Basi di dati II

Prova parziale —20 maggio 2013 — Compito A

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

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

Domanda 1 (25%)
Per ovviare alle conseguenze negative di un guasto del coordinatore, alcune implementazioni del 2PC prevedono la possibilità di comunicazione fra i partecipanti (mentre la versione base prevede solo comunicazione fra il coordinatore e ciascuno dei partecipanti). In particolare, un partecipante che abbia una transazione in stato di "ready" può chiedere agli altri partecipanti informazioni sullo stato di tale transazione (che può essere "prima-del-ready", "ready", "commit" o "abort") presso di loro. In tale contesto indicare, per ciascuna delle situazioni seguenti, se essa si può effettivamente verificare (giustificare brevemente la risposta) e quali conseguenze il partecipante ne può trarre:

1. uno o più partecipanti sono in uno stato"ready" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

2. uno o più partecipanti sono in uno stato"prima-del-ready" e uno o più partecipanti in uno stato "abort"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

3. uno o più partecipanti sono in uno stato"abort" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

4. uno o più partecipanti sono in uno stato"commit" e uno o più partecipanti in uno stato "prima-del-ready"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

Domanda 2 (15%)
Il *semijoin* è un'operazione simile al join, in cui del secondo operando interessano solo gli attributi di join. In concreto, se il join su $A_1 = A_2$ di $R_1(X_1)$ e $R_2(X_2)$ (con $A_1 \in X_1$, $A_2 \in X_2$ e $X_1 \cap X_2 = \emptyset$) è definito come

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = \{\, t \text{ su } X_1X_2 \mid \text{esistono } t_1 \in R_1 \text{ e } t_2 \in R_2 \text{ con } t[X_1] = t_1,\ t[X_2] = t_2 \text{ e } t_1[A_1] = t_2[A_2]\}$$

il semijoin corrispondente è definito come

$$R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2 = \{\, t_1 \mid t_1 \in R_1 \text{ ed esiste } t_2 \in R_2 \text{ con } t_1[A_1] = t_2[A_2]\}$$

Ad esempio:

IMP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1
cf_3	...	...	...	2

DIP

<u>CodD</u>	NomeD	CFDir	...
1	ABC	cf_1	...
3	DEF	cf_4	...

IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1

Dimostrare (anche in modo informale, ma ragionevolmente preciso) che, con riferimento alle definizioni date all'inizio, è sempre vero che

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = (R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2) \textbf{ Join}_{A_1=A_2} R_2$$

Per comodità, fare riferimento all'esempio e dimostrare che

$$\text{IMP } \textbf{Join}_{\text{CDip=CodD}} \text{ DIP} = (\text{IMP } \textbf{SemiJoin}_{\text{CDip=CodD}} \text{ DIP}) \textbf{ Join}_{\text{CDip=CodD}} \text{ DIP}$$

Domanda 3 (15%)
Il semijoin e la sua proprietà illustrata nell'esempio della domanda precedente sono alla base di tecniche per l'esecuzione di join in ambiente distribuito che cercano di minimizzare le quantità di dati che vengono trasferiti. Considerare una base di dati con lo schema mostrato nell'esempio, assumendo che le due relazioni siano memorizzate su nodi di rete diversi, come segue:

- IMP è memorizzata nel nodo A e ha $L_I = 200.000$ ennuple di $N_I = 100$ byte, di cui $N_{CF} = 10$ per il codice fiscale (attributo CF) e $N_{CD} = 4$ per il riferimento al dipartimento (attributo CDip)
- DIP è memorizzata nel nodo B e ha $L_D = 100$ ennuple di $N_D = 50$ byte, di cui (per coerenza con quanto sopra) $N_{CD} = 4$ per il codice (attributo CodD) e $N_{CF} = 10$ per il codice fiscale del direttore (attributo CFDir)

Indicare, per le due equivalenti strategie di esecuzione dell'interrogazione la quantità di dati che vengono trasferiti, supponendo che solo il 25% delle ennuple della relazione IMP partecipa effettivamente al join.

1. IMP **Join**$_{\text{CDip=CodD}}$ DIP (eseguita sul nodo B)

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

2. (IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP) **Join**$_{\text{CDip=CodD}}$ DIP (con il semijoin eseguito sul nodo A e il join sul nodo B);

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

Domanda 4 (30%)
Si consideri la base di dati seguente, relativa alle telefonate di una singola giornata effettuate da clienti di telefonia fissa di una azienda telefonica:

- TELEFONATE(Orario, Chiamante, Chiamato, Durata) dove Orario indica l'istante preciso (ora, minuti, secondi) di inizio della telefonata, Chiamante è un'utenza (cioè un numero, comprensivo di prefisso) dell'azienda e Chiamato è un'utenza, dell'azienda stessa oppure di un'altra azienda
- UTENZE(Numero, SubDistretto, PianoTariffario) (le utenze dell'azienda, tutte nazionali), con vincolo di riferimento da SubDistretto verso SUBDISTRETTI
- UTENZEESTERNE(Numero, Tipo, Azienda) (le utenze di altre aziende) dove Tipo assume uno dei tre valori: nazionale fissa, nazionale mobile, estera
- SUBDISTRETTI(Codice, Descrizione, Distretto) con vincolo di riferimento da Distretto verso DISTRETTI
- DISTRETTI(Codice, Descrizione, Provincia)

