

# CORRECTION

## TD1 : PROBLEMATIQUE DES BDD

Algèbre et Modèle relationnelle

### 1. SENSIBILISATION A LA PROBLEMATIQUE DES BASES DE DONNEES

#### 1 APPROCHE NAÏVE

Certains d'entre vous ont une expérience des bases de données (il s'agit vraiment de quelque chose d'incontournable aujourd'hui) ou une expérience importante en développement logiciel. Dans le cadre de cet exercice, oubliez toutes vos connaissances et vos réflexions sur le sujet.

Supposons que nous adoptions la solution naïve et naturelle suivante :

- Nous créons un fichier texte comportant à l'origine une ligne par livre.
- Dans chaque ligne, on trouve les informations *titre*, auteur, éditeur, *numéro du livre* séparées par une tabulation.
- Quand une personne emprunte un livre, on complète la ligne du livre en question par les champs *nom*, *prénom*, *téléphone*, *adresse* et *date-emprunt* toujours en séparant ces informations par une tabulation.
- Lorsqu'une personne retourne un livre, il suffit d'ajouter un dernier champs *date-retour* sur la ligne du livre en question.
- Quand un livre est emprunté une nouvelle fois, on crée une nouvelle ligne avec toutes les informations concernant le livre et la personne qui l'emprunte. Bien entendu, le bibliothécaire ne ressaisit pas tout, l'application va chercher la plupart de ces informations dans le fichier.

En fait, on peut voir ce fichier texte comme un tableau de chaînes de caractères dont l'entête des colonnes serait la suivante :

Titre	Auteur	Éditeur	N° Livre	Nom	Prénom	Téléphone	Adresse	Date-emprunt	Date-retour

Supposons que l'application de gestion de bibliothèque fonctionne correctement et stocke toutes ses données dans un fichier comme celui que nous venons de décrire. Nous allons nous pencher sur les inconvénients et les conséquences inhérentes à une telle approche.

L'application fonctionne maintenant **depuis 10 ans**. Le nombre de personnes inscrites à la bibliothèque est relativement constant (bien que l'on constate un roulement) et de **5000 abonnés en moyenne par an**. Un abonné emprunte en moyenne **5 livres par mois**.

1. Quel est, approximativement, le nombre de lignes du fichier des données ?

$$5000 \text{ abonnés} \times 5 \text{ livres} \times 12 \text{ mois} \times 10 \text{ ans} = 3 \times 10^6 \text{ soit } 3 \text{ millions de lignes}$$

2. Quelle est la taille approximative du fichier sachant que chaque caractère occupe 1 octet et qu'une ligne contient, en moyenne, 150 caractères ?

$$3 \times 10^6 \times 150 = 450 \times 10^6 \text{ soit environs } 450 \text{ Mo}$$

3. Supposons qu'une personne est abonnée depuis l'origine de l'application. Elle prévient le bibliothécaire que son prénom est mal orthographié. Combien de lignes, approximativement, doivent être modifiées pour corriger cette erreur dans tout le fichier de données ?

$$12 \text{ mois} \times 5 \text{ livres} \times 10 \text{ ans} = 600 \text{ lignes}$$

# CORRECTION

## TD1 : PROBLEMATIQUE DES BDD

Algèbre et Modèle relationnelle

4. Lorsqu'un abonné emprunte un livre, le bibliothécaire saisit simplement le numéro du livre, le nom et le prénom de l'abonné. L'application se charge alors de parcourir le fichier pour rechercher les informations manquantes concernant le livre et l'abonné afin d'écrire, à la fin du fichier, la nouvelle ligne concernant l'emprunt. Dans le pire des cas, l'application doit parcourir tout le fichier. Supposons qu'un accès au fichier coûte 10ms, qu'une lecture de ligne coûte 6ms et qu'une recherche sur la ligne pour trouver le numéro du livre ou le nom et le prénom de l'abonné coûte 1ms. Quel est, dans le pire des cas, le temps mis par l'application pour compléter les informations saisies par le bibliothécaire ?

Titre	Auteur	Éditeur	N° Livre	Nom	Prénom	Téléphone	Adresse	Date-emprunt	Date-retour

Accès fichier 10ms

lecture de ligne coûte 6ms

recherche sur la ligne pour trouver le numéro du livre ou le nom et le prénom de l'abonné coûte 1ms

l'application existe depuis 10 ans soit 30 millions d'enregistrements

*Dans le pire des cas*

$3 \times 10^6 \times (6\text{ms} + 2\text{ms})$  (1 pour le N° du livre et 1 pour le nom et prénom)

$24 \times 10^6 \times 10^{-3} = 24 \times 10^3 \text{ s}$  soit 400 minutes soit 6 heures et 40 minutes !!

