Basi di dati, primo modulo Tecnologia delle basi di dati
30 giugno 2004 — Compito A
Tempo a disposizione: due ore

Domanda 1 (40%)
L'ufficio statistico dell'ateneo riceve spesso, dai presidi di facoltà e da altri docenti, richieste volte a conoscere:

- Il numero di immatricolati per la prima volta nel sistema universitario nazionale in un corso di laurea in un certo anno accademico (nel seguito si usa il termine *coorte* per fare riferimento a tale insieme di studenti);
- Per una specifica coorte e uno specifico anno accademico, il numero degli studenti che appartengono a ciascuna delle seguenti categorie:
  - iscritti a ciascuno degli anno di corso;
  - iscritti fuori corso;
  - laureati;
  - non più iscritti al corso di laurea, non laureati, ma iscritti ad altro corso dello stesso ateneo;
  - non più iscritti al corso di laurea, non laureati, e non iscritti ad altro corso dello stesso ateneo.
- Per una specifica coorte, con riferimento ad una certa data, con riferimento agli studenti ancora iscritti (in quella data), numero di studenti che hanno conseguito crediti in numero compreso in un certo intervallo (supponendo di interesse gli intervalli multipli di 10 e quelli multipli di 15).
- Per una specifica coorte, per uno specifico corso, il numero di studenti che hanno, ad una certa data, superato il relativo esame.
- Per una specifica coorte, per uno specifico corso, il voto medio riportato dagli studenti che hanno, ad una certa data, superato il relativo esame.

Progettare uno o più data mart che permettano di rispondere alle esigenze sopra formulate, supponendo che le informazioni necessarie allo scopo siano nella base di dati delle segreterie studenti. In particolare,

1. mostrare i frammenti di schema (ER e relazionale) della base di dati delle segreterie che si suppone di utilizzare come sorgente dei dati;
2. mostrare gli schemi a stella dei data mart;
3. mostrare (anche in modo schematico) le trasformazioni necessarie per passare dalla sorgente ai data mart.

Domanda 2 (15%)
Considerare le seguenti richieste di lettura e scrittura ricevute da un gestore del controllo di concorrenza basato su timestamp (assumendo che si tratti delle prime richieste ricevute dopo l'avvio del sistema):

$$r_1(z), w_1(z), r_2(x), r_8(x), r_5(x), r_3(v), r_4(v), w_3(v), w_4(v), w_7(x), r_6(x), w_9(x), w_8(x)$$

Indicare quali vengono accettate e quali rifiutate e, di conseguenza, quali transazioni vengono uccise.

Domanda 3 (10%)
Nel controllo di concorrenza basato su timestamp una transazione viene uccisa se essa, avendo un timestamp pari a *ts*, richiede una scrittura su un elemento x tale che $ts < \text{WTM}(x)$. Alcuni autori hanno notato che questa uccisione non è in effetti necessaria. Spiegare perché (commentando anche con riferimento alla risposta alla domanda precedente).

Domanda 4 (20%)
Si supponga di disporre di una base di dati con i saldi dei conti correnti gestiti dalle varie agenzie di una banca e di dover eseguire su di essa l'interrogazione che calcola, per ciascuna agenzia, la somma dei saldi dei conti correnti. Indicare (con un breve commento, non più di tre righe) 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'interrogazione è eseguita mentre vengono inseriti alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di acquisire informazioni anche approssimate sugli andamenti complessivi
2. l'interrogazione è eseguita mentre vengono modificati i saldi di alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di stilare una classifica delle agenzie, sulla base della somma dei saldi
3. l'interrogazione è eseguita mentre vengono modificati i saldi di tutti i conti correnti; l'operazione ha la finalità di acquisire informazioni anche approssimate sugli andamenti complessivi
4. l'interrogazione è eseguita in un momento in cui non ci sono aggiornamenti
5. l'interrogazione è eseguita mentre vengono inseriti alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di stilare una classifica delle agenzie, sulla base della somma dei saldi

Domanda 5 (15%)
Si consideri una relazione STUDENTE(Matricola,Cognome,Nome,DataNascita,Residenza) con un numero di ennuple pari a N e una dimensione di ciascuna ennupla (a lunghezza fissa) pari a L byte, di cui K per la chiave.
Si supponga di avere a disposizione un DBMS che permetta strutture fisiche disordinate (heap), ordinate e hash e che preveda la possibilità di definire indici secondari e operi su un sistema operativo che utilizza blocchi di dimensione B e con puntatori ai blocchi di P caratteri.
Si supponga che la relazione sia *utilizzata in sola lettura* e che siano le seguenti le operazioni principali:

1. ricerca sul numero di matricola con frequenza f_1
2. ricerca sul cognome anche approssimata (sottostringa iniziale) con frequenza f_2

Individuare per tale relazione le organizzazioni fisiche che possono essere ritenute valide sulla base di una analisi qualitativa e scegliere la migliore, sulla base di una analisi quantitativa, supponendo $N = 5.000.000$, $L = 125$, $K = 5$, $B = 1.000$, $P = 4$, $f_1 = 100$, $f_2 = 1.000$.

