Activité de découverte: insuffisance des structures de données plates

Certains d'entre vous ont probablement déjà entendu parler des bases de données. Dans le cadre de cette activité, oublions, dans un premier temps, toutes nos connaissances et nos réflexions sur le sujet.

Contexte

Supposons que nous ayons à développer une application de gestion d'une bibliothèque. Tous les livres de la 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**, son **numéro de téléphone**, son **adresse**, 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 de type `csv` pour enregistrer les données. Pour cela, nous adoptons naïvement la solution simple et naturelle suivante :

— Nous créons un fichier texte `csv` comportant à l'origine une ligne par livre.
— Dans chaque ligne, nous renseignons les informations: `idEx` (*numéro d'exemplaire*), `Titre`, `Auteur` et `Éditeur`, séparées par une virgule.
— Quand une personne emprunte un livre, nous complétons la ligne du livre emprunté par les champs: `Nom`, `Prénom`, `Téléphone`, `Adresse` et `Date-emprunt` toujours en séparant ces informations par une virgule.
— Lorsqu'une personne ramène un livre, il suffit de compléter le dernier champ `Date-retour` sur la ligne du livre emprunté.
— 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 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. Nous parlerons également de table plutôt que de tableau ou de fichier.

À faire

Ouvrir le fichier `biblio.csv` pour avoir une idée plus précise du contenu.

idEx	Titre	Auteur	Editeur	Nom	Prenom	tel
1	La volonté de puissance	Nietzsche	Gallimard			
2	Espoir-du-cerf	O Scott Card	Denoël	Michel	Tom	
3	Vendredi ou la vie sauvage	M. Tournier	Poche	Moreau	J. Batiste	
4	Élévation	D. Brin	J'AI LU	Laurent	Camille	
5	Vendredi ou la vie sauvage	M. Tournier	Poche	Moreau	J. Batiste	
6	Vendredi ou la vie sauvage	M. Tournier	Poche			
7	Ainsi parlait Zarathoustra	F. Nietzsche	Poche			
8	Humain trop humain	F. NIETZSCHE	POCHE			
9	Les maîtres chanteurs	O.S. Card	Denoël	Moreau	J. Batiste	
10	Les maîtres chanteurs	O.S. Card	Denoel			
11	Les maîtres chanteurs	O.S. Card	Denoël			
12	St-Exupery Terre des hommes	F. Brin	Broché			

idEx	Titre	Auteur	Editeur	Nom	Prenom	tel
13	Rédemption	Brin	J’ai lu			
2	Espoir-du-cerf	O Scott Card	Denoël	Roux	Sarah	
4	Élévation	D. Brin	J’AI LU	Dubois	Mathis	

...Mais pas sans conséquence

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 20 ans. Le nombre de personnes s'inscrivant à la bibliothèque est d'environ 5000 par an. Un abonné emprunte en moyenne 2 livres par mois.

1. Montrer qu'une année de fonctionnement entraînera la création de $120\ 000$ lignes. En déduire le nombre de lignes approximatif du fichier des données, au bout de 20 ans?

2. Quelle est la taille approximative du fichier sachant que chaque caractère est codé sur 1 octet et qu'une ligne contient, en moyenne, 200 caractères? *Donner le résultat en Mo*.

3. Lorsqu'un abonné emprunte un livre, le 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. Dans le pire des cas, l'application doit parcourir tout le fichier.

Supposons:

- qu'un accès au fichier coûte 8 ms (*c'est le temps d'accès moyen au disque dur*);
- qu'une lecture de ligne coûte 0.1 ms (*temps pour lire les 200 caractères de la ligne*);
- 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 bibliothécaire ? *Donner le résultat en seconde puis en minutes secondes*.

4. 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?

5. Cette base de données permet-elle vraiment de retrouver des informations ? Par exemple, en se rappelant du travail fait en 1re sur l'exploitation des fichiers `csv` (*en python*), réfléchir au moyen de retrouver les informations suivantes :

- Quels sont les livres édités chez Poche?
- Quels sont les livres édités chez Denoël?
- Quels sont les livres écrits par Orson Scott Card?
- Quels sont les livres écrits par Friedrich Nietzsche?
- Quels sont les livres écrits par David Brin?

Indice

On pourra utiliser le cadeau du professeur:

```
import csv


with open('biblio.csv', newline='', encoding='utf8') as fp:
    table = [dict(ligne) for ligne in csv.DictReader(fp, delimiter=',')]
```

6. La mise à jour.

Analyse d'une situation

M. Moreau Jean-Batiste et son fils, dont le prénom est également Jean-Batiste, ont tous les deux emprunté un exemplaire d'un livre: *Vendredi ou la vie sauvage de Michel Tournier chez Poche* et *Vendredi ou la vie sauvage de Michel Tournier chez Poche*.
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, n'est pas écrit par `Michel Tournier`, mais qu'il est coécrit par `Michel Tournier et Gérard Franquin`.

Est-il possible de corriger ces erreurs dans notre fichier ?

7. Énumérer ou résumer tous les problèmes que la représentation des données choisie (c'est-à-dire utilisation d'un fichier de données) 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.
Pour résoudre les problèmes d'incohérences concernant les auteurs, nous proposons de décomposer le tableau de départ en deux sous-tableaux (voir les fichiers `EMPRUNTS_Q8.csv` et `AUTEURS.csv`).
Les colonnes `idAu` permettent de faire le lien entre les deux tables. La redondance sur les noms des auteurs est-elle toujours présente ?

Table EMPRUNTS

idEx	Titre	idAu	Editeur	Nom	Prenom	tel
1	La volonté de puissance	1	Gallimard			
2	Espoir-du-cerf	2	Denoël	Michel	Tom	
3	Vendredi ou la vie sauvage	3	Poche	Moreau	J. Batiste	
4	Élévation	4	J'AI LU	Laurent	Camille	
5	Vendredi ou la vie sauvage	3	Poche	Moreau	J. Batiste	
6	Vendredi ou la vie sauvage	3	Poche			
7	Ainsi parlait Zarathoustra	1	Poche			
8	Humain trop humain	1	POCHE			
9	Les maîtres chanteurs	2	Denoël	Moreau	J. Batiste	
10	Les maîtres chanteurs	2	Denoël			
11	Les maîtres chanteurs	2	Denoël			
12	St-Exupery Terre des hommes	5	Broché			

idEx	Titre	idAu	Editeur	Nom	Prenom	tel
13	Rédemption	4	J’ai lu			
2	Espoir-du-cerf	2	Denoël	Roux	Sarah	
4	Élévation	4	J’AI LU	Dubois	Mathis	

Table AUTEURS

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

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 :

- Quels sont les livres écrits par Orson Scott Card?
- Quels sont les livres écrits par Friedrich Nietzsche?
- Quels sont les livres écrits par David Brin?

10. Sur le même principe, proposer une solution pour que le titre de chaque livre ne soit représenté qu'une seule fois dans notre base de données.

Indication

On pourra introduire une table supplémentaire `LIVRES`.

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 le livre *Les maîtres chanteurs* est le père.

12. Notre base de données comporte encore des redondances. Où se situent-elles ?

Aide

On pourra lire les fichiers avec python suivant la méthode habituelle:

```
import csv


with open('EMPRUNTS_Q8.csv', newline='', encoding='utf8') as emprunts:
    tab_emp = [dict(ligne) for ligne in csv.DictReader(emprunts, delimiter=',')]
with open('AUTEURS.csv', newline='', encoding='utf8') as auteurs:
    tab_aut = [dict(ligne) for ligne in csv.DictReader(auteurs, delimiter=',')]
```