Numéro d'anonymat :

Systèmes de Gestion de Bases de Données – 3I009

EXAMEN - 2è session du 10 juin 2016
Durée : 2 heures – CORRIGÉ
Documents autorisés

Les téléphones mobiles doivent être éteints et rangés dans les sacs. Le barème sur 20 points (20 questions) n'a qu'une valeur indicative.

1 Indexation : arbres B+ et tables de hachage (3 pts)

Dans cet exercice, on considère des arbres B+ d'ordre 2 (les noeuds et les feuilles ont entre 2 et 4 valeurs). La racine a entre 1 et 4 valeurs. Soit un arbre A1, contenant les valeurs suivantes : 2, 9, 13, 16, 18, 23, 24, 27, 30, 38, 41, 42, 49, 55, 58.

Question 1 (1 point)
Dessinez l'arbre A1, sachant que la racine contient les valeurs 18, 26, 34, 55.

Réponse :

Solution:
Racine (18, 26, 34, 55)
F1 (2, 9, 13, 16), F2 (18, 23, 24), F3 (27, 30), F4 (38, 41, 42, 49), F5 (55, 58)

Question 2 (1 point)
On insère successivement dans A1 les valeurs 28, 31, puis 36. Dessinez l'arbre A2 après insertion de ces valeurs.

Réponse :

Solution: Insertion de 28 et 31 dans F3 : F3 (27,28, 30, 31). Pas d'autre modification
Insertion de 36 dans F4. Plus de place. On éclate F4 en F4a (36, 38, 41) et F4b (42, 49).
On remonte 42 dans la racine. Plus de place, il faut créer un niveau supplémentaire. On éclate la racine en deux noeuds N1 (18, 26) et N2 (42, 55) et on crée une racine avec une seule valeur, 34, qui est la valeur médiane.

Résultat : R(34), N1 (18, 26) , N2 (42, 55)

F1 (2, 9, 13, 16), F2 (18, 23, 24), F3 (27, 28, 30, 31), F4a (36, 38, 41), F4b (42, 49), F5 (55, 58)

Question 3 (1 point)

Dans l'arbre A1 d'origine, on supprime successivement les valeurs 55, 38, puis 42. Dessinez l'arbre A3 après suppression de ces valeurs. En cas de fusion nécessaire, on considérera d'abord la fusion avec le voisin de gauche, puis si impossible, avec le voisin de droite.

Réponse :

Solution:

Suppression de 55, F5 n'est plus assez pleine, on redistribue à gauche : F4 devient (38, 41, 42), F5 contient (49, 58). On remonte 49 à la place de 55 dans la racine.

On a donc R (18, 26, 34, 49) F1 (2, 9, 13, 16), F2 (18, 23, 24), F3 (27, 30), F4 (38, 41, 42), F5 (49, 58).

On supprime 38, F4 devient F4 (41, 42). Le reste ne change pas.
Suppression de 42 : F4 pas assez remplie, on fusionne avec F3. On a F3a (27, 30, 41). Il faut supprimer une entrée dans la racine, qui est 34. La racine devient R (18, 26, 49).

2 Optimisation de requêtes (4 pts)

Un concert se déroule dans une ville un certain jour. Les personnes sont identifiées par leur numéro. Elles achètent des billets de concert à un prix donné.

Concert (numC, ville, jour)

Personne (numP, nom, prénom, âge, profil)

Billet (numC*, numP*, prix)

Il y a 400 villes différentes : $v_1, v_2, ..., v_{400}$.
Les nombres sont entiers : l'âge est dans $[20, 100[$ et le prix est dans $[2, 202[$

Question 4 (1 point)
Soit la requête R1 :

```
select p.prénom, c.jour, b.prix
from Concert c, Personne p, Billet b,
where p.numP = b.numP and b.numC = c.numC
and b.prix>=50 and c.ville= 'Aix'
```

P1 est le plan d'exécution de R1 pour lequel on traite les sélections puis les projections le plus tôt possible. Dessiner l'arbre de P1. Inscrire les feuilles de l'arbre sur les traits pointillés en bas du dessin, la racine est en haut.

Réponse :

....

