Université d' Ottawa
CSI 3530 / CSI 3130 – Examen Final
Professeur: Herna L. Viktor et Iluju Kiringa

11 décembre, 2007
14:h00-17h00
Durée: 3 hrs

A livre fermé; aucune aide permise, sauf une feuille “style légal” écrite au recto et verso.
Tout cas de tricherie sera traité avec sévérité.

Nom de famille: _____

Prénom: _____

Numéro d' étudiant: _____

Il y a 8 questions et un total de 100 points.

Cet examen doit contenir 14 pages,
incluant cette page couverture.

1 – Evaluation des Opérateurs Relationnels	/ 14
2 – Tri Externe	/ 10
3 – Optimisation des Requêtes	/ 10
4 – Acces Simultané	/ 15
5 – Reprise	/ 14
6 – Bases de Données Distributées	/ 13
7 – Entreposage de Données	/ 14
8 – Exploration des Données	/ 10
Total	/ 100

1 Evaluation des Opérateurs Relationnels — 14 points

A. (2 points) Qu'est ce qu'un *plan incliné vers la gauche*?

B. (2 points) Qu'est ce qu'un *index a sacannage seulement ("index only scan")*?

C. (2 points) Considerez le schéma relationnel suivant pour la relation Employee:

$$Employee(\underline{eid : \texttt{int}}, ename : \texttt{string}, address : \texttt{string}, age : \texttt{real})$$

Pour chacun des indexes suivants, dites si cet index correspond a la condition de sélection donnée. Donnez les conjoints primaires de chaque correspondance. I.e. repondez par **oui** ou **non** et si oui donnez les conjoints primaires.

(1) Un arbre B+ avec clé de recherche $\langle Employee.eid, Employee.age \rangle$

(a) $\sigma_{Employee.age=20 \wedge Employee.eid<40}(Employee)$

(2) Un index a hachage avec clé de recherche $\langle Employee.eid, Employee.age \rangle$

(b) $\sigma_{Employee.eid=20{,}000}(Employee)$

... suite

D. (8 points) Repondez a **SEULEMENT DEUX** des cinq questions suivantes:

- (4 points) Expliquez/donnez l' algorithme "**page** nested loop join".
- (4 points) Expliquez/donnez l' algorithme "**sort-merge** join".
- (4 points) Expliquez/donnez l' algorithme "double buffering".
- (4 points) Expliquez/donnez l' algorithme de l' opérateur de l' union basé sur le tri.
- (4 points)Expliquez/donnez l' algorithme des opérateurs d' aggrégat avec GROUP BY basé sur le hachage.

... suite

2 Tri Externe — 10 points

Supposez un fichier avec 9 pages ainsi que un certain nombre de pages tampons mis a votre disposition. Les 9 pages du fichier sur disque contiennent les données numériques suivantes:

5,6	8, 4	11, 6	10, 9	7, 8	5, 3	4, 5	7,3	4

Utilisez l' algorithme général de tri externe pour répondre aux questions suivantes:

1 (6 points) Illustrez l' algorithme de tri externe a 3 voies pour trier ce fichier de 9 pages.

2 (4 points) Combien de passages faudra-t-il pour trier le fichier au complet?

... suite

3 Optimisation des Requêtes — 10 points

Partie A —2 points Qu'est ce que un *facteur de r éduction* d'une condition de sélection?

Partie B —8 points Considérez le schéma relationnel suivant:

$$Player(\underline{pid} : \texttt{int}, pname : \texttt{string}, age : \texttt{real}, nationality : \texttt{string})$$
$$PlaysIn(\underline{pid} : \texttt{int}, tid : \texttt{integer}, years : \texttt{real})$$
$$Team(\underline{tid : \texttt{int}}, tname : \texttt{string}, owner : \texttt{string})$$

Supposez la requête SQL suivante:

```
SELECT P.pname, T.tname
FROM Player P, Team T, PlaysIn I
WHERE P.pid = I.pid AND T.tid = I.tid
            AND P.nationality = 'Greek'
            AND T.owner ='tiller'
```

Supposez que nous ayions les indexes suivants: un index B+ sur la colonne *pid* de la relation *Player*, un index B+ sur la colonne *pid* de la relation *PlaysIn*, et un index a hachage sur la colonne *owner* de la relation *Team*.

