

--	--	--

LI341 Bases de données – CORRIGÉ

EXAMEN - 1ère session du 7 janvier 2015 – 2 heures – documents autorisés

Les téléphones mobiles doivent être éteints et rangés dans les sacs. Le barème sur 60 points (18 questions) n'a qu'une valeur indicative.

1 Requêtes (13 pts)

Considérez le schéma relationnel *BDPedia* pour stocker des extraits (incomplets) de Wikipedia dans les six tables suivantes :

LANGUE (CodeL, Langue)

ARTICLE (Aid, CodeL, Titre, Contenu)

CATEGORIE (Cid, CodeL, NomC)

REFERENCE (Aid, Cid)

PERSONNE (Pid, NomP, Prenom, Email)

ECRIT (Pid, Aid, CodeL)

Les attributs de la clé primaire de chaque table sont soulignés. Les attributs ayant le même nom dans les différentes tables portent sur les mêmes entités (articles, catégories, personnes, langues). Les différentes langues utilisées sont stockées dans la table **LANGUE** avec un code à deux lettres (par ex. 'fr') et le nom de la langue (par ex. 'français'). Chaque article (table **ARTICLE**) a un identifiant unique (Aid) et des contenus et titres dans plusieurs langues. Les articles sont organisés en catégories. Chaque catégorie (table **CATEGORIE**) a un identifiant unique (Cid) et des noms dans plusieurs langues. Un article appartient à une ou plusieurs catégories (table **REFERENCE**) indépendamment de sa langue. Le contenu d'un article dans une langue donnée peut être rédigé par un ou plusieurs auteurs (table **ECRIT**). Les identifiants (Pid), les noms, les prenom et les adresses mail des auteurs sont stockés dans la table **PERSONNE**.

Voici un extrait des trois tables **CATEGORIE**, **ARTICLE** et **REFERENCE** :

CATEGORIE		
<u>Cid</u>	<u>CodeL</u>	NomC
c12	en	Indian art
c12	fr	Art du monde indien
c45	fr	Monuments religieux

ARTICLE			
<u>Aid</u>	<u>CodeL</u>	Titre	Contenu
a23	en	The Taj Mahal	The Taj Mahal...
a23	fr	Le Taj Mahal	Le Taj Mahal...
a23	de	Das Taj Mahal	Das Taj Mahal...

REFERENCE	
<u>Aid</u>	<u>Cid</u>
a23	c12
a23	c45

Question 1 (3 points)

Formuler en **SQL** et en **algèbre relationnelle** la requête qui retourne les noms d'auteurs d'articles en français (codeL='fr') dans la catégorie 'c12'.

ARTICLE (Aid, CodeL, Titre, Contenu)

CATEGORIE (Cid, CodeL, NomC)

REFERENCE (Aid, Cid)

PERSONNE (Pid, NomP, Prenom, Email)

ECRIT (Pid, Aid, CodeL)

LANGUE (CodeL, Langue)

Solution:

SQL

```
select distinct p.NomP
from PERSONNE p, ECRIT e, REFERENCE r, ARTICLE a
where p.Pid = e.Pid
and e.Aid = a.Aid
and e.CodeL = a.CodeL
and r.Aid = a.Aid
and r.Cid = 'c12'
and a.CodeL = 'fr'
```

ou

```

select distinct p.NomP
from PERSONNE p, ECRIT e, REFERENCE r
where p.Pid = e.Pid
      and e.Aid = r.Aid
      and e.CodeL = 'fr'
      and r.Cid = 'c12'

```

Algèbre

$$Q1 = \pi_{NomP}(PERSONNE \bowtie ECRIT \bowtie \sigma_{codeL='fr'} ARTICLE \bowtie \sigma_{Cid='c12'} REFERENCE)$$

ou

$$Q1 = \pi_{NomP}(PERSONNE \bowtie \sigma_{codeL='fr'} ECRIT \bowtie \sigma_{Cid='c12'} REFERENCE)$$

Question 2 (4 points)

Formuler en **calcul relationnel** et en **SQL** la requête qui retourne *le titre et la langue des articles qui n'ont qu'un seul auteur*.