Solution:
jointure de Billet et Concert AVANT la jointure avec Personne

$\pi_{prenom,jour,prix}(\pi_{jour,prix,numP}((\pi_{numC,jour}\sigma_{ville='Aix'}Concert)$
$\bowtie (\pi_{numC,numP,prix}\sigma_{prix\geq 50}Billet)) \bowtie \pi_{numC,prenom}Personne)$

Question 5 (1 point)

Pour chaque prédicat de sélection s1 à s3, quel est son facteur de sélectivité ?

s1 : *age >= 92*

s2 : *prix >=15 and prix < 25*

s3 : *ville = 'Aix' or ville = 'Nice'*

Réponse :

s1 facteur de sélectivité=

s2 facteur de sélectivité=

s3 facteur de sélectivité=

Solution:

$age >= 92$: SF = (100 - 92) / (100 - 20) = 8 / 80 = 10%

$prix >= 15\ and\ prix < 25$: SF = (25 - 15) / (202-2) = 10/200 = 5%

$ville =' Aix'\ or\ ville =' Nice'$: SF = (1+1) / 400 = 0,5%

Question 6 (1 point)

Coût d'une sélection. Les données sont stockées sans ordre particulier. Le coût pour lire une page d'une relation vaut 1. L'accès à <u>un</u> nuplet indexé se fait en <u>une</u> lecture de page .

Il y a 3000 billets stockées dans 300 pages. L'attribut prix est indexé.

Il y a 400 personnes stockées dans 100 pages. L'attribut âge est indexé.

On considère les requêtes :

R2 : $\sigma_{prix=99} Billet$

R3 : $\sigma_{age<80} Personne$

Pour chaque requête, quel est son coût avec et sans utiliser d'index ?

Réponse :

coût R2 avec index = coût R2 sans index =

coût R3 avec index = coût R3 sans index =

Solution:
Sans index : lire Billet : coût= **300** pages
Avec index : lire $1/(202-2)*3000 = 15$ nuplets donc **15** pages à lire.
Sans index : lire Personne : coût= **100** pages
Avec index : lire $(80-20)/(100-20)*400 = 300$ nuplets donc **300** pages à lire.

Question 7 (1 point)
Les attributs age, $prix$, $Personne.numP$ et $Billet.numP$ sont indexés. On considère les plans :
P1 : $\sigma_{age<80}(\sigma_{prix=99}Billet \bowtie Personne)$
P2 : $\sigma_{prix=99}(\sigma_{age<80}Personne \bowtie Billet)$

Les jointures sont traitées par boucle imbriquées. Penser à utiliser des index lorsque cela permet de diminuer le coût d'un plan. Donner le coût des plans P1 et P2 en précisant les index utilisés.

Réponse :
P1 : index utilisés : $cout_{P1} =$

P2 : index utilisés : $cout_{P2} =$

Solution:
P1 :
Lire les Billets à 99 euros avec index sur le prix : 15 pages à lire
Jointure avec Personne : 15 *1 (chaque billet est associé à une personne)
Sélection (age<60) : 0
coût total = 30
P2 :
Lire les Personne de moins de 80 ans sans index : 100 pages à lire
On obtient 300 personnes
Jointure avec Billet : $300 * 7,5 = 2250$(car on a en moyenne 7,5 billets par personne)
Sélection 0
coût total = $100 + 2250 = 2350$

3 Transactions et concurrence (4 pts)

Soient T_1, T_2, T_3, T_4 et T_5 cinq transactions et x, y, z et t quatre granules d'une base de données. On note :

— $L_i(g)$ la <u>lecture</u> de la transaction T_i du granule g
— $E_i(g)$ l'<u>écriture</u> de la transaction T_i du granule g
— V_i l'opération de validation de la transaction T_i

Question 8 (1 point)

Soit S_1 la séquence d'opérations donnée comme suit

$$S_1 = L_2(t), L_4(y), L_4(z), E_3(t), E_4(z), L_3(z), E_3(z), L_1(x), E_1(y), E_1(x), L_1(y), L_2(x), V_1, V_2, V_3, V_4$$

On suppose que les opérations sont exécutées dans l'ordre indiqué.

