Les Grandes Découvertes en Bases de Données

Jef Wijsen
Université de Mons-Hainaut
Service de Science des Systèmes d'Information
jef.wijsen@umh.ac.be
http://saturn.umh.ac.be/~wijsen/

Journée de Mathématique et de Sciences
29 mars 2001

1 Introduction

Dans ma jeunesse, j'étais passionné par les livres racontant les grandes découvertes scientifiques. Les découvertes dans le domaine des bases de données (BD) n'étaient pas parmi eux... Dans cet exposé, j'essaie néanmoins de montrer que les BD sont devenues un domaine de recherche intéressant et important en informatique depuis les années 60. J'explique quelles étaient les étapes principales de ces recherches et quels sont les problèmes restant à résoudre. Deux questions fondamentales posées par cette discipline sont:

1. Comment les données peuvent-elles être structurées?
2. Comment les données peuvent-elles être interrogées?

Le terme "requête" est utilisé pour une question posée à une BD en un langage interprété par l'ordinateur.

2 Les BD Hiérarchiques

Le premier système de BD a été conçu pour la gestion des données du projet Apollo de la NASA. Les données étaient structurées dans des hiérarchies, comparables à l'organisation des répertoires sur un PC. La figure 1 donne un exemple d'une telle hiérarchie; elle montre des

Classe
  Ordre
    Espèce

Mammifères
  Carnivores
    Lion
    Loup
    Tigre
    Hyène
  Artiodactyles
    Girafe
    Hippopotame
Reptiles
  Serpents
    Cobra

Figure 1: Classification hiérarchique des animaux.

animaux (Lion, Loup, Tigre,...) groupés dans des ordres (Carnivores, Artiodactyles, Serpents) qui eux-mêmes sont groupés dans des classes (Mammifères, Reptiles). Une telle structuration des données permet de répondre facilement aux questions de type:

Quels animaux sont carnivores?

Supposons maintenant qu'on veuille ajouter des informations sur la répartition géographique des animaux. Au moins deux possibilités se présentent. La figure 2 (*gauche*) ajoute les continents au plus bas niveau de la hiérarchie. Grâce

à cette organisation, il est très facile de répondre à la question:

Où peut-on trouver des lions?

Par contre, pour répondre à la question:

Quels sont les carnivores d'Afrique?

la hiérarchie montrée par la figure 2 (*droite*) convient mieux, parce que les carnivores africains (Lion et Hyène) se retrouvent groupés.

Bien que personne ne mettra en doute la hiérarchie représentée par la figure 1, on observe qu'ajouter les continents peut se faire de plusieurs manières et qu'il n'y a pas d'organisation idéale pour toutes les requêtes. Il est facile de comprendre pourquoi: la relation entre les continents et les espèces n'est pas de nature hiérarchique, dans le sens où une espèce n'est pas limitée à un seul continent (et inversement, bien sûr, un continent contient plusieurs espèces). Il n'est donc pas naturel de vouloir stocker une telle relation dans une hiérarchie. On est enclin à croire qu'une structuration des données en réseau est plus naturelle qu'une organisation hiérarchique. Ce sont sans doute de telles considérations qui ont mené aux BD de type réseau.

3 Les BD de Type Réseau

Ce modèle de données sera toujours associé au nom de C.W. Bachman. La figure 3 montre les mêmes données zoo-géographiques structurées en réseau. Les rectangles contiennent les données et les "circuits" représentent les relations entre les données. Par exemple, on reconnaît facilement le circuit qui relie les carnivores; celui-ci permet de répondre à la question:

Quels animaux sont carnivores?

Pour répondre à la question:

Quels sont les animaux d'Asie?

il faut parcourir un chemin qui contient plusieurs circuits (lesquels?). Finalement, pour répondre à la question:

Quels sont les carnivores d'Asie?

plusieurs parcours sont possibles. Tout d'abord, on peut traverser le circuit qui relie les carnivores et sélectionner ceux qui sont liés à l'Asie. Alternativement, on peut partir du nœud "Asie", parcourir les animaux asiatiques et sélectionner ceux qui se trouvent dans le circuit qui relie les carnivores.

Il faut comprendre qu'il n'est pas évident d'exprimer ces parcours en un langage de programmation. A titre d'exemple, le programme montré par la figure 4 décrit le parcours qui trouve les carnivores asiatiques à partir du nœud "Asie". Un tel programme est appelé "navigationnel": le programmeur doit diriger de manière détaillée le parcours à travers les données en indiquant pas à pas les opérations à réaliser [1].

En 1973, C.W. Bachman a reçu le Prix Turing pour sa contribution à l'informatique. Ce prix signifie pour un informaticien ce qui signifie le Prix Nobel pour un physicien ou un chimiste.