Con riferimento a tale base di dati, progettare uno schema dimensionale (specificando grana, misure e dimensioni e illustrando brevemente come le tabelle possono essere ottenute partendo dai dati disponibili) che permettano di rispondere facilmente ad interrogazioni quali ad esempio (la descrizione non ha pretesa di essere esaustiva):

1. calcolare il numero e la durata complessiva delle telefonate fatte da una specifica utenza (o da ciascuna utenza di un certo insieme, specificato attraverso il distretto, oppure il subdistretto oppure il piano tariffario), in ciascuna fascia oraria (0-6,6-12,...,18-24), in ciascun mese dell'anno, e in ciascun giorno della settimana, verso utenze dei vari tipi (nazionali fisse, nazionali mobili, estere)
2. calcolare il numero e la durata complessiva delle telefonate fatte da utenze di un certo distretto con un certo piano tariffario, verso utenze di ciascuno dei distretti (o verso utenze di altre aziende e di altri tipi), in ciascuna ora della giornata (0, 1, 2, ..., 23)

Si noti che:

- per ragioni di privacy la granularità non può essere eccessiva (ma deve permettere le interrogazioni sopra citate)
- a causa della portabilità dei numeri telefonici, un numero può passare da un operatore all'altro (e un'utenza può cambiare numero, per gestire i periodo transitori nel cambio di operatore); spiegare come si tiene conto di questo aspetto
- il piano tariffario di un'utenza può cambiare nel tempo; spiegare come si tiene conto di questo aspetto

(rispondere nel riquadro della pagina a fianco)

Domanda 5 (15%)
Elencare le più importanti motivazioni che portano, nei datawarehouse, a preferire l'utilizzo di identificatori ad-hoc, generati nella fase di caricamento, rispetto all'uso di chiavi provenienti dalle applicazioni.

Basi di dati II

Prova parziale —20 maggio 2013 — Compito B

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

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

Domanda 1 (25%)
Per ovviare alle conseguenze negative di un guasto del coordinatore, alcune implementazioni del 2PC prevedono la possibilità di comunicazione fra i partecipanti (mentre la versione base prevede solo comunicazione fra il coordinatore e ciascuno dei partecipanti). In particolare, un partecipante che abbia una transazione in stato di "ready" può chiedere agli altri partecipanti informazioni sullo stato di tale transazione (che può essere "prima-del-ready", "ready", "commit" o "abort") presso di loro. In tale contesto indicare, per ciascuna delle situazioni seguenti, se essa si può effettivamente verificare (giustificare brevemente la risposta) e quali conseguenze il partecipante ne può trarre:

1. uno o più partecipanti sono in uno stato"abort" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

2. uno o più partecipanti sono in uno stato"commit" e uno o più partecipanti in uno stato "prima-del-ready"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

3. uno o più partecipanti sono in uno stato"ready" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

4. uno o più partecipanti sono in uno stato"prima-del-ready" e uno o più partecipanti in uno stato "abort"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

Domanda 2 (15%)
Il *semijoin* è un'operazione simile al join, in cui del secondo operando interessano solo gli attributi di join. In concreto, se il join su $A_1 = A_2$ di $R_1(X_1)$ e $R_2(X_2)$ (con $A_1 \in X_1$, $A_2 \in X_2$ e $X_1 \cap X_2 = \emptyset$) è definito come

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = \{\, t \text{ su } X_1X_2 \mid \text{esistono } t_1 \in R_1 \text{ e } t_2 \in R_2 \text{ con } t[X_1] = t_1,\ t[X_2] = t_2 \text{ e } t_1[A_1] = t_2[A_2]\}$$

il semijoin corrispondente è definito come

$$R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2 = \{\, t_1 \mid t_1 \in R_1 \text{ ed esiste } t_2 \in R_2 \text{ con } t_1[A_1] = t_2[A_2]\}$$

Ad esempio:

IMP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1
cf_3	...	...	...	2

DIP

<u>CodD</u>	NomeD	CFDir	...
1	ABC	cf_1	...
3	DEF	cf_4	...

IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1

Dimostrare (anche in modo informale, ma ragionevolmente preciso) che, con riferimento alle definizioni date all'inizio, è sempre vero che

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = (R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2) \textbf{ Join}_{A_1=A_2} R_2$$

Per comodità, fare riferimento all'esempio e dimostrare che

$$\text{IMP } \textbf{Join}_{\text{CDip=CodD}} \text{ DIP} = (\text{IMP } \textbf{SemiJoin}_{\text{CDip=CodD}} \text{ DIP}) \textbf{ Join}_{\text{CDip=CodD}} \text{ DIP}$$

Domanda 3 (15%)
Il semijoin e la sua proprietà illustrata nell'esempio della domanda precedente sono alla base di tecniche per l'esecuzione di join in ambiente distribuito che cercano di minimizzare le quantità di dati che vengono trasferiti. Considerare una base di dati con lo schema mostrato nell'esempio, assumendo che le due relazioni siano memorizzate su nodi di rete diversi, come segue:

- IMP è memorizzata nel nodo A e ha $R_I = 400.000$ ennuple di $N_I = 100$ byte, di cui $N_{CF} = 10$ per il codice fiscale (attributo CF) e $N_{CD} = 4$ per il riferimento al dipartimento (attributo CDip)
- DIP è memorizzata nel nodo B e ha $R_D = 100$ ennuple di $N_D = 50$ byte, di cui (per coerenza con quanto sopra) $N_{CD} = 4$ per il codice (attributo CodD) e $N_{CF} = 10$ per il codice fiscale del direttore (attributo CFDir)

Indicare, per le due equivalenti strategie di esecuzione dell'interrogazione la quantità di dati che vengono trasferiti, supponendo che solo il 25% delle ennuple della relazione IMP partecipa effettivamente al join.

1. IMP **Join**$_{\text{CDip=CodD}}$ DIP (eseguita sul nodo B)

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

2. (IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP) **Join**$_{\text{CDip=CodD}}$ DIP (con il semijoin eseguito sul nodo A e il join sul nodo B);

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

Domanda 4 (30%)
Si consideri la base di dati seguente, relativa alle telefonate di una singola giornata effettuate da clienti di telefonia fissa di una azienda telefonica:

- TELEFONATE(Orario, Chiamante, Chiamato, Durata) dove Orario indica l'istante preciso (ora, minuti, secondi) di inizio della telefonata, Chiamante è un'utenza (cioè un numero, comprensivo di prefisso) dell'azienda e Chiamato è un'utenza, dell'azienda stessa oppure di un'altra azienda
- UTENZE(Numero, SubDistretto, PianoTariffario) (le utenze dell'azienda, tutte nazionali), con vincolo di riferimento da SubDistretto verso SUBDISTRETTI
- UTENZEESTERNE(Numero, Tipo, Azienda) (le utenze di altre aziende) dove Tipo assume uno dei tre valori: nazionale fissa, nazionale mobile, estera
- SUBDISTRETTI(Codice, Descrizione, Distretto) con vincolo di riferimento da Distretto verso DISTRETTI
- DISTRETTI(Codice, Descrizione, Provincia)

Con riferimento a tale base di dati, progettare uno schema dimensionale (specificando grana, misure e dimensioni e illustrando brevemente come le tabelle possono essere ottenute partendo dai dati disponibili) che permettano di rispondere facilmente ad interrogazioni quali ad esempio (la descrizione non ha pretesa di essere esaustiva):

1. calcolare il numero e la durata complessiva delle telefonate fatte da una specifica utenza (o da ciascuna utenza di un certo insieme, specificato attraverso il distretto, oppure il subdistretto oppure il piano tariffario), in ciascuna fascia oraria (0-6,6-12,...,18-24), in ciascun mese dell'anno, e in ciascun giorno della settimana, verso utenze dei vari tipi (nazionali fisse, nazionali mobili, estere)
2. calcolare il numero e la durata complessiva delle telefonate fatte da utenze di un certo distretto con un certo piano tariffario, verso utenze di ciascuno dei distretti (o verso utenze di altre aziende e di altri tipi), in ciascuna ora della giornata (0, 1, 2, ..., 23)

Si noti che:

- per ragioni di privacy la granularità non può essere eccessiva (ma deve permettere le interrogazioni sopra citate)
- a causa della portabilità dei numeri telefonici, un numero può passare da un operatore all'altro (e un'utenza può cambiare numero, per gestire i periodo transitori nel cambio di operatore); spiegare come si tiene conto di questo aspetto
- il piano tariffario di un'utenza può cambiare nel tempo; spiegare come si tiene conto di questo aspetto

(rispondere nel riquadro della pagina a fianco)

Domanda 5 (15%)
Elencare le più importanti motivazioni che portano a costruire datawarehouse come basi di dati separate rispetto a quelle utilizzate nelle applicazioni che gestiscono le attività ordinarie.