Donnez un plan d'évaluation de la requête ci-dessus et motivez votre choix de la précédence des opérateurs.

... suite

4 Acces Simultané — 15 points

Partie A —3 points Pourquoi le protocole “two-phase locking” est appelé “two-phase”?

Partie B —8 points Considerez le plan suivant impliquant trois transactions T_1, T_2 et T_3:

$$R_1(X), R_2(Z), R_1(Z), R_3(X), R_3(Y), W_1(X), C_1, W_3(Y), C_3, R_2(Y), W_2(Z), W_2(Y), C_2$$

A. (4 points) Dessinez le graphe de sérialisabilité de ce plan.

B. (2 points) Déterminez si ce plan est érialisable par rapport aux conflits.

C. (2 points) Déterminez si ce plan est recouvrable. Motivez votre réponse.

Partie C —4 points Considerez le plan suivant qui, en plus des actions “read” et “write”, contient des actions de verrouillage S(X) et X(A).

$$\mathtt{S_1(U)}, \mathtt{R_1(U)}, \mathtt{X_2(Y)}, \mathtt{W_2(Y)}, \mathtt{S_1(Y)}, \mathtt{S_3(Z)}, \mathtt{R_3(Z)}, \mathtt{X_2(Z)}, \mathtt{X_4(Y)}, \mathtt{X_3(Y)}$$

Dessinez le graphe “Wait-for” pour ce plan et indiquez un éventuel deadlock.

... suite

5 Reprise — 14 points

Partie A —6 points

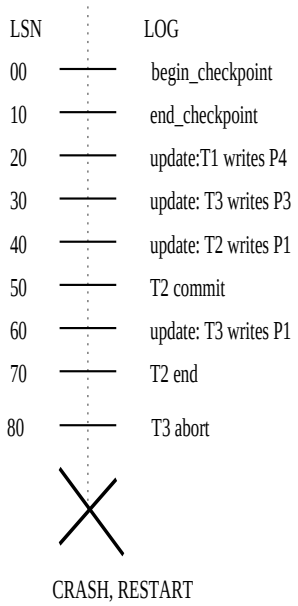
A. (2 points) Décrivez l' approche ' “steal/no-force” dans la reprise.

B. (4 points) Décrivez les quatre propriétés des transactions que un SGBD doit garrantir afin de gérer les données en présence de l' acces simultané et des pannes du systeme.

... suite

Partie B —8 points

Considérez l'exécution illustrée dans la figure ci-dessous:



Supposez que le systeme se plante pendant la reprise apres avoir écrit deux enregistrements du log vers le disque et tombe en panne a nouveau apres avoir écrit deux autres enregistrements.

Répondez aux questions suivantes:

A. (3 points) Montrez quels pas sont entrepris durant la phase de l'analyse.

B. (3 points) Montrez quels pas sont entrepris durant la phase REDO.

C. (2 points) Montrez quels pas sont entrepris durant la phase UNDO.

... suite

6 Bases de Données Distributées — 13 points

Partie A —4 points Expliquez la différence entre les reproductions synchrone et asynchrone.

Partie B —9 points

Considérez un systeme de bases de données distribuées contenant deux tables, a savoir Customer (qui contient de l'info sur les clients qui louent des maisons de vacances dans les Caraibes) ainsi que RentalHome (qui contient de l'info sur les maisons louées). L'information au sujet de tous les clients (i.e. la table Customer) est stockée a Ottawa et toute l'information sur les maisons louées (i.e. la table RentalHome) est stockée a New Providence aux Bahamas. La base de données a le schéma suivant:

- Customer(Cid: integer, Rentid: integer, Income: real)
- RentalHome(Rentid: integer, OwnerId: integer, RentAmount: integer)

La relation Customer contient 100,000 pages et la relation RentalHome en contient 5,000. Il n'y a aucun index de join et un join a tri-fusion (“sort-merge join”) est utilisé locallement sur les sites.

