Basi di dati II

Prova parziale — 9 maggio 2012 — Compito A

Cenni sulle soluzioni

Rispondere su questo fascicolo.
Tempo a disposizione: un'ora e quindici minuti.

Cognome _____ **Nome** _____ **Matricola** _____ **Ordin.** _____

Domanda 1 (25%)
Considerare un sistema distribuito su cui viene eseguita una transazione che coinvolge tre nodi, un coordinatore A e due partecipanti B e C. Dopo la richiesta di `prepare` da parte del coordinatore, il partecipante C va in crash senza ricevere il messaggio, mentre il partecipante B fa in tempo a ricevere il messaggio e a rispondere positivamente ma va in crash poco dopo, non ricevendo la prima notifica della decisione. Indicare, nello schema sottostante, una possibile sequenza di scritture sui log e invio di messaggi, supponendo che entrambi i nodi siano ripristinati abbastanza presto. Per semplicità, si fa riferimento ad una sola transazione e quindi non c'è bisogno di indicarla. Per i messaggi si usi la notazione *tipo*→*destinatari* (come nell'esempio: `prepare`→`B,C`). Supporre che timeout per le varie fasi scattino all'incirca negli istanti indicati a sinistra della tabella.

	Nodo A		Nodo B		Nodo C	
	Log	Messaggi	Log	Messaggi	Log	Messaggi
	`prepare(B,C)`	`prepare`→`B,C`				
			`ready`	`ready`→`A`		
			crash			
t_1	`abort`	`abort`→`B,C`				
t_2		`abort`→`B,C`				
			restart			
			`abort`	`ack`→`A`		
t_3		`abort`→`C`				
					crash	
					restart	
					`abort`	
						`ack`→`A`
	`complete`					

Eventuali commenti:
A deve ricevere un `ack` anche da C, prima di concludere con `complete`, perché C potrebbe essere in stato di `ready`

Domanda 2 (25%)
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 inseriti alcuni nuovi clienti (per ciascun negozio pochi rispetto a quelli già presenti), 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 acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
3. L'operazione è eseguita mentre vengono inseriti alcuni nuovi clienti (per ciascun negozio pochi rispetto a quelli già presenti), con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
4. 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 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.
RC	S	S	RR	RU

Domanda 3 (25%)
Considerare uno schema dimensionale relativo alla valutazione della didattica (realizzata attraverso i questionari distribuiti agli studenti, come nella nostra università) che utilizzi, come tabella dei fatti e come una delle dimensioni, relazioni come le seguenti (KAA denota la chiave della dimensione anno accademico e KCDS di quella relativa al corso di studio):

KAA	KCorso	KCDS	NumStudenti	Chiarezza	...
301	201	405	25	7,5	...
301	202	406	20	6,2	...
302	201	405	100	9,1	...
302	203	407	75	7,8	...
...	...	...	...	...	...

KCorso	Titolo	Area	...
201	Calcolo	Matematica	...
202	Fotonica	Elettronica	...
203	Geometria	Matematica	...
...	...	...	...

Con riferimento a questo contesto, considerare le esigenze di modifica mostrate nella pagina seguente e, per ciascuna, proporre una modifica allo scehma e rispondere alle eventuali altre domande.

- i corsi cambiano nome nel tempo: per esempio, il corso nella prima ennupla potrebbe ad un certo punto cambiare nome da “Calcolo” in “Analisi Matematica”; interessano selezioni e aggregazioni relative agli esami tanto con riferimento al nome del corso (al momento della valutazione) 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 corso); le modifiche sono rare, ma è possibile che ci siano corsi 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 la dimensione può essere utile la tecnica della“slowly changing dimension” di tipo 2 (e contemporaneamente 3), con un nuovo elemento (quindi una ennupla nella relazione) per ogni modifica e, viste le specifiche, un codice aggiuntivo, che non cambia nel tempo, e due attributi per il nome, con quello attuale e quello “dell’epoca,” rispettivamente.

KCorso	Codice	Titolo	TitoloAttuale	Crediti	...
201	FIS01	Fisica I	Meccanica	6	...
202	CHIM01	Chimica	Chimica	9	...
203	GEOM01	Geometria	Geometria	6	...
...	...	...	...	...	...
209	FIS01	Meccanica	Meccanica	6	...

- per ogni corso, interessa rappresentare anche il docente, per supportare analisi sulla valutazione complessiva di ciascun docente; i docenti cambiano nel tempo e passano da un corso all’altro (e possono anche tenere più corsi nello stesso momento, ma ogni corso ha, in ciascun anno accademico, un solo docente); è disponibile l’informazione relativa ai docenti dei corsi nel tempo (per tutto il periodo, anche passato, di interesse).

