TD #3 — Introduction aux bases de données

Un TD de Nicolas Pécheux

L'objectif de ce TD est de vous sensibiliser à l'intérêt d'une théorie autour des bases de données et des avantages de l'utilisation d'un système de gestion de bases de données (SGDB). En d'autres termes, nous allons essayer d'apporter des éléments de réponse à la question suivante : *Pourquoi toute une théorie et un cours sur les bases de données alors que l'on sait très bien gérer des données avec un langage comme* Python ?

Contexte

Supposons que nous ayons à développer une application de gestion d'une bibliothèque. Tous les livres de cette bibliothèque possèdent un numéro d'exemplaire, un titre, un ou plusieurs auteurs et un éditeur. Le numéro d'exemplaire est un identifiant unique permettant de différencier les exemplaires d'un même livre. Lorsqu'une personne emprunte un livre, il faut mémoriser son nom, son prénom, la date de l'emprunt et la date de retour une fois ce dernier réalisé. Toutes les informations doivent être conservées pour garder un historique des emprunts.

Une solution simple et naïve

Notre application va devoir stocker toutes les informations précisées ci-dessus de manière persistante. Nous choisissons donc d'utiliser un fichier texte pour enregistrer cette information. Pour cela, nous adoptions naïvement la solution simple et naturelle suivante :

- Nous créons un fichier texte comportant à l'origine une ligne par livre.
- Dans chaque ligne, nous renseignons les informations *Numéro exemplaire (idEx), Titre, Auteur*, et *Éditeur* séparés par une tabulation.
- Quand une personne emprunte un livre, nous complétons la ligne du livre en question par les champs *Nom, Prénom, Renseignements (téléphone, adresse, etc.)* et *Date-emprunt* toujours en séparant ces informations par une tabulation.
- Lorsqu'une personne retourne un livre, il suffit d'ajouter une tabulation puis un dernier champ *Date-retour* sur la ligne du livre en question.
- Quand un livre est emprunté une nouvelle fois, nous ajoutons une nouvelle ligne avec toutes les informations concernant le livre et la personne qui l'emprunte. Bien entendu, le ou la bibliothécaire ne ressaisit pas tout, l'application va chercher la plupart de ces informations dans le fichier.

Le fichier en question peut donc être perçu comme un tableau de chaînes de caractères tel celui représenté Figure 1. Nous parlerons également de *table* (ou plus tard, de *relation*) plutôt que de tableau ou de fichier.

Emprunts								
idEx	Titre	Auteur	Éditeur	Nom	Prénom	Rens.	Date-emp.	Date-ret.
1	La volonté de puissance	Nietzsche	Gallimard					
2	Espoir-du-cerf	O Scott Card	Denoël	Michel	Tom	...	20 oct. 2008	07 nov. 2008
3	Vendredi ou la vie sauvage	M. Tournier	Poche	Moreau	J. Batiste	...	02/10/2009	
4	Élevation	D. Brin	J'AI LU	Laurent	Camille	...	02/10/2009	03/10/2009
5	Vendredi ou la vie sauvage	M. Tournier	Poche	Moreau	J. Batiste	...	02/10/2009	
6	Vendredi ou la vie sauvage	M. Tournier	Poche					
7	Ainsi parlait Zarthoustra	F. Nietzshe	Poche					
8	Humain, trop humain	F NIETSCHE	POCHE					
9	Les maîtres chanteurs	O.S. CARD	Denoel	Moreau	J. Batiste	...	08/10/2009	
10	Les maîtres chanteurs	O.S. CARD	Denoel					
11	Les maîtres chanteurs	O.S. CARD	Denoel					
12	St-Exupéry, Terre des hommes	F. Brin	Broché					
13	Rédemption	Brin	J'ai lu					
2	Espoir-du-cerf	O Scott Card	Denoël	Roux	Sarah	...	20/11/2009	
4	Élevation	D. Brin	J'AI LU	Dubois	Mathis	...	25/11/2009	

FIGURE 1 – Exemple illustrant le format et le contenu du fichier utilisé pour enregistrer les données de notre application de gestion d'une bibliothèque.

