Basi di dati II — Esame — 16 febbraio 2016 — Compito A

Tempo a disposizione: due ore e quindici minuti.

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

Domanda 1 (20%) Si consideri una base di dati sulle seguenti relazioni, ognuna delle quali ha una struttura heap e un indice secondario sulla chiave (si suppongano le relazioni tutte molto grandi, con R_1 più piccola delle altre due)

- $R_1(\underline{A}BC)$
- $R_2(\underline{D}EF)$
- $R_3(\underline{G}HL)$

e con una vista definita come segue:

- `CREATE VIEW V AS SELECT * FROM (R1 JOIN R2 ON B=D) LEFT JOIN R3 ON C=G WHERE B>10`

In tale contesto, supponendo che il sistema esegua join solo con nested loop, utilizzando gli indici ove definiti, mostrare un possibile piano di esecuzione (in termini di operatori dell'algebra relazionale e loro realizzazioni, quindi indicando come vengono eseguiti i join) per ciascuna delle seguenti interrogazioni

```
SELECT A, C, E, F FROM V WHERE A<>25
```

```
SELECT A, B, G, H, L FROM V WHERE A=25
```

Nel caso in cui il sistema sia in grado di eseguire anche hash join, indicare quali dei join mostrati nelle risposte precedenti possa convenire eseguire con tale tecnica.

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 2 (20%) Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco. Quindi R1 occupa $B_1 = 6$ blocchi e R2 ne occupa $B_2 = 8$.

Relazione R1

20		21		22		23		24		25	
X01	AA	Y01	DA	Z03	AB	K03	AB	Z03	AB	Z03	AB
Y42	CA	X42	CC	W05	EF	W07	EF	W08	EF	W09	EF
W73	CC	W93	CB	X52	HA	X59	HA	X50	HA	X56	HA
Z55	GC	W54	LB	Y55	EA	Y54	EA	Y51	EA	Y57	EA

Relazione R2

40		41		42		43		44		45		46		47	
AA	3	BC	4	LB	7	AA	8	AC	3	EA	7	BA	5	EF	6
DA	7	GB	7	HB	3	EC	2	CB	5	LB	8	BB	4	GA	8

Si supponga di disporre di un buffer di p pagine.
Considerare l'esecuzione del join di R1 ed R2, sulla base dei valori del secondo attributo di R1 e del primo di R2, con il metodo nested loop senza utilizzo di indici. Supporre che non serva memorizzare il risultato e che quindi esso possa essere prodotto una ennupla alla volta (approccio "pipelining"). Rispondere alle domande seguenti, con riferimento a due casi diversi per il numero di pagine di buffer disponibili, $p = 2$ e $p = 8$

Indicare, nell'ordine, le prime quattro ennuple che vengono prodotte

$p = 2$:

$p = 8$:

Indicare gli indirizzi dei blocchi che si trovano nel buffer dopo che sono state prodotte le prime quattro ennuple.

$p = 2$:

$p = 8$:

Indicare gli indirizzi dei blocchi effettivamente letti da memoria secondaria e caricati nel buffer (nell'ordine) per produrre le prime quattro ennuple del risultato.

$p = 2$:

$p = 8$:

Indicare il numero complessivo di accessi a memoria secondaria necessari per eseguire il join (indicare formula e numero)

$p = 2$:

$p = 8$:

Domanda 3 (20%) Dimostrare, per grandi linee, che la conflict-seriallizzabilità implica la view-serializzabilità (ma non necessariamente viceversa) e che il 2PL stretto implica la conflict-serializzabilità (ma non necessariamente viceversa)

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 4 (20%) Si supponga che i dirigenti di una organizzazione siano spesso coinvolti in riunioni, che vengono organizzate secondo una procedura che si basa sul principio del commit a due fasi, nel senso che, se viene data disponibilità a partecipare, poi questa non può essere ritirata. Le comunicazioni fra i dirigenti possono essere sincrone o asincrone (ad esempio via telefono o via email) e può succedere che qualcuno sia irraggiungibile per un periodo anche lungo. Con riferimento a questo contesto, rispondere alle seguenti domande:

1. Supponendo che ogni dirigente abbia diverse segretarie, che si alternano nella gestione delle richieste e delle risposte, indicare che cosa è assolutamente necessario per coordinare le loro attività.

