale, con frequenza oraria $f_3 = 10$; supporre che il valore parziale sia molto selettivo e porti alla identificazione, in media, di $s = 2$ ennuple;
4. ricerca di una ennupla sulla base del valore parziale (una sottostringa iniziale) dell'attributo *Cognome*, con frequenza oraria $f_4 = 1$; supporre che il valore parziale sia poco selettivo e porti alla identificazione, in media, di $s = 40$ ennuple.

Progettare l'organizzazione fisica della relazione, individuando 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, (iii) lettura e scrittura abbiano lo stesso costo. 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)
Indici utilizzati			
Costo Op. 1			
Costo Op. 2			
Costo Op. 3			
Costo Op. 4			
Costo tot			

Basi di dati II — 9 febbraio 2022

Domanda 3 (15%)
Considerare ancora il caso illustrato nella domanda precedente, ma con riferimento ad una fase in cui le frequenze siano completamente diverse:

1. $f_1 = 1$
2. $f_2 = 1000$
3. $f_3 = 1$
4. $f_4 = 1000$

Indicare quale soluzione si sceglierebbe in questo caso

	Alternativa 1	Alternativa 2	Alternativa 3 (eventuale)
Indici utilizzati			
Costo Op. 1			
Costo Op. 2			
Costo Op. 3			
Costo Op. 4			
Costo tot			

Basi di dati II — 9 febbraio 2022

Domanda 4 (30%) Si consideri la seguente porzione dello schema dell'archivio delle carriere degli studenti di una anagrafe ministeriale:

- STUDENTI (CodiceFiscale, Cognome, Nome, DataNascita, TipoMaturità, VotoMaturità)
- ISCRIZIONI (CodiceFiscale, AnnoAccademico, CodiceCdS), (si noti che la chiave include anche AnnoAccademico perché, in anni diversi, lo studente potrebbe essere iscritto a corsi di studio diversi)
- CORSIDISTUDIO (CodiceCdS, Titolo, Livello, Classe, CodiceUniv), che contiene informazioni su tutti i corsi di laurea, triennali e magistrali
- LAUREE CodiceFiscale, CodiceCdS, Data, Voto)
- UNIVERSITÀ (CodiceUniv, NomeUniversità), che contiene informazioni su tutte le università

Supporre (i) che l'anno accademico sia rappresentato in modo semplice e sempre nello stesso modo; (ii) che il titolo di un corso di studio possa cambiare da un anno all'altro.
Progettare uno schema dimensionale che permetta di rispondere, fra le altre, alle seguenti interrogazioni:

- calcolare il numero di studenti (con la relativa media dei voti e l'età media) che si sono laureati in un certo corso di studio (inteso come corso di studio presso una università) in un certo anno accademico (si supponga che, per la data di laurea, l'unico dettaglio rilevante sia l'anno accademico e che esista un modo univoco per associare un anno accademico ad una data di laurea)
- calcolare il numero di laureati per una classe di corsi studio, distinto per tipo di maturità e per numero di anni impiegati per conseguire il titolo (ad esempio, 2, 3, 4, 5, 6, più di 6)
- calcolare il numero di laureati per una classe di corsi studio, distinto per "età alla laurea" (ad esempio, 21, 22, 23, ...26, più di 26)

Assumere che, per ragioni di privatezza e di compattezza, sia opportuno limitare la cardinalità della tabella dei fatti, a patto di permettere la risposta alle precedenti interrogazioni. Per le fasce di voto, supporre che interessino fasce di 5 in 5 oppure di 10 in 10.
Indicare esplicitamente la grana dei fatti.

Grana dei fatti:

Schema dimensionale:

Basi di dati II — 9 febbraio 2022

Descrivere, informalmente, ma in modo strutturato e comprensibile, il processo di ETL che porta alla tabella dei fatti mostrata in risposta alla domanda precedente

Basi di dati II — 9 febbraio 2022

Domanda 5 (15%)
Per ciascuno degli schedule sotto riportati, indicare, scrivendo **sì** o **no** nelle varie caselle, a quali classi appartiene: S (seriale, rispetto a letture e scritture, ignorare commit e abort), CSR (conflict-serializzabile), S2PL (generabile da uno scheduler basato su 2PL stretto), MV (generabile da uno scheduler multiversion con controllo di serializzabilità: "a serializable transaction cannot modify or lock rows changed by other transactions after the serializable transaction began"). Negli schedule, s_i indica l'inizio della transazione i e c_i il suo commit.

	S	CSR	S2PL	MV
$s_1, s_2, r_1(x), r_2(x), w_1(x), c_1, w_2(x), c_2$				
$s_2, r_2(x), s_1, w_2(x), r_1(x), c_2, w_1(x), c_1$				
$s_2, r_2(x), w_2(x), s_1, c_2, r_1(x), w_1(x), c_1$				
$s_1, s_2, r_1(x), r_2(x), w_2(x), r_2(y), w_2(y), c_2, r_1(y), c_1$				