Solution:

CALCUL

$$\{a.Titre, a.codeL \mid ARTICLE(a) \wedge \exists p(ECRIT(e) \wedge e.Aid = a.Aid \wedge e.codeL = a.codeL \\ \wedge \neg \exists p2(ECRIT(p2) \wedge e.Aid = p2.Aid \wedge e.codeL = p2.codeL \wedge e.Pid \neq p2.Pid))\}$$

ou

$$\{a.Titre, l.Langue \mid ARTICLE(a) \wedge LANGUE(l) \wedge a.codeL = l.codeL \wedge \exists p(ECRIT(e) \\ \wedge e.Aid = a.Aid \wedge e.codeL = a.codeL \wedge \neg \exists p2(ECRIT(p2) \\ \wedge e.Aid = p2.Aid \wedge e.codeL = p2.codeL \wedge e.Pid \neq p2.Pid))\}$$

SQL

```

select a.titre , a.codeL
from ARTICLE a, ECRIT e
where e.Aid=a.Aid
      and e.codeL=a.codeL
      and not exists (select *
                     from ECRIT e2
                     where e.Aid=e2.Aid
                       and e.codeL=e2.codeL
                       and e.Pid <> e2.Pid)

```

ou

```

select a.titre , l.Langue
from ARTICLE a, ECRIT e, LANGUE l
where e.Aid=a.Aid
      and e.codeL=a.codeL
      and e.codeL=l.codeL
      and not exists (select *
                     from ECRIT e2
                     where e.Aid=e2.Aid
                       and e.codeL=e2.codeL
                       and e.Pid <> e2.Pid)

```

ARTICLE (Aid, CodeL, Titre, Contenu) **CATEGORIE** (Cid, CodeL, NomC) **REFERENCE** (Aid, Cid)
PERSONNE (Pid, NomP, Prenom, Email) **ECRIT** (Pid, Aid, CodeL) **LANGUE** (CodeL, Langue)

Question 3 (3 points)

Formuler en **SQL** la requête qui trouve *les noms des auteurs qui ont rédigé des articles dans au moins deux langues (CodeL) et deux catégories (Cid) différentes.*

Solution:

sémantique : plusieurs articles (Aid) dans deux langues et deux catégories différentes **SQL**

```
select p.NomP
  from PERSONNE p, ECRIT e, REFERENCE r
 where p.Pid = e.Pid and e.Aid = r.Aid
 group by p.Pid, p.NomP
 having count(distinct CodeL) >= 2
        and count(distinct Cid) >= 2;
```

ou

```
select p.NomP
  from PERSONNE p
 where exists (select * from ECRIT e1, ECRIT e2
                where p.Pid = e1.Pid and p.Pid = e2.Pid
                and e1.CodeL <> e2.CodeL)
        and exists (select * from ECRIT e1, ECRIT e2, REFERENCE r1, REFERENCE r2
                where p.Pid = e1.Pid and p.Pid = e2.Pid
                and e1.Aid = r1.Aid and e2.Aid = r2.Aid
                and r1.Cid <> r2.Cid)
```

ou

```
select p.NomP
  from PERSONNE p, ECRIT e1, ECRIT e2, ECRIT e3, ECRIT e4, REFERENCE r1, REFERENCE r2
 where p.Pid = e1.Pid and p.Pid = e2.Pid
        and e1.CodeL <> e2.CodeL
        and p.Pid = e3.Pid and p.Pid = e4.Pid
        and e3.Aid = r1.Aid and e4.Aid = r2.Aid
        and r1.Cid <> r2.Cid
```

sémantique : le même article (Aid) dans deux langues et deux catégories différentes

```
select p.NomP
  from PERSONNE p, ECRIT e, REFERENCE r
 where p.Pid = e.Pid and e.Aid = r.Aid
        and exists (select * from ECRIT e1, REFERENCE r1
                where p.Pid = e1.Pid
                and e.Aid = r1.Aid and
                and e.CodeL <> e1.CodeL
                and r.Cid <> r1.Cid)
```