Basi di dati II

Prova parziale —20 maggio 2013 — Compito C

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

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

Domanda 1 (25%)
Per ovviare alle conseguenze negative di un guasto del coordinatore, alcune implementazioni del 2PC prevedono la possibilità di comunicazione fra i partecipanti (mentre la versione base prevede solo comunicazione fra il coordinatore e ciascuno dei partecipanti). In particolare, un partecipante che abbia una transazione in stato di "ready" può chiedere agli altri partecipanti informazioni sullo stato di tale transazione (che può essere "prima-del-ready", "ready", "commit" o "abort") presso di loro. In tale contesto indicare, per ciascuna delle situazioni seguenti, se essa si può effettivamente verificare (giustificare brevemente la risposta) e quali conseguenze il partecipante ne può trarre:

1. uno o più partecipanti sono in uno stato"commit" e uno o più partecipanti in uno stato "prima-del-ready"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

2. uno o più partecipanti sono in uno stato"ready" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

3. uno o più partecipanti sono in uno stato"prima-del-ready" e uno o più partecipanti in uno stato "abort"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

4. uno o più partecipanti sono in uno stato"abort" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

Domanda 2 (15%)
Il *semijoin* è un'operazione simile al join, in cui del secondo operando interessano solo gli attributi di join. In concreto, se il join su $A_1 = A_2$ di $R_1(X_1)$ e $R_2(X_2)$ (con $A_1 \in X_1$, $A_2 \in X_2$ e $X_1 \cap X_2 = \emptyset$) è definito come

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = \{\, t \text{ su } X_1X_2 \mid \text{esistono } t_1 \in R_1 \text{ e } t_2 \in R_2 \text{ con } t[X_1] = t_1,\ t[X_2] = t_2 \text{ e } t_1[A_1] = t_2[A_2]\}$$

il semijoin corrispondente è definito come

$$R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2 = \{\, t_1 \mid t_1 \in R_1 \text{ ed esiste } t_2 \in R_2 \text{ con } t_1[A_1] = t_2[A_2]\}$$

Ad esempio:

IMP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1
cf_3	...	...	...	2

DIP

<u>CodD</u>	NomeD	CFDir	...
1	ABC	cf_1	...
3	DEF	cf_4	...

IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1

Dimostrare (anche in modo informale, ma ragionevolmente preciso) che, con riferimento alle definizioni date all'inizio, è sempre vero che

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = (R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2) \textbf{ Join}_{A_1=A_2} R_2$$

Per comodità, fare riferimento all'esempio e dimostrare che

$$\text{IMP } \textbf{Join}_{\text{CDip=CodD}} \text{ DIP} = (\text{IMP } \textbf{SemiJoin}_{\text{CDip=CodD}} \text{ DIP}) \textbf{ Join}_{\text{CDip=CodD}} \text{ DIP}$$

Domanda 3 (15%)
Il semijoin e la sua proprietà illustrata nell'esempio della domanda precedente sono alla base di tecniche per l'esecuzione di join in ambiente distribuito che cercano di minimizzare le quantità di dati che vengono trasferiti. Considerare una base di dati con lo schema mostrato nell'esempio, assumendo che le due relazioni siano memorizzate su nodi di rete diversi, come segue:

- IMP è memorizzata nel nodo A e ha $N_I = 200.000$ ennuple di $L_I = 100$ byte, di cui $L_{CF} = 10$ per il codice fiscale (attributo CF) e $L_{CD} = 4$ per il riferimento al dipartimento (attributo CDip)
- DIP è memorizzata nel nodo B e ha $N_D = 100$ ennuple di $L_D = 50$ byte, di cui (per coerenza con quanto sopra) $L_{CD} = 4$ per il codice (attributo CodD) e $L_{CF} = 10$ per il codice fiscale del direttore (attributo CFDir)

Indicare, per le due equivalenti strategie di esecuzione dell'interrogazione la quantità di dati che vengono trasferiti, supponendo che solo il 25% delle ennuple della relazione IMP partecipa effettivamente al join.

1. IMP $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (eseguita sul nodo B)

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

2. (IMP $\mathbf{SemiJoin}_{\text{CDip=CodD}}$ DIP) $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (con il semijoin eseguito sul nodo A e il join sul nodo B);

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

Domanda 4 (30%)
Si consideri la base di dati seguente, relativa alle telefonate di una singola giornata effettuate da clienti di telefonia fissa di una azienda telefonica:

- TELEFONATE(Orario, Chiamante, Chiamato, Durata) dove Orario indica l'istante preciso (ora, minuti, secondi) di inizio della telefonata, Chiamante è un'utenza (cioè un numero, comprensivo di prefisso) dell'azienda e Chiamato è un'utenza, dell'azienda stessa oppure di un'altra azienda
- UTENZE(Numero, SubDistretto, PianoTariffario) (le utenze dell'azienda, tutte nazionali), con vincolo di riferimento da SubDistretto verso SUBDISTRETTI
- UTENZEESTERNE(Numero, Tipo, Azienda) (le utenze di altre aziende) dove Tipo assume uno dei tre valori: nazionale fissa, nazionale mobile, estera
- SUBDISTRETTI(Codice, Descrizione, Distretto) con vincolo di riferimento da Distretto verso DISTRETTI
- DISTRETTI(Codice, Descrizione, Provincia)

Con riferimento a tale base di dati, progettare uno schema dimensionale (specificando grana, misure e dimensioni e illustrando brevemente come le tabelle possono essere ottenute partendo dai dati disponibili) che permettano di rispondere facilmente ad interrogazioni quali ad esempio (la descrizione non ha pretesa di essere esaustiva):