. . .Mais pas sans conséquences

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 inhérents à une telle approche.
L'application fonctionne maintenant depuis 20 ans. Le nombre de personnes inscrites à la bibliothèque est relativement constant et de 5 000 personnes en moyenne par an. Un abonné emprunte en moyenne 2 livres par mois.

1. Quel est, approximativement, le nombre de lignes du fichier de données ?
2. Quelle est, en mébioctets (c.-à.-d., en multiple de 2^{20} octets), la taille approximative du fichier sachant que chaque caractère occupe 1 octet et qu'une ligne contient, en moyenne, 200 caractères ?
3. Lorsqu'un abonné emprunte un livre, le ou la bibliothécaire saisit simplement le numéro de l'exemplaire et 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 de préremplir, à la fin du fichier, la nouvelle ligne concernant l'emprunt. Supposons qu'un accès au fichier coûte 8 *ms* (c'est le temps d'accès moyen au disque dur), qu'une lecture d'une ligne coûte 0.1 *ms* (temps pour lire les 200 caractères de la ligne) et qu'une recherche sur la ligne pour trouver le numéro de l'exemplaire ou le nom et le prénom de l'abonné coûte 0.01 *ms*. Quel est, dans le pire des cas, le temps mis par l'application pour compléter les informations saisies par le ou la bibliothécaire ?
4. Supposons qu'une personne est abonnée depuis l'origine de l'application. Elle prévient le ou la bibliothécaire que son prénom est mal orthographié. Combien de lignes, approximativement, doivent être modifiées pour corriger cette erreur dans tous le fichier de données ? Combien de temps cela prendrait-il en supposant que corriger une ligne prend le même temps que de la lire ?
5. La base de données permet-elle vraiment de retrouver les informations ? Par exemple, en se référant au tableau de la Figure 1, réfléchir et décrire précisément un moyen algorithmique de retrouver les informations suivantes :
   - **i.** Quels sont les livres édités chez *Poche* ?
   - **ii.** Quels sont les livres édités chez *Denoël* ?
   - **iii.** Quels sont les livres écrits par *Orson Scott Card* ?
   - **iv.** Quels sont les livres écrits par *Friedrich Nietzsche* ?
   - **v.** Quels sont les livres écrits par *David Brin* ?
6. Supposons la situation suivante. M. Jean-Batiste Moreau et son fils, Jean-Batiste également, ont tous les deux emprunté un exemplaire de deux livres différents, de même titre *Vendredi ou la vie sauvage*, de même auteur *Michel Tournier*, tous les deux édités chez *Poche* mais dont l'un est illustré (par *Gérard Franquin*) mais pas l'autre. Lorsqu'il vient rendre les deux livres, le père précise que le prénom de son fils est *Jean-Batiste Junior* et non pas *Jean-Batiste*. Il remarque également que le livre qu'il (le père) vient d'emprunter, qui porte le numéro 3, est coécrit par *Michel Tournier* et *Gérard Franquin* et pas simplement par *Michel Tournier*. Est-il possible de corriger ces erreurs dans notre fichier ?
7. Énumérer tous les problèmes que la représentation des données choisie semble poser.

Affinement de la solution

Il est maintenant 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 dans les fichiers bibliographiques LaTeX).
Pour résoudre les problèmes d'incohérence concernant les auteurs, nous proposons de décomposer le tableau de départ en deux sous-tableaux comme illustré par la Figure 2. Les colonnes *idAu* permettent de faire le lien entre les deux tables. Observer comment la redondance sur les noms des auteurs a été éradiquée dans cette solution.