Question 4 (3 points)

Formuler en **SQL** la requête qui retourne *pour chaque catégorie et langue, le nombre d'articles dans cette langue* (on suppose que pour une langue et une catégorie il y a au moins un article de la même langue). Le résultat a la forme suivante :

CodeL	Cid	Categorie	NbArticles
en	67	American TV Series	1245
fr	67	Série TV Americaines	87
en	78	French Cuisine	145
fr	78	La Cuisine Française	745

Solution:

```
select c.CodeL, c.Cid, c.NomC as Categorie, count(distinct e.Aid) as NbArticles
  from ECRIT e, REFERENCE r, CATEGORIE c
 where e.Aid = r.Aid
       and c.Cid = r.Cid
       and e.codeL = c.codeL
 group by c.Cid, c.CodeL, c.NomC
```

ou (remplacer ECRIT par ARTICLE et enlever distinct dans le count)

```
select c.CodeL, c.Cid, c.NomC as Categorie, count(e.Aid) as NbArticles
  from ARTICLE a, REFERENCE r, CATEGORIE c
 where a.Aid = r.Aid
       and c.Cid = r.Cid
       and a.codeL = c.codeL
 group by c.Cid, c.CodeL, c.NomC;
```

2 Dépendances fonctionnelles et normalisation (20 pts)

On considère un schéma de relation $R(A, B, C, D, E, H)$ et les deux ensembles \mathcal{F} et \mathcal{G} de dépendances fonctionnelles :

$$\mathcal{F} = \{A \rightarrow C; AH \rightarrow D; BC \rightarrow D; AC \rightarrow E; DE \rightarrow A; H \rightarrow B\}$$

$$\mathcal{G} = \{A \rightarrow C; BC \rightarrow D; A \rightarrow E; DE \rightarrow A; H \rightarrow B\}$$

Question 5 (3 points)

On sait que \mathcal{G} est un ensemble minimal. Montrez que \mathcal{G} est une couverture minimale de \mathcal{F} .

Solution:

Solution 1 : calculer l'ensemble minimal de \mathcal{F} . \mathcal{F} est déjà en forme canonique.

Elimination de DF redondantes :

- La DF $AH \rightarrow D$ est redondante : $D \subseteq [D]_{\mathcal{F}-\{AH \rightarrow D\}}^+ = AHCEBD$

Elimination de redondances à gauche :

- La DF $AC \rightarrow E$ est redondante à gauche $\Rightarrow A \rightarrow E$

Résultat : $\mathcal{G} = \{A \rightarrow C, BC \rightarrow D, A \rightarrow E, DE \rightarrow A, H \rightarrow B\}$

Solution 2 : montrer que $\mathcal{F} \equiv \mathcal{G}$:

— $AH \rightarrow D \in \mathcal{F} - \mathcal{G} : D \in [AH]_{\mathcal{G}}^+ = AHCEBD$

— $AC \rightarrow E \in \mathcal{F} - \mathcal{G} : E \in [AC]_{\mathcal{G}}^+ = ACE$

— $A \rightarrow E \in \mathcal{G} - \mathcal{F} : E \in [A]_{\mathcal{F}}^+ = ACE$

Question 6 (5 points)

Donnez les trois clés de la table $R(A, B, C, D, E, H)$ par rapport à \mathcal{F} (ou \mathcal{G}) et montrez comment vous les obtenez (vous n'êtes pas obligés de montrer qu'il n'y a pas d'autres clés).

Solution: On peut utiliser \mathcal{F} ou \mathcal{G} .

Les trois clés sont : $[HA], [HCE], [HDE]$.

Justification :

$$[HA]_{\mathcal{F}}^+ = [HCE]_{\mathcal{F}}^+ = [HDE]_{\mathcal{F}}^+ = ABCDEH \text{ clé}$$

H fait partie de toutes les clés.