Préciser pour chaque granule la séquence d'opérations qui le concerne ainsi que les arcs de précédence $T_i \longrightarrow T_j$.

Réponse :

granule	LA séquence	LES arcs
x	$L_1 \ldots$	$T_{\ldots} \longrightarrow T_{\ldots}$
y		
z		
t		

Solution: Pour x : L1 E1 L2 . $T_1 \longrightarrow T_2$
Pour y : L4 E1 L1 . $T_4 \longrightarrow T_1$
Pour z : L4 E4 L3 E3. $T_4 \longrightarrow T_3$
Pour t : L2 E3 . $T_2 \longrightarrow T_3$
graphe **a**cyclique. donc sérialisable pour les conflits.

Cette séquence est-elle sérialisable ? **Justifier.**

Réponse : Entourer la bonne réponse

S_2 est SERIALISABLE NON SERIALISABLE

Justification

Question 9 (1 point)

Soit S_2 la séquence d'opérations donnée comme suit

$$S_2 = L_3(t), L_5(t), L_4(z), E_4(t), E_4(z), L_2(z), E_2(z), L_1(z), L_1(y), E_1(x), E_5(x), L_4(x), V_1, V_2, V_3, V_4, V_5$$

On suppose que les opérations sont exécutées dans l'ordre indiqué.
Préciser pour chaque granule la séquence d'opérations qui le concerne ainsi que les arcs de précédence $T_i \longrightarrow T_j$.

Réponse :

granule	LA séquence	LES arcs
x	$E_1 \dots$	$T_{\dots} \longrightarrow T_{\dots}$
y		
z		
t		

Solution:
Pour x : E1 E5 L4 . $T_1 \longrightarrow T_5 \longrightarrow T_4$
Pour y : L1 . aucun
Pour z : L4 E4 L2 E2 L1. $T_4 \longrightarrow T_2 \longrightarrow T_1$
Pour t : L3 L5 E4 . $(T_3, T_5) \longrightarrow T_4$
graphe cyclique. donc non sérialisable pour les conflits.

Cette séquence est-elle sérialisable ? **Justifier.**

Réponse : Entourer la bonne réponse

S_1 est SERIALISABLE NON SERIALISABLE

Justification

Pour les questions suivantes, on ne considère que trois transactions T_1, T_2 et T_3 travaillant sur trois granules x, y, z

Question 10 (1 point)
Soit S_3 la séquence d'opérations donnée comme suit

$$S_3 = L_1(x), L_2(y), L_2(z), E_3(y), L_3(y), V3, E_1(z), V1, E_2(x), V2$$

On voudrait appliquer le protocole de verrouillage en deux phases strict (2PL strict). Indiquer la séquence d'actions obtenue en sortie lorsqu'il n y pas d'interblocage. S'il y a interblocage, dessiner le graphe des attentes à la place.

Réponse : Entourer la bonne réponse

INTERBLOCAGE PAS D'INTERBLOCAGE

Ordre en sortie <u>ou</u> Graphe des attentes

Solution: interblocage car Graphe d'attente comporte un cycle.

— $L1(x), L2(y), L2(z), E3(y)$ puis T3 mise en attente de T2 sur y, donc $T_3 \longrightarrow T_2$.
— $VX1(z)$ pour faire $E1(z)$, T1 mise en attente de T2 sur z, donc $T_1 \longrightarrow T_2$.
— T2, la seule qui s'execute demande $VX_2(x)$ et rentre en attente de T_1, donc $T_2 \longrightarrow T_1$.

Question 11 (1 point)

Soit S_3 la séquence d'opérations donnée comme suit

$$S_4 = L_2(x), L_3(x), E_2(z), L_1(y), L_1(z), V2, E_3(x), E_1(z), V1, E_3(y), V3$$

On voudrait appliquer le protocole de verrouillage en deux phases strict (2PL strict). Indiquer la séquence d'actions obtenue en sortie lorsqu'il n y pas d'interblocage. S'il y a interblocage, dessiner le graphe des attentes à la place.