8. Cette décomposition a-t-elle engendré une perte d'information ? Autrement dit, est-il possible de reconstituer la table originale à partir de cette décomposition ?
9. Pouvons-nous maintenant répondre aux requêtes suivantes :
   - **i.** Quels sont les livres écrits par *Orson Scott Card* ?
   - **ii.** Quels sont les livres écrits par *Friedrich Nietzsche* ?
   - **iii.** Quels sont les livres écrits par *David Brin* ?
10. Sur le même principe, proposer une solution pour que chaque livre ne soit représenté qu'une seule fois dans notre base de données. Dans cette perspective, nous précisons que deux livres distincts portent le même titre : *Vendredi ou la vie sauvage*. Le premier livre existe en deux exemplaires à la bibliothèque, le 3 et le 6, le second en un seul exemplaire, le 5.
11. Toujours en appliquant la même méthode, supprimer les redondances concernant la mention des éditeurs et les informations associées aux abonnés. Pour ce faire, nous précisons que l'abonné qui a emprunté *Les maîtres chanteurs* est le père.

Emprunts								
idEx	Titre	idAu	Éditeur	Nom	Prénom	Rens.	Date-emp.	Date-ret.
1	La volonté de puissance	1	Gallimard					
2	Espoir-du-cerf	2	Denoël	Michel	Tom	...	20 oct. 2008	07 nov. 2008
3	Vendredi ou la vie sauvage	3	Poche	Moreau	J. Batiste	...	02/10/2009	
4	Élevation	4	J'AI LU	Laurent	Camille	...	02/10/2009	03/10/2009
5	Vendredi ou la vie sauvage	3	Poche	Moreau	J. Batiste	...	02/10/2009	
6	Vendredi ou la vie sauvage	3	Poche					
7	Ainsi parlait Zarthoustra	1	Poche					
8	Humain, trop humain	1	POCHE					
9	Les maîtres chanteurs	2	Denoel	Moreau	J. Batiste	...	08/10/2009	
10	Les maîtres chanteurs	2	Denoel					
11	Les maîtres chanteurs	2	Denoel					
12	St-Exupéry, Terre des hommes	5	Broché					
13	Rédemption	4	J'ai lu					
2	Espoir-du-cerf	2	Denoël	Roux	Sarah	...	20/11/2009	
4	Élevation	4	J'AI LU	Dubois	Mathis	...	25/11/2009	

Auteurs		
idAu	Prénom	Nom
1	Friedrich	Nietzsche
2	Orson Scott	Card
3	Michel	Tournier
4	David	Brin
5	Françoise	Brin

FIGURE 2 – Décomposition initiale de la table en deux sous-tables pour résoudre les problèmes d'incohérences concernant les auteurs.

12. Notre base de données est-elle encore largement entachée de redondances ? Où se situent-elles ?
13. Proposer une solution pour corriger la ou les redondances détectées dans la question précédente.
14. Proposer une solution pour tenir compte du cas des livres coécrits par plusieurs auteurs. Vous devriez obtenir un ensemble de tables semblable à celui de la Figure 3.
15. Reprenons la situation décrite à la question 6. On suppose que le père et le fils possèdent chacun leur carte d'abonné où figure leur numéro d'abonné (respectivement $idAb = 2$ et $idAb = 4$) et chacun des exemplaires empruntés possède un autocollant où figure le numéro d'exemplaire (respectivement $idEx = 3$ et $idEx = 5$). Montrer que ces corrections ne posent plus de problème pour notre nouvelle base.
16. Reprenons la situation de la question 4. Combien de lignes faut-il modifier et combien de temps cela prendrait-il pour effectuer la correction ?
17. À partir de la base de donnée de la Figure 3, expliquer très précisément comment on pourrait s'y prendre pour répondre aux requêtes suivantes. On demande de décrire un algorithme (si possible efficace), en gardant en tête que ces tables peuvent comporter plusieurs milliers de lignes.