Les sous-ensembles (avec H) ne sont pas des clés :

— $[H]_{\mathcal{F}}^+ = HB$

— $[HB]_{\mathcal{F}}^+ = HB$

— $[HC]_{\mathcal{F}}^+ = CHBD$

— $[HD]_{\mathcal{F}}^+ = HDB$

— $[HE]_{\mathcal{F}}^+ = HEB$

$R(A, B, C, D, E)$

$$\mathcal{G} = \{A \rightarrow C; BC \rightarrow D; A \rightarrow E; DE \rightarrow A; H \rightarrow B\}$$

Question 7 (3 points)

Est-ce que le schéma $R(A, B, C, D, E, H)$ est en 3e Forme Normale (3FN) par rapport à \mathcal{G} ? Est-ce qu'il est en Forme Normale Boyce-Codd (FNBC) ? Justifiez vos réponses.

Solution: Le schéma n'est pas en 3FN, dans la DF $H \rightarrow B$, H n'est pas surclé et B n'est pas prime. Comme il n'est pas en 3FN, il n'est pas non plus en FNBC.

Question 8 (6 points)

Est-ce que la décomposition

$$R_1(A, B, C)$$

$$R_2(A, B, E, H)$$

$$R_3(A, D, E)$$

$$R_4(C, D, E)$$

de $R(A, B, C, D, E, H)$ est Sans Perte d'Information (SPI) ? Justifiez votre réponse avec la méthode du tableau en montrant le tableau initial et le tableau final avec les DF utilisées.

Solution:

Tableau initial :

	A	B	C	D	E	H
R1	A	B	C	d0	e0	g0
R2	A	B	c1	d1	E	H
R3	A	b2	c2	D	E	g2
R4	a3	b3	C	D	E	g3

- $A \rightarrow C : c1=c2=C$ - $BC \rightarrow D : d0=d1$ - $A \rightarrow E : e0=E$ - $DE \rightarrow A : a3=A$

Tableau final :

	A	B	C	D	E	H
R1	A	B	C	d1	E	g0
R2	A	B	C	d1	E	H
R3	A	b2	C	D	E	g2
R4	A	b3	C	D	E	g3

Il n'y a pas de ligne complètement définie : la décomposition n'est pas SPI.

R(A,B,C,D,E) $\mathcal{G} = \{A \rightarrow C; BC \rightarrow D; A \rightarrow E; DE \rightarrow A; H \rightarrow B\}$ **Question 9** (3 points)

Utilisez la couverture minimal \mathcal{G} pour donner une décomposition de R qui est SPI, SPD et où chaque relation est en 3FN. Donnez les clés de chaque relation.

Solution: On utilise la couverture minimale $\mathcal{G} = \{A \rightarrow C, BC \rightarrow D, A \rightarrow E, DE \rightarrow A, H \rightarrow B\}$

1. Regrouper les DF qui ont la même partie gauche :

 $\{A \rightarrow CE, BC \rightarrow D, DE \rightarrow A, H \rightarrow B\}$

2. Faire une relation par groupe de DF :

 $R_1(\mathbf{A}, C, E), R_2(\mathbf{B}, \mathbf{C}, D), R_3(\mathbf{D}, \mathbf{E}, A), R_4(\mathbf{H}, B)$

3. Aucune clé de la relation de départ n'est contenue dans l'une des relations obtenues :

Ajouter $R_5(\mathbf{A}, \mathbf{H})$ ou $R_5(\mathbf{D}, \mathbf{E}, \mathbf{H})$ ou $R_5(\mathbf{C}, \mathbf{E}, \mathbf{H})$

4. Si une relation est contenue dans une autre, la supprimer.

Il y en a pas.

3 Optimisation (12 pts)

On considère à nouveau le schéma relationnel *BDPedia*. On suppose qu'il existe 5 langues différentes (attribut *CodeL*) et qu'il y a le même nombre de nuplets pour chaque langue dans la table **ARTICLE** (distribution uniforme). Il y a 400 articles distincts (attribut *Aid*). Tous les articles sont rédigés en exactement deux langues (le même nombre d'articles pour chaque couple distincts de langues). Il y a 100 personnes (auteurs) différentes et chaque personne a un nom différent (*nomP* est unique). La table **ECRIT** contient 5000 tuples et il y a le même nombre d'articles par auteur (distribution uniforme).