1. calcolare il numero e la durata complessiva delle telefonate fatte da una specifica utenza (o da ciascuna utenza di un certo insieme, specificato attraverso il distretto, oppure il subdistretto oppure il piano tariffario), in ciascuna fascia oraria (0-6,6-12,...,18-24), in ciascun mese dell'anno, e in ciascun giorno della settimana, verso utenze dei vari tipi (nazionali fisse, nazionali mobili, estere)
2. calcolare il numero e la durata complessiva delle telefonate fatte da utenze di un certo distretto con un certo piano tariffario, verso utenze di ciascuno dei distretti (o verso utenze di altre aziende e di altri tipi), in ciascuna ora della giornata (0, 1, 2, ..., 23)

Si noti che:

- per ragioni di privacy la granularità non può essere eccessiva (ma deve permettere le interrogazioni sopra citate)
- a causa della portabilità dei numeri telefonici, un numero può passare da un operatore all'altro (e un'utenza può cambiare numero, per gestire i periodo transitori nel cambio di operatore); spiegare come si tiene conto di questo aspetto
- il piano tariffario di un'utenza può cambiare nel tempo; spiegare come si tiene conto di questo aspetto

(rispondere nel riquadro della pagina a fianco)

Domanda 5 (15%)
Elencare le più importanti motivazioni che portano, nei datawarehouse, a preferire l'utilizzo di identificatori ad-hoc, generati nella fase di caricamento, rispetto all'uso di chiavi provenienti dalle applicazioni.

Basi di dati II

Prova parziale —20 maggio 2013 — Compito D

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

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

Domanda 1 (25%)
Per ovviare alle conseguenze negative di un guasto del coordinatore, alcune implementazioni del 2PC prevedono la possibilità di comunicazione fra i partecipanti (mentre la versione base prevede solo comunicazione fra il coordinatore e ciascuno dei partecipanti). In particolare, un partecipante che abbia una transazione in stato di "ready" può chiedere agli altri partecipanti informazioni sullo stato di tale transazione (che può essere "prima-del-ready", "ready", "commit" o "abort") presso di loro. In tale contesto indicare, per ciascuna delle situazioni seguenti, se essa si può effettivamente verificare (giustificare brevemente la risposta) e quali conseguenze il partecipante ne può trarre:

1. uno o più partecipanti sono in uno stato"prima-del-ready" e uno o più partecipanti in uno stato "abort"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

2. uno o più partecipanti sono in uno stato"abort" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

3. uno o più partecipanti sono in uno stato"commit" e uno o più partecipanti in uno stato "prima-del-ready"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

4. uno o più partecipanti sono in uno stato"ready" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)?
Perchè?

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?

Domanda 2 (15%)
Il *semijoin* è un'operazione simile al join, in cui del secondo operando interessano solo gli attributi di join. In concreto, se il join su $A_1 = A_2$ di $R_1(X_1)$ e $R_2(X_2)$ (con $A_1 \in X_1$, $A_2 \in X_2$ e $X_1 \cap X_2 = \emptyset$) è definito come

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = \{\, t \text{ su } X_1X_2 \mid \text{esistono } t_1 \in R_1 \text{ e } t_2 \in R_2 \text{ con } t[X_1] = t_1,\ t[X_2] = t_2 \text{ e } t_1[A_1] = t_2[A_2]\}$$

il semijoin corrispondente è definito come

$$R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2 = \{\, t_1 \mid t_1 \in R_1 \text{ ed esiste } t_2 \in R_2 \text{ con } t_1[A_1] = t_2[A_2]\}$$

Ad esempio:

IMP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1
cf_3	...	...	...	2

DIP

<u>CodD</u>	NomeD	CFDir	...
1	ABC	cf_1	...
3	DEF	cf_4	...

IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1

Dimostrare (anche in modo informale, ma ragionevolmente preciso) che, con riferimento alle definizioni date all'inizio, è sempre vero che

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = (R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2) \textbf{ Join}_{A_1=A_2} R_2$$

Per comodità, fare riferimento all'esempio e dimostrare che

$$\text{IMP } \textbf{Join}_{\text{CDip=CodD}} \text{ DIP} = (\text{IMP } \textbf{SemiJoin}_{\text{CDip=CodD}} \text{ DIP}) \textbf{ Join}_{\text{CDip=CodD}} \text{ DIP}$$

Domanda 3 (15%)
Il semijoin e la sua proprietà illustrata nell'esempio della domanda precedente sono alla base di tecniche per l'esecuzione di join in ambiente distribuito che cercano di minimizzare le quantità di dati che vengono trasferiti. Considerare una base di dati con lo schema mostrato nell'esempio, assumendo che le due relazioni siano memorizzate su nodi di rete diversi, come segue:

- IMP è memorizzata nel nodo A e ha $R_I = 400.000$ ennuple di $L_I = 100$ byte, di cui $L_{CF} = 10$ per il codice fiscale (attributo CF) e $L_{CD} = 4$ per il riferimento al dipartimento (attributo CDip)
- DIP è memorizzata nel nodo B e ha $R_D = 100$ ennuple di $L_D = 50$ byte, di cui (per coerenza con quanto sopra) $L_{CD} = 4$ per il codice (attributo CodD) e $L_{CF} = 10$ per il codice fiscale del direttore (attributo CFDir)

Indicare, per le due equivalenti strategie di esecuzione dell'interrogazione la quantità di dati che vengono trasferiti, supponendo che solo il 25% delle ennuple della relazione IMP partecipa effettivamente al join.