Réponse : Entourer la bonne réponse

INTERBLOCAGE PAS D'INTERBLOCAGE

Ordre en sortie <u>ou</u> Graphe des attentes

Solution: pas d'interblocage. $L2(x), L3(x), E2(z), L1(y), V2, L1(z), E3(x), E1(z), V1, E3(y), V3$

4 Algèbre relationnelle (4 pts)

On considère le schéma relationnel suivant qui permet de stocker des données géographiques :

PAYS (idP, NomP, Superficie, Population, idC^*)
CONTINENT (idC, NomC, Superficie)
FRONTIERES ($\underline{idP^*, idPF^*}$, longueur)
LANGUE (CodeL, NomL)
LANGUESPAYS ($\underline{idP^*, CodeL^*}$, pourcentage)
MONTAGNE (idM, NomM, Altitude)
MONTAGNESPAYS ($\underline{idP^*, idM^*}$)

Les attributs soulignés représentent les clés primaires, les attributs en italique et avec astérisque représentent les clés étrangères. Un pays a un identifiant, un nom, on connaît sa population et sa superficie (en km^2) et le continent auquel il appartient. Un continent a un identifiant, on connaît son nom et sa superficie. La table **FRONTIERES** stocke les couples de pays frontaliers, avec la longueur de chaque frontière. La relation **FRONTIERES** n'est pas symétrique, chaque couple de pays est stocké une seule fois, *i.e* si (IdPays1, IdPays2, long) existe dans FRONTIERES alors (idPays2, idPays1, long) n'y est pas. Pour chaque langue on connaît son code (identifiant) et son nom. Les langues parlées dans chaque pays sont enregistrées dans la table **LANGUESPAYS**, pour chaque langue on connaît le pourcentage de la population du pays qui la parle. Pour chaque montagne on connaît son identifiant, son nom et son altitude. Une montagne peut se trouver dans plusieurs pays (table **MONTAGNESPAYS**).

Question 12 (1 point)
Les noms des pays ayant au moins deux montagnes.

Réponse :

Solution:
$R_1 = \pi_{NomP}(Pays \bowtie (MontagnePay_{mp1} \bowtie_{mp1.idM=mp2.idM \wedge mp1.idP \neq mp2.idp} MontagnePays_{mp2})))$

Question 13 (1 point)
Les noms des montagnes qui se trouvent dans un seul pays.

Réponse :

Solution: $R_1 = \sigma_{idP_1 \neq idP_2}(\rho_{idP \to idP_1}(MontagnesPays) \bowtie \rho_{idP \to idP_2}(MontagnesPays))$
$\pi_{NomM}(Montagne \bowtie (\pi_{idM}(MontagnesPays) - \pi_{idM}(R_1)))$

Question 14 (1 point)

Les noms des pays frontaliers de la France, avec la longueur de la frontière (rappel : la relation **FRONTIERES** n'est pas symétrique).

Réponse :

Solution: $R_1 = \pi_{NomP,longueur}(\pi_{idPF,longueur}(\sigma_{NomP='France'}(Pays) \bowtie Frontieres) \bowtie_{idPF=idP} (Pays))$
$R_2 = \pi_{NomP,longueur}(\pi_{idP,longueur}(\rho_{idP \to idPF}(\sigma_{NomP='France'}(Pays)) \bowtie Frontieres) \bowtie (Pays))$
$R = R_1 \cup R_2$

Question 15 (1 point)

Les noms des langues qui sont parlées par au moins 10 % de la population dans tous les pays dont la population est supérieure à 10 millions.

Réponse :

Solution:
$R_1 = (\sigma_{pourcentage > 10'} LanguesPays) \div \pi_{idP}(\sigma_{population > 10000000} Pays)$
$R_2 = \pi_{NomL}(R_1 \bowtie Langue)$

5 Optimisation de schéma (5 pts)

Soit donné une table **R(A,B,C,D,E,F)** et deux ensemble de dépendances fonctionnelles $\mathcal{F}$ et $\mathcal{G}$:

$$\mathcal{F} = \{AD \to C; A \to B; C \to AD; D \to A; E \to A\} \quad (1)$$
$$\mathcal{G} = \{A \to B; C \to D; D \to C; D \to A; E \to A\} \quad (2)$$
$$(3)$$

$\mathcal{G}$ **est un ensemble minimal.**

Question 16 (1 point)

Montrez que $\mathcal{G}$ est équivalent à $\mathcal{F}$.