2. Si può osservare che questo contesto è in effetti poco adatto al 2PC, per via delle possibili "irraggiungibilità" e del fatto che le riunioni sono programmate ognuna in un certo orario. Spiegare brevemente perché questo problema sussiste qui e non sussiste nel 2PC per la gestione delle transazioni.

vspace4cm

3. Mostrare, sinteticamente, il susseguirsi delle azioni relative al tentativo di fissare una riunione da parte di un dirigente con altri due (contattati in parallelo), il primo dei quali dà la propria disponibilità e l'altro no (ma rispondendo subito).

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 5 (20%) Si consideri la seguente base di dati, relativa alle ricette acquisite da un insieme di farmacie:

- Ricette(Numero, CodFarmacia, CFPaziente, Data)
- Farmacie(CodFarmacia, Nome, Via, NumeroCivico, Città)
- ElementiRicetta(NumeroRicetta, CodFarmaco, Quantità)
- Farmaci(Codice, Descrizione, CodMolecola, CodCasa, Prezzo, Fascia)
- Molecole(CodMolecola, Descrizione)
- Pazienti(CF, Cognome, Nome, DataNascita, Via, NumeroCivico, Città)
- CaseFarmaceutiche(CodCasa, Nome)
- ASL(Codice, Nome)
- Territorio(Via, Città, NumeroCivico, 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 (molecola e casa farmaceutica);
- 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 — Esame — 16 febbraio 2016 — Compito B

Tempo a disposizione: due ore e quindici minuti.

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

Domanda 1 (20%) Si consideri una base di dati sulle seguenti relazioni, ognuna delle quali ha una struttura heap e un indice secondario sulla chiave (si suppongano le relazioni tutte molto grandi, con S_1 più piccola delle altre due)

- $S_1(\underline{A}BC)$
- $S_2(\underline{D}EF)$
- $S_3(\underline{G}HL)$

e con una vista definita come segue:

- `CREATE VIEW V AS SELECT * FROM (S1 JOIN S2 ON B=D) LEFT JOIN S3 ON C=G WHERE B>10`

In tale contesto, supponendo che il sistema esegua join solo con nested loop, utilizzando gli indici ove definiti, mostrare un possibile piano di esecuzione (in termini di operatori dell'algebra relazionale e loro realizzazioni, quindi indicando come vengono eseguiti i join) per ciascuna delle seguenti interrogazioni

```
SELECT A, B, H, L FROM V WHERE A<>25
```

```
SELECT A, B, C, E, F FROM V WHERE A=25
```

Nel caso in cui il sistema sia in grado di eseguire anche hash join, indicare quali dei join mostrati nelle risposte precedenti possa convenire eseguire con tale tecnica.

Basi di dati II — 16 febbraio 2016 — Compito B

Domanda 2 (20%) Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco. Quindi R1 occupa $N_1 = 6$ blocchi e R2 ne occupa $N_2 = 8$.

Relazione R1

40		41		42		43		44		45	
X01	AA	Y01	DA	Z03	AB	K03	AB	Z03	AB	Z03	AB
Y42	CA	X42	CC	W05	EF	W07	EF	W08	EF	W09	EF
W73	CC	W93	CB	X52	HA	X59	HA	X50	HA	X56	HA
Z55	GC	W54	LB	Y55	EA	Y54	EA	Y51	EA	Y57	EA

Relazione R2

50			51			52			53			54			55			56			57		
AA	3		BC	4		LB	7		AA	8		AC	3		EA	7		BA	5		EF	6	
DA	7		GB	7		HB	3		EC	2		CB	5		LB	8		BB	4		GA	8	

Si supponga di disporre di un buffer di p pagine.
Considerare l'esecuzione del join di R1 ed R2, sulla base dei valori del secondo attributo di R1 e del primo di R2, con il metodo nested loop senza utilizzo di indici. Supporre che non serva memorizzare il risultato e che quindi esso possa essere prodotto una ennupla alla volta (approccio "pipelining"). Rispondere alle domande seguenti, con riferimento a due casi diversi per il numero di pagine di buffer disponibili, $p = 2$ e $p = 8$

Indicare, nell'ordine, le prime quattro ennuple che vengono prodotte

$p = 2$:

$p = 8$:

Indicare gli indirizzi dei blocchi che si trovano nel buffer dopo che sono state prodotte le prime quattro ennuple.