1. IMP $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (eseguita sul nodo B)

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

2. (IMP $\mathbf{SemiJoin}_{\text{CDip=CodD}}$ DIP) $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (con il semijoin eseguito sul nodo A e il join sul nodo B);

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

Domanda 4 (30%)
Si consideri la base di dati seguente, relativa alle telefonate di una singola giornata effettuate da clienti di telefonia fissa di una azienda telefonica:

- TELEFONATE(Orario, Chiamante, Chiamato, Durata) dove Orario indica l'istante preciso (ora, minuti, secondi) di inizio della telefonata, Chiamante è un'utenza (cioè un numero, comprensivo di prefisso) dell'azienda e Chiamato è un'utenza, dell'azienda stessa oppure di un'altra azienda
- UTENZE(Numero, SubDistretto, PianoTariffario) (le utenze dell'azienda, tutte nazionali), con vincolo di riferimento da SubDistretto verso SUBDISTRETTI
- UTENZEESTERNE(Numero, Tipo, Azienda) (le utenze di altre aziende) dove Tipo assume uno dei tre valori: nazionale fissa, nazionale mobile, estera
- SUBDISTRETTI(Codice, Descrizione, Distretto) con vincolo di riferimento da Distretto verso DISTRETTI
- DISTRETTI(Codice, Descrizione, Provincia)

Con riferimento a tale base di dati, progettare uno schema dimensionale (specificando grana, misure e dimensioni e illustrando brevemente come le tabelle possono essere ottenute partendo dai dati disponibili) che permettano di rispondere facilmente ad interrogazioni quali ad esempio (la descrizione non ha pretesa di essere esaustiva):

1. calcolare il numero e la durata complessiva delle telefonate fatte da una specifica utenza (o da ciascuna utenza di un certo insieme, specificato attraverso il distretto, oppure il subdistretto oppure il piano tariffario), in ciascuna fascia oraria (0-6,6-12,...,18-24), in ciascun mese dell'anno, e in ciascun giorno della settimana, verso utenze dei vari tipi (nazionali fisse, nazionali mobili, estere)
2. calcolare il numero e la durata complessiva delle telefonate fatte da utenze di un certo distretto con un certo piano tariffario, verso utenze di ciascuno dei distretti (o verso utenze di altre aziende e di altri tipi), in ciascuna ora della giornata (0, 1, 2, ..., 23)

Si noti che:

- per ragioni di privacy la granularità non può essere eccessiva (ma deve permettere le interrogazioni sopra citate)
- a causa della portabilità dei numeri telefonici, un numero può passare da un operatore all'altro (e un'utenza può cambiare numero, per gestire i periodo transitori nel cambio di operatore); spiegare come si tiene conto di questo aspetto
- il piano tariffario di un'utenza può cambiare nel tempo; spiegare come si tiene conto di questo aspetto

(rispondere nel riquadro della pagina a fianco)

Domanda 5 (15%)
Elencare le più importanti motivazioni che portano a costruire datawarehouse come basi di dati separate rispetto a quelle utilizzate nelle applicazioni che gestiscono le attività ordinarie.

Basi di dati II

Prova parziale —20 maggio 2013 — Compito A

Cenni sulle soluzioni

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

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

Domanda 1 (25%)
Per ovviare alle conseguenze negative di un guasto del coordinatore, alcune implementazioni del 2PC prevedono la possibilità di comunicazione fra i partecipanti (mentre la versione base prevede solo comunicazione fra il coordinatore e ciascuno dei partecipanti). In particolare, un partecipante che abbia una transazione in stato di "ready" può chiedere agli altri partecipanti informazioni sullo stato di tale transazione (che può essere "prima-del-ready", "ready", "commit" o "abort") presso di loro. In tale contesto indicare, per ciascuna delle situazioni seguenti, se essa si può effettivamente verificare (giustificare brevemente la risposta) e quali conseguenze il partecipante ne può trarre:

1. uno o più partecipanti sono in uno stato"ready" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)? *Risposta:* SÌ
Perchè?
Risposta Alcuni nodi hanno ricevuto il messaggio di commit e altri no

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?
Risposta: La transazione è certamente andata in commit

2. uno o più partecipanti sono in uno stato"prima-del-ready" e uno o più partecipanti in uno stato "abort"

Si può verificare (SÌ o NO)? *Risposta:* SÌ
Perchè?
Risposta L'abort può essere deciso in qualunque momento, escluso il caso in cui sia stato già deciso un commit.

