Basi di dati II

Esame — 20 febbraio 2020

Tempo a disposizione: due ore.

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

Basi di dati II — 20 febbraio 2020

Domanda 1 (20%) Si considerino un sistema con blocchi di dimensione $B = 2000$ byte e una relazione *R(ID, CodiceFiscale, Cognome, ...)* di cardinalità pari circa a $N = 200.000$, 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 $f_1 = 10$;
2. ricerca di una ennupla sulla base del valore completo di *ID*, frequenza oraria $f_2 = 1000$
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 $n = 2$ ennuple;
4. ricerca di una ennupla sulla base del valore parziale (una sottostringa iniziale) dell'attributo *Cognome*, con frequenza oraria $f_4 = 10.000$; supporre che il valore parziale sia poco selettivo e porti alla identificazione, in media, di $n = 40$ 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 — 20 febbraio 2020

Domanda 2 (10%)
Considerare ancora il caso illustrato nella domanda precedente, ma con riferimento ad una fase in cui le frequenze siano completamente diverse:

1. $f_1 = 10.000$
2. $f_2 = 10$
3. $f_3 = 10$
4. $f_4 = 1$

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 — 20 febbraio 2020

Domanda 3 (30%) 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, CodCasa, Prezzo, Fascia)
- Pazienti(CF, Cognome, Nome, DataNascita, CodIndirizzo)
- CaseFarmaceutiche(CodCasa, Nome, Nazione)
- Nazione(Codice, 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 (casa farmaceutica e nazione);
- 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 — 20 febbraio 2020

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 — 20 febbraio 2020

Domanda 4 (20%)
Si consideri uno schema dimensionale con la seguente tabella dei fatti relativa alle carriere degli studenti presso una università:

- FattiEsami(KMateria, KSessione, KCorsoDiStudio, NumeroEsami, Media)

Con riferimento a tale schema, si supponga che

- complessivamente la tabella contenga 100.000 ennuple
- vi siano 1000 materie e 10 sessioni e per ogni sessione e per ogni materia ci sia almeno un esame verbalizzato
- vi siano 100 corsi di studio
- ogni materia abbia studenti di mediamente 10 corsi di studio (e che in ogni sessione e per ogni materia ci sia almeno uno studente per ciascuno di tali corsi di studio)
- le operazioni più frequenti siano quelle che producono:
  1. numero esami per una materia e ciascuna sessione (ad esempio, Basi di dati II, in ciascuna delle 10 sessioni; quindi il risultato contiene 10 ennuple) con frequenza $f_1 = 100$
  2. numero esami complessivo per ciascuna materia (quindi il risultato contiene 1000 ennuple) con frequenza $f_2 = 1$
  3. numero esami in una sessione (ad esempio, la sessione estiva del 2017) per ciascun corso di studio (quindi vanno prodotte 100 ennuple, una per corso di studio) con frequenza $f_3 = 100$
- sia possibile realizzare una sola vista materializzata
- nelle selezioni, il costo sia pari al numero di ennuple estratte dalla relazione (prima dell'eventuale aggregazione)

Scegliere, fra le tre viste sostanzialmente corrispondenti alle tre interrogazioni, quella che si ritiene supporti meglio il carico applicativo. Indicare lo schema di ciascuna vista.

Basi di dati II — 20 febbraio 2020

Domanda 5 (10%)
Considerare una relazione definita con il seguente comando:

```
create table conti (numero integer primary key, saldo integer not null);
```

Considerare il seguente scenario in cui due client 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. Supporre che, in caso di abort, le transazioni non vengano rilanciate

`start transaction`
`  isolation level repeatable read;`
`insert into conti values (8, 1000) ;`
 	
`commit;`
	
	`start transaction`
	`  isolation level repeatable read;`
	`insert into conti values (8, 3000) ;`
	`commit`

Considerare uno scheduler con controllo di concorrenza basato su **Multversioni** (come in Postgres). Mostrare il comportamento dello scheduler,

Basi di dati II — 20 febbraio 2020

Domanda 6 (10%)
Considerare ancora la relazione considerata nella domanda precedetnte:

```
create table conti (numero integer primary key, saldo integer not null);
```

e il seguente scenario

`start transaction`
`  isolation level serializable;`
`select * from conti where numero = 3;`
 	
`update conti`
`  set saldo = 20 where numero = 3;`
 	
`commit;`
	
	`start transaction`
	`  isolation level serializable;`
	`select * from conti where numero = 3;`
	
	`update conti`
	`  set saldo = 10 where numero = 3;`
	
	`commit;`

Considerare ancora uno scheduler con controllo di concorrenza basato su **Multversioni** (come in Postgres). Mostrare il comportamento dello scheduler,