ARTICLE (Aid, CodeL, Titre, Contenu) **CATEGORIE** (Cid, CodeL, NomC) **REFERENCE** (Aid, Cid)
PERSONNE (Pid, NomP, Prenom, Email) **ECRIT** (Pid, Aid, CodeL) **LANGUE** (CodeL, Langue)

Justifiez vos réponses aux questions suivantes en précisant le calcul effectué.

Question 10 (1 point)

Soit D le nombre de couples distincts de deux langues différentes. Remarque : un couple ('fr', 'en') ou inversement ('en', 'fr') compte pour un seul couple. Combien vaut D ?

Solution: $D = 10 = 5 \text{ langues} * 4 \text{ langues} / 2 = 10 \text{ couples de 2 langues}$

Question 11 (1 point)

Combien d'articles (ayant des valeurs *Aid* distinctes) sont rédigés à la fois en 'fr' et en 'en' ? Répondre en fonction de D .

Solution: Articles en français et en anglais : $A = 40 = (400 \text{ articles} / D)$

Question 12 (3 points)

- Quelle est la cardinalité de $\pi_{Aid}(ARTICLE)$?
- Combien y a-t-il de nuplets dans la table **ARTICLE** ($card(ARTICLE)$) ?
- Quelle est la cardinalité de $\pi_{Aid}(\sigma_{codeL='fr'}(ARTICLE))$?

Solution:

$$card(\pi_{Aid}(Article)) = 400$$

$$card(ARTICLE) = 800 = 400 \text{ articles} * 2 \text{ langues par article}$$

$$card(\pi_{Aid}(\sigma_{codeL='fr'}(ARTICLE))) = 160 = 400 * 4/10 \text{ ou } 800 * 1/5$$

ARTICLE (Aid, CodeL, Titre, Contenu) **CATEGORIE** (Cid, CodeL, NomC) **REFERENCE** (Aid, Cid)
PERSONNE (Pid, NomP, Prenom, Email) **ECRIT** (Pid, Aid, CodeL) **LANGUE** (CodeL, Langue)

Question 13 (3 points)

Donnez la cardinalité des requêtes suivantes :

- $Q_1 = ECRIT \bowtie (\sigma_{NomP='Dupont'}(PERSONNE))$
- $Q_2 = (\sigma_{codeL='fr'}(ECRIT)) \bowtie PERSONNE$
- $Q_3 = \sigma_{codeL='fr'}(ECRIT) \bowtie \sigma_{NomP='Dupont'}(PERSONNE)$

Solution:

$$card(Q_1) = 50 = 5000 \text{ redactions} / 100 \text{ auteurs}$$

$$card(Q_2) = 1000 = 5000 \text{ redactions} / 5 \text{ langues}$$

$$card(Q_3) = 10 = (5000 \text{ redactions} / 5 \text{ langue}) / 100 \text{ auteurs}$$

Question 14 (1 point)

On considère les deux requêtes :

- $Q_4 = \pi_{Aid, codeL}(\sigma_{codeL='fr' \vee NomP='Dupont'}(ECRIT \bowtie PERSONNE))$
 - $Q_5 = \pi_{Aid, codeL}(\sigma_{codeL='fr'}(ECRIT) \bowtie \sigma_{NomP='Dupont'}(PERSONNE))$
- Est ce que Q_4 est équivalente à Q_5 ?

Solution: non, pas équivalente

Question 15 (3 points)

Transformez Q_4 pour traiter les sélections le plus tôt possible.

Solution:

$\pi_{Aid, codeL}(\sigma_{codeL='fr'}(ECRIT) \bowtie PERSONNE)$
 $\cup \pi_{Aid, codeL}(ECRIT \bowtie \sigma_{NomP='Dupont'}(PERSONNE))$

4 Transactions (15 pts)

Les deux parties suivantes sont indépendantes.