$p = 2$:

$p = 8$:

Indicare gli indirizzi dei blocchi effettivamente letti da memoria secondaria e caricati nel buffer (nell'ordine) per produrre le prime quattro ennuple del risultato.

$p = 2$:

$p = 8$:

Indicare il numero complessivo di accessi a memoria secondaria necessari per eseguire il join (indicare formula e numero)

$p = 2$:

$p = 8$:

Basi di dati II — 16 febbraio 2016 — Compito B

Domanda 3 (20%) Dimostrare, per grandi linee, che la conflict-seriallizzabilità implica la view-serializzabilità (ma non necessariamente viceversa) e che il 2PL stretto implica la conflict-serializzabilità (ma non necessariamente viceversa)

Basi di dati II — 16 febbraio 2016 — Compito B

Domanda 4 (20%) Si supponga che i dirigenti di una organizzazione siano spesso coinvolti in riunioni, che vengono organizzate secondo una procedura che si basa sul principio del commit a due fasi, nel senso che, se viene data disponibilità a partecipare, poi questa non può essere ritirata. Le comunicazioni fra i dirigenti possono essere sincrone o asincrone (ad esempio via telefono o via email) e può succedere che qualcuno sia irraggiungibile per un periodo anche lungo. Con riferimento a questo contesto, rispondere alle seguenti domande:

1. Supponendo che ogni dirigente abbia diverse segretarie, che si alternano nella gestione delle richieste e delle risposte, indicare che cosa è assolutamente necessario per coordinare le loro attività.

2. Si può osservare che questo contesto è in effetti poco adatto al 2PC, per via delle possibili "irraggiungibilità" e del fatto che le riunioni sono programmate ognuna in un certo orario. Spiegare brevemente perché questo problema sussiste qui e non sussiste nel 2PC per la gestione delle transazioni.

vspace4cm

3. Mostrare, sinteticamente, il susseguirsi delle azioni relative al tentativo di fissare una riunione da parte di un dirigente con altri due (contattati in parallelo), il primo dei quali dà la propria disponibilità e l'altro no (ma rispondendo subito).

Basi di dati II — 16 febbraio 2016 — Compito B

Domanda 5 (20%) Si consideri la seguente base di dati, relativa alle ricette acquisite da un insieme di farmacie:

- Ricette(Numero, CodFarmacia, CFPaziente, Data)
- Farmacie(CodFarmacia, Nome, Via, NumeroCivico, Città)
- ElementiRicetta(NumeroRicetta, CodFarmaco, Quantità)
- Farmaci(Codice, Descrizione, CodMolecola, CodCasa, Prezzo, Fascia)
- Molecole(CodMolecola, Descrizione)
- Pazienti(CF, Cognome, Nome, DataNascita, Via, NumeroCivico, Città)
- CaseFarmaceutiche(CodCasa, Nome)
- ASL(Codice, Nome)
- Territorio(Via, Città, NumeroCivico, 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 (molecola e casa farmaceutica);
- 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 — Esame — 16 febbraio 2016 — Compito A

Cenni sulle soluzioni
(solo Compito A, le varianti del testo sono in rosso)

Tempo a disposizione: due ore e quindici minuti.

