Un exemple naïf sans Base de Données

Au sein d'une technopole de biotechnologies, on veut stocker les données utiles pour gérer une plateforme d'achat et de mutualisation de matériels entre plusieurs startups et PME. Chaque matériel, du plus petit au plus gros (de la simple chaise à l'énorme microscope électronique à balayage) appartient à l'entreprise qui l'a acheté (avec l'aide éventuelle de la technopole) et elle sera bien sûr la première à l'utiliser. Par la suite, d'autres entreprises pourront l'emprunter pour une période donnée, surtout bien sûr pour les matériels les plus chers.
Chaque entreprise de la technopole est caractérisée par son nom, son adresse et le numéro de téléphone pour la joindre. Un matériel est caractérisé par son nom exact, son type (meuble, microscope, séquenceur, réfrigérateur, *etc.*), l'entreprise qui en est propriétaire, la date d'achat et une description courte de sa fonction.
La technopole doit conserver l'historique complet des achats et utilisations de chaque matériel, pour optimiser ou cofinancer des achats groupés, rendre des comptes aux investisseurs, *etc.*

Une solution simple pour commencer est d'utiliser un tableur avec 10 colonnes :

nom PME	tél PME	adrs PME	identifiant matériel	type matériel	proprio matériel	date achat	fonction matériel	date emprunt	date rendu
GenoData	404	7, rue Buffon	MS78bis	microscope	GenoData	01/01/10	à balayage, confocal,...	01/01/10	05/02/21 (obsolète)
ProtéoFun	218	2, bd Watson&Crick	VS89	vidéo proj.	ProtéoFun	15/12/10	visualisation 3D, TrueColors,...	15/12/10	14/7/15 (HS)
GenoData	404	7, rue Buffon	VS89	vidéo proj.	ProtéoFun	15/12/10	visualisation 3D, TrueColors,...	04/05/11	07/07/11
...	...		...	...	...	...	...	...	...

La description fonctionnelle des matériels est bien sûr plus conséquente que dans cet exemple. Implicitement, un matériel rendu est de nouveau chez son propriétaire ; il est inutile d'ajouter une ligne pour ça dans le tableau. De même la date de rendu, pour le propriétaire du matériel, est lorsqu'il est mis hors service ou bien lorsque la PME quitte la technopole, inutile d'ajouter des colonnes pour ça. Il y a donc une nouvelle ligne pour chaque nouvel achat ou chaque nouvel emprunt au sein de la technopole.

1 Quelques chiffres...

Supposons maintenant que la technopole soit un succès depuis une vingtaine d'années, avec une moyenne de 50000 matériels inventoriés et environ 2500 startups ou PME[1]. Une PME fait appel au service de prêt environ 5 fois par mois.

Exercice 1 : Quel est approximativement le nombre de lignes du tableau ?

Exercice 2 : Une ligne du tableau fait en moyenne 400 caractères ; quelle est la taille du fichier contenant le tableau, en ignorant la place occupée par les données de mise en page, de fontes, *etc* ?

Exercice 3 : Une startup de la technopole est rachetée par une autre entreprise et change de nom. Pour garder un historique cohérent, il faut mettre à jour toutes les lignes du tableau. Combien de lignes faut-il modifier en moyenne ?

Pour ce faire (de même que pour de nombreuses autres opérations comme par exemple charger le fichier, s'assurer qu'en ajoutant un nouveau matériel on ne lui attribue pas un nom préexistant, *etc.*) il faut que l'application en charge de cette fonction (tableur ou autre) parcoure tout le fichier.

Exercice 4 : Lire une ligne de 400 caractères dans un tableur prend en moyenne 0,1ms sur un ordinateur de bureau standard. Il faut ensuite faire une recherche de mots dans la ligne pour savoir s'il faut la traiter (0,01ms) puis faire le traitement si nécessaire (0,01ms). Une fois toutes les lignes parcourues, il faut réécrire le fichier modifié (0,1ms par ligne). Quel est le temps moyen nécessaire à l'application pour faire ce type d'opération ?
Quel opérateur accepterait d'attendre ce temps entre chaque modification ou recherche ? Et si l'ordinateur était 10 fois plus rapide ?

La technopole souhaite maintenant évaluer quels matériels elle aurait intérêt à acheter en son nom propre plutôt que d'octroyer des aides à une des PME pour qu'elle l'achète. L'idée est que tout matériel qui est plus longtemps en prêt qu'utilisé par la PME propriétaire mérite d'être acheté par la technopole. Pour cela, il va falloir créer un tableau des matériels contenant au moins 3 colonnes :

identifiant matériel	Durée de vie (jours)	Jours empruntés
MS78bis	4045	1589
VS89	2013	1500
XS18ZR30	450	56
...	...	...