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?
Risposta: La transazione è certamente andata in abort

3. uno o più partecipanti sono in uno stato"abort" e uno o più partecipanti in uno stato "commit"

Si può verificare (SÌ o NO)? *Risposta:* NO
Perchè?
Risposta Lo stato di commit o abort può essere raggiunto solo a seguito di una decisione del coordinatore in tale direzione (l'una o l'altra) e il coordinatore non può avere preso entrambe le decisioni.

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?
Risposta: Non applicabile

4. uno o più partecipanti sono in uno stato"commit" e uno o più partecipanti in uno stato "prima-del-ready"

Si può verificare (SÌ o NO)? *Risposta:* NO
Perchè?
Risposta Per esserci un commit, tutti i partecipanti debbono avere inviato il messaggio di ready e quindi nessuno può più trovarsi in uno stato "prima-del-ready".

Quali conseguenze può trarre il partecipante che ha ricevuto queste informazioni?
Risposta: Non applicabile

Domanda 2 (15%)
Il *semijoin* è un'operazione simile al join, in cui del secondo operando interessano solo gli attributi di join. In concreto, se il join su $A_1 = A_2$ di $R_1(X_1)$ e $R_2(X_2)$ (con $A_1 \in X_1$, $A_2 \in X_2$ e $X_1 \cap X_2 = \emptyset$) è definito come

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = \{\, t \text{ su } X_1X_2 \mid \text{esistono } t_1 \in R_1 \text{ e } t_2 \in R_2 \text{ con } t[X_1] = t_1,\ t[X_2] = t_2 \text{ e } t_1[A_1] = t_2[A_2]\}$$

il semijoin corrispondente è definito come

$$R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2 = \{\, t_1 \mid t_1 \in R_1 \text{ ed esiste } t_2 \in R_2 \text{ con } t_1[A_1] = t_2[A_2]\}$$

Ad esempio:

IMP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1
cf_3	...	...	...	2

DIP

<u>CodD</u>	NomeD	CFDir	...
1	ABC	cf_1	...
3	DEF	cf_4	...

IMP **SemiJoin**$_{\text{CDip=CodD}}$ DIP

<u>CF</u>	Cogn.	Nome	...	CDip
cf_1	Bini	Gino	...	1
cf_2	...	...	...	1

Dimostrare (anche in modo informale, ma ragionevolmente preciso) che, con riferimento alle definizioni date all'inizio, è sempre vero che

$$R_1 \textbf{ Join}_{A_1=A_2} R_2 = (R_1 \textbf{ SemiJoin}_{A_1=A_2} R_2) \textbf{ Join}_{A_1=A_2} R_2$$

Per comodità, fare riferimento all'esempio e dimostrare che

$$\text{IMP } \textbf{Join}_{\text{CDip=CodD}} \text{ DIP} = (\text{IMP } \textbf{SemiJoin}_{\text{CDip=CodD}} \text{ DIP}) \textbf{ Join}_{\text{CDip=CodD}} \text{ DIP}$$

Cenni sulla soluzione Dimostriamo nei due versi

- se una ennupla t appartiene al join a primo membro, allora esistono due ennuple t_1 e t_2 una per ciascuna delle due relazioni, sulla base delle quali essa viene costruita; in tal caso, t_1 appartiene al semijoin (a secondo membro), perché esiste t_2 nella seconda relazione e quindi poi partecipa al successivo join, contribuendo al risultato insieme a t_2, generando proprio t
- se una ennupla t appartiene al join a secondo membro, allora esistono una ennupla t_s nel semijoin e una ennupla t_2 nella seconda relazione sulla base delle quali essa viene costruita; se t_s appartiene al semijoin, allora appartiene anche alla prima relazione e quindi, insieme a t_2, può generare t nel join a primo membro

Domanda 3 (15%)
Il semijoin e la sua proprietà illustrata nell'esempio della domanda precedente sono alla base di tecniche per l'esecuzione di join in ambiente distribuito che cercano di minimizzare le quantità di dati che vengono trasferiti. Considerare una base di dati con lo schema mostrato nell'esempio, assumendo che le due relazioni siano memorizzate su nodi di rete diversi, come segue:

- IMP è memorizzata nel nodo A e ha $L_I = 200.000$ ennuple di $N_I = 100$ byte, di cui $N_{CF} = 10$ per il codice fiscale (attributo CF) e $N_{CD} = 4$ per il riferimento al dipartimento (attributo CDip)
- DIP è memorizzata nel nodo B e ha $L_D = 100$ ennuple di $N_D = 50$ byte, di cui (per coerenza con quanto sopra) $N_{CD} = 4$ per il codice (attributo CodD) e $N_{CF} = 10$ per il codice fiscale del direttore (attributo CFDir)

Indicare, per le due equivalenti strategie di esecuzione dell'interrogazione la quantità di dati che vengono trasferiti, supponendo che solo il 25% delle ennuple della relazione IMP partecipa effettivamente al join.