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

Domanda 1 (20%) Si consideri una base di dati sulle seguenti relazioni, ognuna delle quali ha una struttura heap e un indice secondario sulla chiave (si suppongano le relazioni tutte molto grandi, con R_1 più piccola delle altre due)

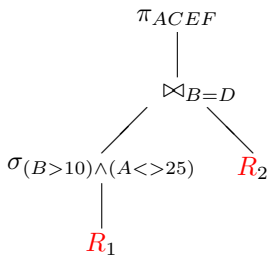
- $R_1(\underline{A}BC)$
- $R_2(\underline{D}EF)$
- $R_3(\underline{G}HL)$

e con una vista definita come segue:

- `CREATE VIEW V AS SELECT * FROM (R1 JOIN R2 ON B=D) LEFT JOIN R3 ON C=G WHERE B>10`

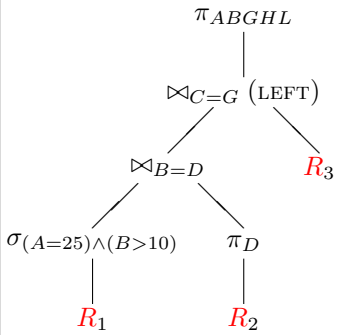
In tale contesto, supponendo che il sistema esegua join solo con nested loop, utilizzando gli indici ove definiti, mostrare un possibile piano di esecuzione (in termini di operatori dell'algebra relazionale e loro realizzazioni, quindi indicando come vengono eseguiti i join) per ciascuna delle seguenti interrogazioni

```
SELECT A, C, E, F FROM V WHERE A<>25
```



Nei join è indicata a sinistra la tabella esterna e viene eseguito un accesso diretto su quella interna.

```
SELECT A, B, G, H, L FROM V WHERE A=25
```



Nel caso in cui il sistema sia in grado di eseguire anche hash join, indicare quali dei join mostrati nelle risposte precedenti possa convenire eseguire con tale tecnica.

Nel primo caso probabilmente sì (anche se sarebbe necessario valutare le dimensioni) nel secondo certamente no, perchè sono coinvolte nel join solo una ennupla di R_2 e una di R_3, accessibili tramite indice.

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 2 (20%) Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco. Quindi R1 occupa $B_1 = 6$ blocchi e R2 ne occupa $B_2 = 8$.

Relazione R1

20			21			22			23			24			25		
	X01	AA		Y01	DA		Z03	AB		K03	AB		Z03	AB		Z03	AB
	Y42	CA		X42	CC		W05	EF		W07	EF		W08	EF		W09	EF
	W73	CC		W93	CB		X52	HA		X59	HA		X50	HA		X56	HA
	Z55	GC		W54	LB		Y55	EA		Y54	EA		Y51	EA		Y57	EA

Relazione R2

40			41			42			43			44			45			46			47		
	AA	3		BC	4		LB	7		AA	8		AC	3		EA	7		BA	5		EF	6
	DA	7		GB	7		HB	3		EC	2		CB	5		LB	8		BB	4		GA	8

Si supponga di disporre di un buffer di p pagine.
Considerare l'esecuzione del join di R1 ed R2, sulla base dei valori del secondo attributo di R1 e del primo di R2, con il metodo nested loop senza utilizzo di indici. Supporre che non serva memorizzare il risultato e che quindi esso possa essere prodotto una ennupla alla volta (approccio "pipelining"). Rispondere alle domande seguenti, con riferimento a due casi diversi per il numero di pagine di buffer disponibili, $p = 2$ e $p = 8$

Indicare, nell'ordine, le prime quattro ennuple che vengono prodotte

$p = 2$: (X01, AA, 3), (X01, AA, 8), (Y01, DA, 7), (W54, LB, 7)

$p = 8$: (X01, AA, 3), (Y01, DA, 7), (W54, LB, 7), (X01, AA, 8)

Indicare gli indirizzi dei blocchi che si trovano nel buffer dopo che sono state prodotte le prime quattro ennuple.

$p = 2$: 21, 42

$p = 8$: 20, 21, 22, 23, 24, 25, 42, 43

Indicare gli indirizzi dei blocchi effettivamente letti da memoria secondaria e caricati nel buffer (nell'ordine) per produrre le prime quattro ennuple del risultato.