5. Énumérez ou résumez tous les problèmes que la représentation des données choisie (le fichier de données) semble poser.

*Redondances d'informations (données du livre et de l'emprunteur),*

*Anomalies de mise à jour (modification de x lignes pour changer le nom d'un abonné,*

*Anomalie de recherche (temps excessivement long)*

## 2 AFFINEMENT DE LA SOLUTION

Il est évident que la solution naïve décrite dans la section précédente pose de nombreux problèmes. Elle est totalement inacceptable pour une application sérieuse bien qu'elle soit encore largement employée dans des cas de petite taille (comme par exemple, dans la plupart des fichiers bibliographiques LaTeX)

Un premier affinage de la solution de la section précédente consiste à utiliser non pas un fichier unique mais quatre fichiers distincts :

- Un premier fichier est dédié au stockage des informations concernant les livres de la bibliothèque.
- Un second fichier est dédié au stockage des informations concernant les abonnés.
- Les informations stockées dans le troisième fichier vont permettre de faire la correspondance entre les deux premiers pour signifier qu'un livre donné est en cours de prêt par un abonné donné depuis une date donnée.
- Enfin, un dernier fichier va permettre de stocker l'historique des prêts. Il est similaire au troisième fichier, mais il comporte en plus une information relative à la date de retour du livre.

1. Précisez le format et les informations stockées dans chacun de ces quatre fichiers.

*Un premier fichier*

Titre	Auteur	Éditeur	N° Livre
-------	--------	---------	----------

*Un second fichier*

NumAbo	Nom	Prénom	Téléphone	Adresse
--------	-----	--------	-----------	---------

*le troisième fichier*

NumAbo	N° Livre	Date-emprunt
--------	----------	--------------

*le quatrième fichier*

NumAbo	N° Livre	Date-emprunt	Date-retour
--------	----------	--------------	-------------

# CORRECTION

## TD1 : PROBLEMATIQUE DES BDD

Algèbre et Modèle relationnelle

2. Quels sont les avantages de cette nouvelle solution ?

*Ajout/Modification aisée d'abonné ou des livres  
Ajout d'un emprunt rapide (n° du livre et de l'abonné)  
Taille des fichiers réduits*

3. Intéressons-nous au premier fichier (celui concernant les livres). Quels problèmes diagnostiquez-vous dans ce fichier ?

Titre	Auteur	Éditeur	N° Livre
-------	--------	---------	----------

*Redondance sur le nom de l'auteur et de l'éditeur*

4. Le format de ce fichier permet-il de prendre en compte des livres co-écrits par plusieurs auteurs ?

*Impossible sans réécrire deux fois la même ligne en changeant l'auteur  
Idem pour les éditeurs*

5. Quelle solution proposez-vous ?

*Création de deux nouvelles relations :  
auteurs(numAuteur, nom, prénom)  
titreauteur(numAuteur, N° livre)*

### Que retenir de ce TD ?

Les problèmes les plus courants rencontrés dans des bases de données mal conçues peuvent être regroupés selon les critères suivants :

#### Redondance des données -

Certains choix de conception entraînent une répétition des données lors de leur insertion dans la base. Cette redondance est souvent la cause d'anomalies provenant de la complexité des insertions.

C'est, par exemple, le cas de la première organisation proposée : dès qu'un abonné emprunte un livre, il faut dupliquer toutes les informations concernant l'abonné et le livre emprunté ! Au contraire, dans la deuxième solution, seuls les numéros indispensables à la distinction d'un livre et d'un abonné sont répétés dans le cas d'un emprunt.

#### Incohérence en modification -

La redondance de l'information entraîne également des risques en cas de modification d'une donnée car on oublie fréquemment de modifier toutes ses occurrences.

#### Anomalie d'insertion -

Une mauvaise conception peut parfois empêcher l'insertion d'une information, faute de connaître la valeur de tous ses champs. Pour remédier à ce problème, certains SGBD introduisent une valeur non typée qui signifie que la valeur d'un attribut est inconnue ou indéterminée. Cette valeur (appelée usuellement NULL) indique réellement une valeur inconnue et non une chaîne de caractères vide ou un entier égal à zéro.

Dans la première solution proposée, insérer un nouvel abonné qui n'a jamais emprunté de livre peut poser des problèmes. Une solution serait d'insérer des champs vides (suite de tabulations consécutives) au début de la ligne.

#### Anomalie de suppression -

Enfin, une mauvaise conception peut entraîner, lors de la suppression d'une information, la suppression d'autres informations, sémantiquement distinctes, mais indissociables dans la modélisation adoptée.

Par exemple, dans la première solution proposée, si l'on désire supprimer toutes les traces d'un livre dans le fichier de données, on fera complètement disparaître tous les abonnés qui n'ont emprunté que ce livre.