Basi di dati, primo modulo Tecnologia delle basi di dati
30 giugno 2004 — Compito B
Tempo a disposizione: due ore

Domanda 1 (40%)
L'ufficio statistico dell'ateneo riceve spesso, dai presidi di facoltà e da altri docenti, richieste volte a conoscere:

- Il numero di immatricolati per la prima volta nel sistema universitario nazionale in un corso di laurea in un certo anno accademico (nel seguito si usa il termine *coorte* per fare riferimento a tale insieme di studenti);
- Per una specifica coorte e uno specifico anno accademico, il numero degli studenti che appartengono a ciascuna delle seguenti categorie:
  - iscritti a ciascuno degli anno di corso;
  - iscritti fuori corso;
  - laureati;
  - non più iscritti al corso di laurea, non laureati, ma iscritti ad altro corso dello stesso ateneo;
  - non più iscritti al corso di laurea, non laureati, e non iscritti ad altro corso dello stesso ateneo.
- Per una specifica coorte, con riferimento ad una certa data, con riferimento agli studenti ancora iscritti (in quella data), numero di studenti che hanno conseguito crediti in numero compreso in un certo intervallo (supponendo di interesse gli intervalli multipli di 10 e quelli multipli di 15).
- Per una specifica coorte, per uno specifico corso, il numero di studenti che hanno, ad una certa data, superato il relativo esame.
- Per una specifica coorte, per uno specifico corso, il voto medio riportato dagli studenti che hanno, ad una certa data, superato il relativo esame.

Progettare uno o più data mart che permettano di rispondere alle esigenze sopra formulate, supponendo che le informazioni necessarie allo scopo siano nella base di dati delle segreterie studenti. In particolare,

1. mostrare i frammenti di schema (ER e relazionale) della base di dati delle segreterie che si suppone di utilizzare come sorgente dei dati;
2. mostrare gli schemi a stella dei data mart;
3. mostrare (anche in modo schematico) le trasformazioni necessarie per passare dalla sorgente ai data mart.

Domanda 2 (15%)
Considerare le seguenti richieste di lettura e scrittura ricevute da un gestore del controllo di concorrenza basato su timestamp (assumendo che si tratti delle prime richieste ricevute dopo l'avvio del sistema):

$$r_1(w), w_1(w), r_2(x), r_8(x), r_5(x), r_3(y), r_4(y), w_3(y), w_4(y), w_7(x), r_6(x), w_9(x), w_8(x)$$

Indicare quali vengono accettate e quali rifiutate e, di conseguenza, quali transazioni vengono uccise.

Domanda 3 (10%)
Nel controllo di concorrenza basato su timestamp una transazione viene uccisa se essa, avendo un timestamp pari a *ts*, richiede una scrittura su un elemento x tale che $ts < \textsc{wtm}(x)$. Alcuni autori hanno notato che questa uccisione non è in effetti necessaria. Spiegare perché (commentando anche con riferimento alla risposta alla domanda precedente).

Domanda 4 (20%)
Si supponga di disporre di una base di dati con i saldi dei conti correnti gestiti dalle varie agenzie di una banca e di dover eseguire su di essa l'interrogazione che calcola, per ciascuna agenzia, la somma dei saldi dei conti correnti. Indicare (con un breve commento, non più di tre righe) 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'interrogazione è eseguita mentre vengono modificati i saldi di tutti i conti correnti; l'operazione ha la finalità di acquisire informazioni anche approssimate sugli andamenti complessivi
2. l'interrogazione è eseguita mentre vengono inseriti alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di acquisire informazioni anche approssimate sugli andamenti complessivi
3. l'interrogazione è eseguita in un momento in cui non ci sono aggiornamenti
4. l'interrogazione è eseguita mentre vengono modificati i saldi di alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di stilare una classifica delle agenzie, sulla base della somma dei saldi
5. l'interrogazione è eseguita mentre vengono inseriti alcuni conti correnti (in ciascuna agenzia pochi rispetto a quelli già presenti); l'operazione ha la finalità di stilare una classifica delle agenzie, sulla base della somma dei saldi

Domanda 5 (15%)
Si consideri una relazione STUDENTE(Matricola,Cognome,Nome,DataNascita,Residenza) con un numero di ennuple pari a N e una dimensione di ciascuna ennupla (a lunghezza fissa) pari a L byte, di cui K per la chiave.
Si supponga di avere a disposizione un DBMS che permetta strutture fisiche disordinate (heap), ordinate e hash e che preveda la possibilità di definire indici secondari e operi su un sistema operativo che utilizza blocchi di dimensione B e con puntatori ai blocchi di P caratteri.
Si supponga che la relazione sia *utilizzata in sola lettura* e che siano le seguenti le operazioni principali:

1. ricerca sul numero di matricola con frequenza f_1
2. ricerca sul cognome anche approssimata (sottostringa iniziale) con frequenza f_2

Individuare per tale relazione le organizzazioni fisiche che possono essere ritenute valide sulla base di una analisi qualitativa e scegliere la migliore, sulla base di una analisi quantitativa, supponendo $N = 1.000.000$, $L = 125$, $K = 5$, $B = 1.000$, $P = 4$, $f_1 = 2.000$, $f_2 = 200$.