$p = 2$: 20, poi tutti i blocchi di R2, poi 21, 40, 41, 42

$p = 8$: 20, 21, 22, 23, 24, 25, 40, 41, 42 , 43

Indicare il numero complessivo di accessi a memoria secondaria necessari per eseguire il join (indicare formula e numero)

$p = 2$: $B_1 + B_1 \times B_2 = 54$

$p = 8$: $B_1 + B_2 = 14$

Domanda 3 (20%) Dimostrare, per grandi linee, che la conflict-seriallizzabilità implica la view-serializzabilità (ma non necessariamente viceversa) e che il 2PL stretto implica la conflict-serializzabilità (ma non necessariamente viceversa)

Vedi libro o lucidi

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 4 (20%) Si supponga che i dirigenti di una organizzazione siano spesso coinvolti in riunioni, che vengono organizzate secondo una procedura che si basa sul principio del commit a due fasi, nel senso che, se viene data disponibilità a partecipare, poi questa non può essere ritirata. Le comunicazioni fra i dirigenti possono essere sincrone o asincrone (ad esempio via telefono o via email) e può succedere che qualcuno sia irraggiungibile per un periodo anche lungo. Con riferimento a questo contesto, rispondere alle seguenti domande:

1. Supponendo che ogni dirigente abbia diverse segretarie, che si alternano nella gestione delle richieste e delle risposte, indicare che cosa è assolutamente necessario per coordinare le loro attività.

Un "registro" o qualcosa del genere per ogni dirigente, che svolga il ruolo del log locale

2. Si può osservare che questo contesto è in effetti poco adatto al 2PC, per via delle possibili "irraggiungibilità" e del fatto che le riunioni sono programmate ognuna in un certo orario. Spiegare brevemente perché questo problema sussiste qui e non sussiste nel 2PC per la gestione delle transazioni.

Nel 2PL non è previsto alcun supporto per la scadenza temporale e il problema si pone soprattutto nella seconda fase: se un partecipante non riceve per tempo la conferma o la smentita per una riunione non sa che cosa fare

3. Mostrare, sinteticamente, il susseguirsi delle azioni relative al tentativo di fissare una riunione da parte di un dirigente con altri due (contattati in parallelo), il primo dei quali dà la propria disponibilità e l'altro no (ma rispondendo subito).

Basi di dati II — 16 febbraio 2016 — Compito A

Domanda 5 (20%) Si consideri la seguente base di dati, relativa alle ricette acquisite da un insieme di farmacie:

- Ricette(Numero, CodFarmacia, CFPaziente, Data)
- Farmacie(CodFarmacia, Nome, Via, NumeroCivico, Città)
- ElementiRicetta(NumeroRicetta, CodFarmaco, Quantità)
- Farmaci(Codice, Descrizione, CodMolecola, CodCasa, Prezzo, Fascia)
- Molecole(CodMolecola, Descrizione)
- Pazienti(CF, Cognome, Nome, DataNascita, Via, NumeroCivico, Città)
- CaseFarmaceutiche(CodCasa, Nome)
- ASL(Codice, Nome)
- Territorio(Via, Città, NumeroCivico, 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 (molecola e casa farmaceutica);
- 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: la grana scelta è "singole prescrizioni"

Schema dimensionale:

- FattiPrescrizioni(KData, KVersioneFarmaco, KRicetta, KASLfarmacia, KASLpaziente, KEtà, Quantità, Importo)
- DimFarmaci(KVersioneFarmaco, Codice, Descrizione, CodMolecola, DescrizioneMolecola, CodCasa, NomeCasa, Fascia)
- DimASL(KASL, CodiceASL, NomeASL)
- DimEtà(KEtà, Età, Fascia ...)
- DimData(KData, ...)

Commenti:

- sono indicate chiavi ad hoc per le dimensioni
- *Ricetta* è una dimensione "degenere," cioè senza attributi
- per la privatezza, si eliminano tutte le informazioni personali, indicando solo ASL del paziente e fascia d'età; quindi la grana scelta è "singole prescrizioni;" in effetti, vista la presenza di *KRicetta*, tutte le dimensioni, a parte *Ricetta* e *Farmaci*, sono secondarie
- *Farmaci* è una slowly changing dimension rispetto alla fascia
- si usa *Età* invece di fascia di età, per dare un po' di flessibilità
- è opportuno avere due viste su DimASL, poiché la dimensione è utilizzata due volte