La tabella dei fatti può essere estesa aggiungendo una dimensione supplementare Docente, che dipende da Corso e Anno Accademico e può quindi essere facilmente aggiunta anche ad una tabella dei fatti esistente (i dati, come detto, sono disponibili e le ennuple rimangono le stesse, con un attributo in più; si noti che Kdocente dipende funzionalmente da KCorso e KAA):

KAA	KCorso	KCDS	KDocente	NumStudenti	Chiarezza	...
301	201	405	...	25	7,5	...
301	202	406	...	20	6,2	...
302	201	405	...	100	9,1	...
302	203	407	...	75	7,8	...
...	...	...	...	...	...	...

- interessano analisi più raffinate, relative all’anno di corso cui lo studente è iscritto al momento della valutazione; si supponga che nei questionari sia riportato l’anno di corso; indicare che cosa sarebbe necessario modificare nello schema dimensionale e quali dati debbono essere disponibili nella staging area per poter riorganizzare il data mart.

Va cambiata la grana dei fatti, aggiungendo una dimensione anno di corso. La riorganizzazione è possibile solo se nella staging area sono disponibili dati con maggiore dettaglio rispetto a quelli utilizzati nello schema dimensionale preesistente

Domanda 4 (25%)
Considerare la seguente interrogazione in SQL:

```
SELECT A, D, H
FROM R, S, T
WHERE E = B AND C = G AND H < 40
```

definita con riferimento a tre relazioni, definite e frammentate come segue (per essere poi distribuite):

- $R(\underline{A}, C, E)$ frammentata orizzontalmente:
  - $R_1 = \sigma_{C>100}(R)$;
  - $R_2 = \sigma_{C\leq 100}(R)$
- $S(\underline{B}, D, F)$ frammentata verticalmente:
  - $S_1 = \pi_{B,D}(S)$;
  - $S_2 = \pi_{B,F}(S)$
- $T(\underline{G}, H)$ frammentata orizzontalmente:
  - $T_1 = \sigma_{H>200}(T)$;
  - $T_2 = \sigma_{H\leq 200}(T)$

Mostrare (ad esempio sotto forma di albero) l'espressione dell'algebra relazionale definita sui frammenti (cioè su R_1, R_2, S_1, ...) che realizza in tale interrogazione tralasciando i frammenti che non contribuiscono.

Join dell'unione di R_1 e R_2 con S_1 e T_2 oppure unione di due join ciascuno di tre relazioni

- R_1 con S_1 e con T_2
- R_2 con S_1 e con T_2

S_2 e T_1 non contribuiscono in alcun modo

Basi di dati II

Prova parziale — 9 maggio 2012 — Compito B

Cenni sulle soluzioni

Rispondere su questo fascicolo.
Tempo a disposizione: un'ora e quindici minuti.

Cognome _____ **Nome** _____ **Matricola** _____ **Ordin.** _____

Domanda 1 (25%)
Considerare un sistema distribuito su cui viene eseguita una transazione che coinvolge tre nodi, un coordinatore X e due partecipanti Y e Z. Dopo la richiesta di `prepare` da parte del coordinatore, il partecipante Z va in crash senza ricevere il messaggio, mentre il partecipante Y fa in tempo a ricevere il messaggio e a rispondere positivamente ma va in crash poco dopo, non ricevendo la prima notifica della decisione. Indicare, nello schema sottostante, una possibile sequenza di scritture sui log e invio di messaggi, supponendo che entrambi i nodi siano ripristinati abbastanza presto. Per semplicità, si fa riferimento ad una sola transazione e quindi non c'è bisogno di indicarla. Per i messaggi si usi la notazione *tipo*→*destinatari* (come nell'esempio: `prepare`→`Y,Z`). Supporre che timeout per le varie fasi scattino all'incirca negli istanti indicati a sinistra della tabella.

	Nodo X		Nodo Y		Nodo Z	
	Log	Messaggi	Log	Messaggi	Log	Messaggi
	`prepare(Y,Z)`	`prepare`→`Y,Z`				
			`ready`			
				`ready`→`X`		
					crash	
t_1	`abort`					
		`abort`→`Y,Z`				
t_2		`abort`→`Y,Z`				
			restart			
			`abort`			
				`ack`→`X`		
t_3		`abort`→`Z`			*restart*	
					`abort`	
						`ack`→`X`
	`complete`					

Eventuali commenti:
X deve ricevere un `ack` anche da Z, prima di concludere con `complete`, perché Z potrebbe essere in stato di `ready`

Domanda 2 (25%)
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 inseriti alcuni nuovi clienti (per ciascun negozio pochi rispetto a quelli già presenti), 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 acquisire informazioni approssimate ma ragionevolmente indicative sugli andamenti complessivi.
3. L'operazione è eseguita mentre vengono inseriti alcuni nuovi clienti (per ciascun negozio pochi rispetto a quelli già presenti), con la finalità di individuare i primi tre negozi da premiare in una campagna promozionale sulla base dei punti acquisiti dai rispettivi clienti.
4. 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 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.
RC	S	S	RR	RU

Domanda 3 (25%)
Considerare uno schema dimensionale relativo alla valutazione della didattica (realizzata attraverso i questionari distribuiti agli studenti, come nella nostra università) che utilizzi, come tabella dei fatti e come una delle dimensioni, relazioni come le seguenti (KAA denota la chiave della dimensione anno accademico e KCDS di quella relativa al corso di studio):

KAA	KCorso	KCDS	NumStudenti	Chiarezza	...
301	201	405	25	7,5	...
301	202	406	20	6,2	...
302	201	405	100	9,1	...
302	203	407	75	7,8	...
...	...	...	...	...	...