Réponse :

Solution: On montre que $\mathcal{G}^+ = \mathcal{F}^+$:

Forme canonique de $\mathcal{F}$:

$\mathcal{F} = \{AD \to C, A \to B, C \to A, C \to D, D \to A, E \to A\}$

Forme canonique de $\mathcal{G}$:

$\mathcal{G} = \{A \to B; C \to D; D \to C; D \to A; E \to A\}$

$\mathcal{F} - \mathcal{G} = \{AD \to C, C \to A\}$: on montre que $C \in [AD]+_{\mathcal{G}} = ADCB$ et $A \in [C]+_{\mathcal{G}} = CDAB$

$\mathcal{G} - \mathcal{F} = \{D \to C\}$: on montre que $C \in [D]+_{\mathcal{F}} = DABC$

Question 17 ($^1\!/_2$ point)

Donnez toutes les clés de la table R avec $\mathcal{G}$. Justifiez votre réponse.

Réponse :

Solution:

E et F font partie de toute clé car jamais à droite de DF.

$[EF]^+ = EFAB$: A et B ne font pas partie d'une clé)

On ajoute C : $[CEF]^+ = CEFABD$

On ajoute D : $[DEF]^+ = DEFABC$

Les clés sont les : [CEF] [DEF]

Question 18 ($^1\!/_2$ point)

Est-ce que la table **R** est en 3e forme normale (3FN) ? Justifiez formellement votre réponse.

Réponse :

Solution:
La relation R(A,B,C,D,E) n'est pas en 3FN.
Par exemple, dans la DF $A \rightarrow B$, A n'est pas surclé et B n'est pas prime.

Question 19 (2 points)
Est-ce que la décomposition de **R** en **R1(A,B,C), R2(C,D), R3(D,E,F)** avec $\mathcal{G}$ est sans perte d'information (SPI) ? Justifiez votre réponse en utilisant le méthode du tableau et en indiquant pour chaque DF utilisée le tableau intermédiaire généré.

Réponse :

Solution:
Décomposition : R1(A,B,C) R2(C,D) R3(D,E,F)

	A	B	C	D	E	F
R1	A	B	C	d0	e0	f0
R2	a1	b1	C	D	e1	f1
R3	a2	b2	c2	D	E	F

- $C \rightarrow AD$: a1=A, d0=D

	A	B	C	D	E	F
R1	A	B	C	D	e0	f0
R2	A	b1	C	D	e1	f1
R3	a2	b2	c2	D	E	F

- $D \rightarrow A$: a2=A

	A	B	C	D	E	F
R1	A	B	C	D	e0	f0
R2	A	b1	C	D	e1	f1
R3	A	b2	c2	D	E	F

- $AD \rightarrow C$: c2=C

	A	B	C	D	E	F
R1	A	B	C	D	e0	f0
R2	A	b1	C	D	e1	f1

- $A \rightarrow B$: b1=B, b2=B

	A	B	C	D	E	F
R1	A	B	C	D	e0	f0
R2	A	B	C	D	e1	f1

Il y a une ligne complètement définie : la décomposition est SPI

Question 20 (1 point)

Donnez une décomposition SPI et SPD de R avec $\mathcal{G}$ (ensemble minimal) en tables qui sont en 3FN. **Donnez pour chaque table les dépendances fonctionnelles associées et les clés.**

Réponse :

Solution:

— R1(**A**,B), $A \rightarrow B$, clés : A

— R2(**D**,A,C), $D \rightarrow AC, C \rightarrow D$; clés : D et C

— R3(**E**,A), $E \rightarrow A$; clé : E

— R4(**C,E, F**), clé : CEF (table pour la clé CEF ; choix de l'autre clé DEF possible)

Question	Points	Score
1	1	
2	1	
3	1	
4	1	
5	1	
6	1	
7	1	
8	1	
9	1	
10	1	
11	1	
12	1	
13	1	
14	1	
15	1	
16	1	
17	$^1/_2$	
18	$^1/_2$	
19	2	
20	1	
Total:	20	