1. C'est la taille approximative de la technopole de Sophia Antipolis.

Exercice 5 : Calculez, très approximativement, quel temps de calcul est nécessaire pour créer ce tableau, sachant que la lecture ou écriture d'une ligne de ce nouveau tableau (environ 20 caractères) prendrait 0,01ms ?
Quel opérateur accepterait d'attendre ce temps ? Et si l'ordinateur était 10 ou même 100 fois plus rapide ?
Indications : on peut par exemple successivement *(1)* lire chaque ligne du tableau d'origine *(2)* y récupérer le nom du matériel et les dates puis parcourir le tableau du matériel et *(2bis)* si le matériel est déjà dedans, mettre à jour le nombre de jours de prêt, sinon ajouter une ligne à la fin avec la durée de vie du matériel et *(3)* sauver le tableau du matériel. On négligera les temps de calcul des nombres de jours à partir des dates.

Exercice 6 : Quelles sont, selon vous, les principales maladresses et les principaux problèmes engendrés par le tableau de données choisi initialement ?

2 Une meilleure solution

Aussi naïve et inacceptable en temps de calcul qu'elle soit, cette solution fondée sur l'accumulation des données récoltées dans un unique tableau est malheureusement celle qui est le plus souvent employée dans les laboratoires de biologie (mais rarement dans les autres industries, heureusement)... Pourtant, sous réserve de pouvoir « naviguer » facilement d'un tableau à un autre, et s'il existait des programmes qui le font, on pourrait répartir les données dans plusieurs tableaux afin de ne plus recopier plusieurs fois les mêmes informations.

Exercice 7 : Avec plusieurs tableaux, comment pourriez-vous faire pour réduire, et même supprimer, les redondances dans les cases ? On peut par exemple réduire la taille des lignes trop longues en isolant les grosses données dans un tableau à part, moins souvent parcouru, mais on peut faire plus.

Pour simplifier les calculs, on va considérer dans toute la suite que le temps de lecture ou d'écriture d'une ligne raisonnablement courte d'un tableau de la question précédente est de 0,01ms. Par exemple les nom, tél et adrs d'une entreprise font moins de 40 caractères, et les identifiant, type, proprio et dates d'achat/obsolescence d'un matériel également, ainsi que des dates d'emprunt/retour. En revanche il faut bien compter 300 caractères pour décrire la fonction d'un matériel, ce qu'on ne va pas considérer comme « raisonnablement court ».

Exercice 8 : Quelle est la taille cumulée de tous ces tableaux réunis ?

Exercice 9 : Comment modifieriez-vous vos tableaux pour qu'il puisse y avoir plusieurs exemplaires d'un même matériel (même type, même modèle) ?

Exercice 10 : Combien de lignes faut-il modifier pour changer le nom d'une entreprise ? quel temps de calcul cela prendra-t-il ? Peut-on modifier les tableaux pour réduire ce temps de calcul ?

Exercice 11 : Quel temps de calcul approximatif est nécessaire pour calculer le tableau de l'exercice 5, en suivant les mêmes étapes ? Pourquoi n'a-t-on rien gagné ?

3 Discussion

Les problèmes les plus courants rencontrés dans la gestion des données résultent souvent de l'ignorance des principes élémentaires des bases de données. Ces problèmes peuvent être regroupés selon les sujets suivants :

- **Redondance des données.** Certains choix de stockage entraînent une répétition des données. Cette redondance produit souvent des tableaux de taille inutilement gigantesque et cause d'énormes pertes de temps, jusqu'à rendre l'exploitation de ces données impossible.
- **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.
- **Les données incomplètes.** Un cas typique peut être de ne pas connaître la valeur de tous les champs d'une insertion. Par exemple, on ne connaît pas la date d'obsolescence d'un matériel lors de son achat. Pour remédier à ce problème, les Systèmes de Gestion de Bases de Données (SGBD) introduisent une valeur appelée `NULL` qui signifie « valeur inconnue ou indéterminée ». Cette valeur indique réellement une valeur *inconnue* et *non pas* une chaîne de caractères vide, une date initiale particulière, ou un entier égal à zéro.
- **Anomalie de suppression.** Enfin, une mauvaise gestion peut entraîner, lors de la suppression d'une information, la création d'incohérences. Typiquement supprimer un matériel en oubliant de supprimer ses informations d'emprunts.
- **Parcours linéaires à répétition de tableaux.** En fait, les SGBD mémorisent leurs tableaux sous une forme dite « arborescente », bien plus intelligente que des tableaux, et cela diminue considérablement les temps de recherche.

La gestion de données en grand nombre, fréquente en biologie, dépasse donc largement les capacités d'un tableur. Les outils de gestion de bases de données, qui permettent de gérer correctement des données découpées en plusieurs tables, sont la solution. Une bonne compréhension de leur usage est indispensable à tout ingénieur. C'est l'objectif de ce cours.