KCorso	Titolo	Area	...
201	Calcolo	Matematica	...
202	Fotonica	Elettronica	...
203	Geometria	Matematica	...
...	...	...	...

Con riferimento a questo contesto, considerare le esigenze di modifica mostrate nella pagina seguente e, per ciascuna, proporre una modifica allo scehma e rispondere alle eventuali altre domande.

- i corsi cambiano nome nel tempo: per esempio, il corso nella prima ennupla potrebbe ad un certo punto cambiare nome da “Calcolo” in “Analisi Matematica”; interessano selezioni e aggregazioni relative agli esami tanto con riferimento al nome del corso (al momento della valutazione) 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 corso); le modifiche sono rare, ma è possibile che ci siano corsi 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 la dimensione può essere utile la tecnica della“slowly changing dimension” di tipo 2 (e contemporaneamente 3), con un nuovo elemento (quindi una ennupla nella relazione) per ogni modifica e, viste le specifiche, un codice aggiuntivo, che non cambia nel tempo, e due attributi per il nome, con quello attuale e quello “dell’epoca,” rispettivamente.

KCorso	Codice	Titolo	TitoloAttuale	Crediti	...
201	FIS01	Fisica I	Meccanica	6	...
202	CHIM01	Chimica	Chimica	9	...
203	GEOM01	Geometria	Geometria	6	...
...	...	...	...	...	...
209	FIS01	Meccanica	Meccanica	6	...

- per ogni corso, interessa rappresentare anche il docente, per supportare analisi sulla valutazione complessiva di ciascun docente; i docenti cambiano nel tempo e passano da un corso all’altro (e possono anche tenere più corsi nello stesso momento, ma ogni corso ha, in ciascun anno accademico, un solo docente); è disponibile l’informazione relativa ai docenti dei corsi nel tempo (per tutto il periodo, anche passato, di interesse).

La tabella dei fatti può essere estesa aggiungendo una dimensione supplementare Docente, che dipende da Corso e Anno Accademico e può quindi essere facilmente aggiunta anche ad una tabella dei fatti esistente (i dati, come detto, sono disponibili e le ennuple rimangono le stesse, con un attributo in più; si noti che Kdocente dipende funzionalmente da KCorso e KAA):

KAA	KCorso	KCDS	KDocente	NumStudenti	Chiarezza	...
301	201	405	...	25	7,5	...
301	202	406	...	20	6,2	...
302	201	405	...	100	9,1	...
302	203	407	...	75	7,8	...
...	...	...	...	...	...	...

- interessano analisi più raffinate, relative all’anno di corso cui lo studente è iscritto al momento della valutazione; si supponga che nei questionari sia riportato l’anno di corso; indicare che cosa sarebbe necessario modificare nello schema dimensionale e quali dati debbono essere disponibili nella staging area per poter riorganizzare il data mart.

Va cambiata la grana dei fatti, aggiungendo una dimensione anno di corso. La riorganizzazione è possibile solo se nella staging area sono disponibili dati con maggiore dettaglio rispetto a quelli utilizzati nello schema dimensionale preesistente

Domanda 4 (25%)
Considerare la seguente interrogazione in SQL:

```
SELECT A, D, H
FROM R, S, T
WHERE E = B AND C = G AND H < 40
```

definita con riferimento a tre relazioni, definite e frammentate come segue (per essere poi distribuite):

- $R(\underline{A}, C, E)$ frammentata orizzontalmente:
  - $R_a = \sigma_{C>100}(R)$;
  - $R_b = \sigma_{C\leq 100}(R)$
- $S(\underline{B}, D, F)$ frammentata verticalmente:
  - $S_a = \pi_{B,D}(S)$;
  - $S_b = \pi_{B,F}(S)$
- $T(\underline{G}, H)$ frammentata orizzontalmente:
  - $T_a = \sigma_{H>200}(T)$;
  - $T_b = \sigma_{H\leq 200}(T)$

Mostrare (ad esempio sotto forma di albero) l'espressione dell'algebra relazionale definita sui frammenti (cioè su R_a, R_b, S_a, ...) che realizza in tale interrogazione tralasciando i frammenti che non contribuiscono.

Join dell'unione di R_a e R_b con S_a e T_b oppure unione di due join ciascuno di tre relazioni

- R_a con S_a e con T_b
- R_b con S_a e con T_b

S_b e T_a non contribuiscono in alcun modo