Partie 1 : Soit une base de données contenant la relation $R(A,B,C)$ avec les trois tuples suivants :

R		
A	B	C
x	y	m
z	t	n
u	v	r

Le gestionnaire de concurrence utilise le verrouillage à deux phases strict. Le verrouillage se fait à la granularité «attribut de n-uplet», ce qui définit 9 *granules* : x, y, z, t, u, v, m, n et r qui sont verrouillées (quand nécessaire) indépendamment des autres granules à chaque opération de lecture et d'écriture dans la base. Les valeurs des granules sont supposées toutes différentes les unes des autres.

Question 16 (4 points)

Quels sont les *granules* verrouillés en lecture (VP) et en écriture (VX) pour les requêtes suivantes :

select A from R where B=y ;

Solution: VP(x,y, t,v)

update R set B=var where A=z ;

Solution: VP(x,z,u) VX(t)

select r1.C from R r1 where not exists (select * from R r2 where r1.A=r2.B) ;

Solution: VP(tous les granules)

select r1.C from R r1 where exists (select * from R r2 where r1.A=r2.B) ;

Solution: VP(x,y,z,t,uv) pas m,n, r car le exists renvoie toujours faux

Partie 2 : Soient T_1, T_2, T_3, T_4 et T_5 cinq transactions et x, y, z, t quatre granules d'une base de données. On note respectivement $L_i(g)$ et $E_i(g)$ la lecture et l'écriture par la transaction T_i du granule g . Soit la séquence d'opérations (exécution) S suivante :

$S = L_4(z), L_1(x), L_3(y), L_4(y), E_1(x), E_3(y), L_4(t), L_3(z), L_4(z), \mathbf{E_4(t)}, \mathbf{L_1(t)}, E_2(t), \mathbf{E_2(x)}, L_5(y), E_5(z), \mathbf{L_3(x)}, \mathbf{E_5(t)}$

On suppose qu'une transaction se termine dès que la dernière opération de cette transaction est effectuée (ces opérations sont notées en gras).

Question 17 (5 points)

En supposant que les opérations sont exécutées dans l'ordre d'arrivée (celui de la séquence S), dire si l'exécution correspondant à la séquence S est sérialisable ou non, en justifiant votre réponse. Si l'exécution est sérialisable, donner un ordre en série équivalent.

Solution: Graphe de précédence :

$z : (T_4, T_3) \rightarrow T_5$

$x : T_1 \rightarrow T_2 \rightarrow T_3$

$y : T_4 \rightarrow T_3 \rightarrow T_5$

$t : T_4 \rightarrow T_1 \rightarrow T_2 \rightarrow T_5$

Pas de circuit, c'est donc sérialisable T_4, T_1, T_2, T_3, T_5

Question 18 (6 points)

On suppose que les opérations arrivent dans l'ordre d'arrivée (de la séquence S) dans un gestionnaire de concurrence contrôlé par verrouillage en deux phases strict.

- Montrez le *nouvel ordre d'exécution* produit jusqu'à la fin de la séquence S ou jusqu'à l'apparition d'un interblocage.
- Pendant l'exécution, une *table de verrous* contient pour chaque granule les transactions T_i avec un verrou en lecture (VP) ou un verrou en écriture (VX) en cours. Représenter la table des verrous lorsque la première transaction qui termine a exécuté sa dernière opération *mais n'a pas encore validé* ou, si un interblocage apparaît avant, au moment de l'interblocage.

Solution:

Ordre obtenu :

$L_4(z), L_1(x), L_3(y), L_4(y), E_1(x), L_4(t), L_4(z), E_4(t), E_3(y), L_3(y), L_3(z), L_1(t), E_2(t), E_2(x), L_3(x), L_5(y), E_5(z), E_5(t)$

Pas d'interblocage

Etat de la table des verrous au moment de $\mathbf{E_4(t)}$:

Granule	x	y	z	t
VX	T1			T4
VP		T3, T4	T4	
Attente		T3X		