1. IMP $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (eseguita sul nodo B)

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

$L_I \times N_I = 20.000.000$ (si deve spostare tutta la relazione IMP da A a B)

2. (IMP $\mathbf{SemiJoin}_{\text{CDip=CodD}}$ DIP) $\mathbf{Join}_{\text{CDip=CodD}}$ DIP (con il semijoin eseguito sul nodo A e il join sul nodo B);

quantità di dati trasferiti (indicare formula, con una breve spiegazione, e valore numerico):

$L_D \times N_C D = 100 \times 4 = 400$ byte (la proiezione di DIP da B ad A)
$1/4 \times L_I \times N_I = 5.000.000$ byte (il 25% di IMP da A a B)
In totale circa 5.000.000 byte

Domanda 4 (30%)
Si consideri la base di dati seguente, relativa alle telefonate di una singola giornata effettuate da clienti di telefonia fissa di una azienda telefonica:

- TELEFONATE(Orario, Chiamante, Chiamato, Durata) dove Orario indica l'istante preciso (ora, minuti, secondi) di inizio della telefonata, Chiamante è un'utenza (cioè un numero, comprensivo di prefisso) dell'azienda e Chiamato è un'utenza, dell'azienda stessa oppure di un'altra azienda
- UTENZE(Numero, SubDistretto, PianoTariffario) (le utenze dell'azienda, tutte nazionali), con vincolo di riferimento da SubDistretto verso SUBDISTRETTI
- UTENZEESTERNE(Numero, Tipo, Azienda) (le utenze di altre aziende) dove Tipo assume uno dei tre valori: nazionale fissa, nazionale mobile, estera
- SUBDISTRETTI(Codice, Descrizione, Distretto) con vincolo di riferimento da Distretto verso DISTRETTI
- DISTRETTI(Codice, Descrizione, Provincia)

Con riferimento a tale base di dati, progettare uno schema dimensionale (specificando grana, misure e dimensioni e illustrando brevemente come le tabelle possono essere ottenute partendo dai dati disponibili) che permettano di rispondere facilmente ad interrogazioni quali ad esempio (la descrizione non ha pretesa di essere esaustiva):

1. calcolare il numero e la durata complessiva delle telefonate fatte da una specifica utenza (o da ciascuna utenza di un certo insieme, specificato attraverso il distretto, oppure il subdistretto oppure il piano tariffario), in ciascuna fascia oraria (0-6,6-12,...,18-24), in ciascun mese dell'anno, e in ciascun giorno della settimana, verso utenze dei vari tipi (nazionali fisse, nazionali mobili, estere)
2. calcolare il numero e la durata complessiva delle telefonate fatte da utenze di un certo distretto con un certo piano tariffario, verso utenze di ciascuno dei distretti (o verso utenze di altre aziende e di altri tipi), in ciascuna ora della giornata (0, 1, 2, ..., 23)

Si noti che:

- per ragioni di privacy la granularità non può essere eccessiva (ma deve permettere le interrogazioni sopra citate)
- a causa della portabilità dei numeri telefonici, un numero può passare da un operatore all'altro (e un'utenza può cambiare numero, per gestire i periodo transitori nel cambio di operatore); spiegare come si tiene conto di questo aspetto
- il piano tariffario di un'utenza può cambiare nel tempo; spiegare come si tiene conto di questo aspetto

(rispondere nel riquadro della pagina a fianco)

Notare che il chiamante deve essere per forza dell'azienda che gestisce il sistema, mentre il chiamato può essere di altra azienda. Peraltro, i dettagli del chiamato, nelle specifiche, interessano molto poco (al massimo interessano i distretti). Quindi sono opportune die dimensioni diverse, per chiamante chiamato.

- FattiTelefonateUtente(KUtente, KFasciaOrariaG, KMese, KGiornoSettimana, KDistrettoDest, Numero, Durata)
- Utente(KUtente, Numero, KSubDistretto, Subdistretto, ..., KDistretto, Distretto, ..., PianoTariffario, ...) attenzione al piano tariffario ...
- FasciaOrariaG(KFasciaOrariaG, FasciaDi6Ore, ...)
- Mese(KMese, Mese, ..., Trimestre, Anno, ...)
- GiornoSettimana(KGiornoSettimana, Giorno, ...)
- DistrettoDest(KDistrettoDest, ...)

Commenti:

- sono indicate chiavi ad hoc per le dimensioni
- per la privatezza, si include solo ciò che è espressamente richiesto e niente di più probabilmente è meglio non avere la data
- il distretto dei destinatari include, con un artificio, le utenze delle altre aziende e degli altri tipi
- la tabella dei fatti si calcola con una aggregazione ...

Domanda 5 (15%)
Elencare le più importanti motivazioni che portano, nei datawarehouse, a preferire l'utilizzo di identificatori ad-hoc, generati nella fase di caricamento, rispetto all'uso di chiavi provenienti dalle applicazioni.

Vedere il materiale didattico