Veuillez répondre aux questions suivantes:

A. (2 points) Expliquez la notion de *interblockage phantome* (“*phantom deadlock*”). Donnez un exemple d'une situation ou ce phénomene apparait dans la base de données ci-dessus.

... suite

B. (3 points) Considérez une requête visant a sélectionner tous les détails des clients qui sont aussi des propriétaires de maisons (I.e. Customer.Cid = RentalHome.OwnerId). Cette requête est posée a Paris en France. En outre, supposez que 1 pourcent des clients sont des propriétaires. Quel plan d' évaluation de la requête utiliseriez vous afin de minimaliser les coûts de transports? Expliquez votre réponse.

C. (4 points) Supposez que tous les tuples de la relation Customer sont toujours stockés a Ottawa, mais que les tuples de Customer dont le revenu est de moins de 100,000 sont reproduits a Paris (I.e. la base de données est maintenant distribuée sur trois sites.) Les verrous sont gérés au *site primaire*, i.e. a Ottawa. Expliquez quels verrous sont enclenchés (et a quels site ils le sont) afin de traiter une requête posée a New York pour lire une page de tuples de Customer dont le revenu est de moins de 50,000.

... suite

7 Entreposage de Données — 14 points

Partie A —2 points Expliquez ce qu'est la matérialisation des vues.

Partie B —12 points Le Directeur de SITE garde et maintient de l'information sur l'utilisation des laboratoires. Cette utilisation est mesurée en terme de nombre d'étudiants qui utilisent les laboratoires. Cette mesure est importante pour la budgétisation. Vous êtes appelé a dévelloper un entrepôt de données pour le maintient des statistiques d'usage des laboratoires. Les principaux prérequis sont:

- Montrer le nombre total d'utilisateurs par différentes périodes de temps.
- Montrer l'utilisation par période de temps, par grade poursuivi et par classsification (sous-gradué ou gradué).
- Comparer l'utilisation pour des grades et des semestres différents.

Pour chaque étudiant, il faut stocker de l'information incluant le numéro d'étudiant, le mot de passe, le nom, l'adresse, le email, les cours aux quels il est enregistré, le majeur, le grade et l'année d'études. Chaque fois que un étudiant ouvre une session ou la termine, son numéro d'étudiant ainsi que le temps de l'action y associée sont capturés dans un log. Pour chaque cours, il faut stocker de l'information incluant le professeur, le début et la fin du semestre, le sujet et les exigences en terme de logiciels.

... suite

A. (6 points) Dessinez le schéma en étoile (“star schema” – modele multidimensionel) de l'utilisation des laboratoires.

B. (3 points) Donnez un exemple d'une hiérarchie d'attibuts (dimensions) qui sera utilisée dans cet entrepôt de données.

C. (3 points) Donnez un exemple d'une requête OLAP qui pourrait être posée a cet entrepôt de données.

... suite

8 Exploration des Données — 10 points

Partie A —2 points Expliquez ce que l'analyse du panier de la ménagere ("**market basket analysis**") est et donnez un exemple d'une méthode pour compter les co-occurrences dans les données.

Partie B —8 points Supposez que vous voulez construire un modele pour déterminer si vous pourrez aller au beach un jour donné. Pour cela, vous avez collecté les données suivantes sur les conditions météorologiques passées ainsi que sur les décisions passées d'aller au beach ou pas ('oui' ou 'non').

Température	Vent	Neige	Soleil	Décision
-20	0	non	oui	non
25	5	non	oui	oui
14	2	non	non	oui
13	0	oui	oui	oui
3	30	non	oui	non

Répondez a **SEULEMENT UNE** des deux questions suivantes:

A. Décrivez l'algorithme a induction pour arbres de décision pour construire le modele et donnez un exemple d'un arbre de décision possible qui peut être construit. Votre arbre de décision doit avoir au moins deux niveaux de noeuds internes.

... suite

B. Décrivez un algorithme pour trouver les regles d' association et donnez au moins trois regles d' association possibles qui peuvent être construites.

... Fin de l' Examen Final