- **i.** Quels sont les titres de livres que possède la bibliothèque ?
- **ii.** Quels sont les prénoms des auteurs dont le nom est *Brin* ?
- **iii.** Quels sont les prénoms qui sont à la fois ceux d'un auteur et d'un abonné ?
- **iv.** Quelles sont les associations possibles et imaginables entre un abonné et un titre de livre qu'il peut emprunter dans cette bibliothèque ?
- **v.** Comment associer chaque auteur avec les titres des livres dont il est auteur (ou coauteur) ?
- **vi.** Quels sont les abonnés qui ont emprunté tous les livres de de la bibliothèque ?
- **vii.** Quels sont les livres dont un emprunt est en cours depuis moins de dix jours et qui sont écrits par un seul auteur, dont de plus le nom commence par la lettre *M* ?
- **viii.** Quels sont les éditeurs qui ont édité au moins un livre emprunté plus de trente jours par un abonné dont le prénom est aussi celui de trois auteurs ayant écrit un livre publié chez un même éditeur ?

18. Proposer encore des améliorations possibles à la base de données obtenue et représentée à la Figure 3.

Conclusion

La base de données à laquelle nous avons abouti (la Figure 3) est une base de données relationnelle. La conception et la gestion des bases de données sont un problème complexe extrêmement important puisque les bases de données se trouvent aujourd'hui au cœur de tous les systèmes d'information. C'est pourquoi tous ces problèmes ont été largement étudiés et des solutions fiables et éprouvées ont été trouvées. De nombreux travaux ont ainsi permis de mettre au

point une théorie permettant la conception de bases de données *bien formées*. La problématique de la gestion des bases de données trouve une solution dans l'utilisation d'un *Système de Gestion de Bases de Données* (SGDB). Ni la conception ni la gestion des bases de données ne sont au programme des classes préparatoires. Dans la suite du cours, nous allons nous intéresser principalement à l'*interrogation* d'une base de données déjà constituée sous forme de *requêtes*, comme par exemple celles des questions 5, 9 et 17.

Corrigé

Une fois le TD terminé, vous devriez avoir abouti à une solution semblable à celle de la Figure 3 ci-dessous :

Exemplaires	
idEx	idLi
1	1
2	2
3	3
4	4
5	5
6	3
7	6
8	7
9	8
10	8
11	8
12	9
13	10

Livres		
idLi	Titre	idEd.
1	La volonté de puissance	1
2	Espoir-du-cerf	2
3	Vendredi ou la vie sauvage	3
4	Élevation	4
5	Vendredi ou la vie sauvage	3
6	Ainsi parlait Zarthoustra	3
7	Humain, trop humain	3
8	Les maîtres chanteurs	2
9	St-Exupéry, Terre des hommes	5
10	Rédemption	4

Éditeurs	
idEd	Nom
1	Gallimard
2	Denoël
3	Poche
4	J'ai lu
5	Broché

AuteurDe	
idLi	idAu
1	1
2	2
3	3
4	4
5	3
6	1
7	1
8	2
9	5
10	4

Auteurs		
idAu	Prénom	Nom
1	Friedrich	Nietzsche
2	Orson Scott	Card
3	Michel	Tournier
4	David	Brin
5	Françoise	Brin

Abonnés			
idAb	Nom	Prénom	Rens.
1	Michel	Tom	...
2	Moreau	J. Batiste	...
3	Laurent	Camille	...
4	Moreau	J. Batiste	...
5	Roux	Sarah	...
6	Dubois	Mathis	...

Emprunts			
idEx	idAb	Date-emp.	Date-ret.
2	1	20/10/2008	07/11/2008
3	2	02/10/2009	
4	3	02/10/2009	03/10/2009
5	4	02/10/2009	
9	2	08/10/2009	
2	5	20/11/2009	
4	6	25/11/2009	

FIGURE 3 – Ensemble de tables obtenues pour représenter la base de données de notre application de gestion d'une bibliothèque.

Références

Bases de données de la modélisation au SQL, Laurent AUDIBERT, Ellipses.

Pour la semaine prochaine

https://sql-island.informatik.uni-kl.de/