4 Les BD Relationnelles

En 1970, au moment où les systèmes basés sur le modèle hiérarchique ou le modèle en réseau étaient en plein développement, E.F. Codd publiait un article [2] où il proposait de stocker des données dans des tables. A l'heure actuelle, cette solution peut nous sembler assez évidente; pensons aux tables utilisées pour afficher les scores des matchs de football ou les listes de prix... Néanmoins, en 1970 cette idée était considérée comme une curiosité intellectuelle. On doutait que les tables puissent jamais être gérées de manière efficace par un ordinateur...

Une table se compose de plusieurs colonnes et rangées. Pour notre exemple, les tables sont celles de la figure 5. En général, les tables et leurs colonnes sont fixées au moment de la conception

Les continents (Afrique, Asie, Europe) sont ajoutés à la base de la hiérarchie:

Classe
Ordre
Espèce
Continent

Mammifères
Carnivores
Lion
Afrique
Asie
Loup
Europe
Asie
Tigre
Asie
Hyène
Afrique
Artiodactyles
Girafe
Afrique
Hippopotame
Afrique
Reptiles
Serpents
Cobra
Asie

Les animaux de même ordre sont groupés par continent:

Classe
Ordre
Continent
Espèce

Mammifères
Carnivores
Afrique
Lion
Hyène
Asie
Loup
Lion
Tigre
Europe
Loup
Artiodactyles
Afrique
Girafe
Hippopotame
Reptiles
Serpents
Asie
Cobra

Figure 2: Deux façons d'ajouter la répartition géographique à la classification hiérarchique des animaux.

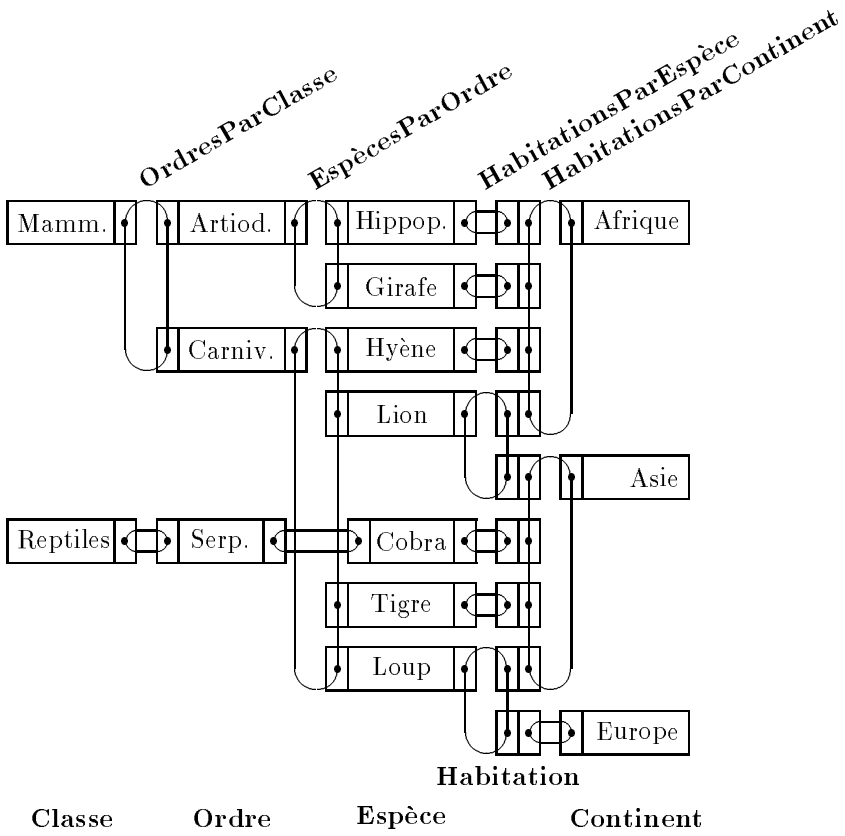

Figure 3: Classification et répartition des animaux dans une BD de type réseau.

```
FIND Continent WITHIN IndexSurContinents USING 'Asie';
FIND FIRST Habitation WITHIN HabitationsParContinent;
WHILE db-rec-found LOOP
      OBTAIN OWNER WITHIN HabitationsParEspèce;
      OBTAIN OWNER WITHIN EspècesParOrdre;
      IF Ordre = 'Carnivores' THEN print Espèce END-IF;
      FIND NEXT Habitation WITHIN HabitationsParContinent;
END-LOOP;
```

Figure 4: Programme navigationnel pour trouver les carnivores asiatiques.

EO

Espèce	Ordre
Lion	Carnivores
Loup	Carnivores
Tigre	Carnivores
Hyène	Carnivores
Girafe	Artiodactyles
Hippopotame	Artiodactyles
Cobra	Serpents

OC

Ordre	Classe
Carnivores	Mammifères
Artiodactyles	Mammifères
Serpents	Reptiles

VIT

Espèce	Continent
Lion	Afrique
Lion	Asie
Loup	Europe
Loup	Asie
Tigre	Asie
Hyène	Afrique
Girafe	Afrique
Hippopotame	Afrique
Cobra	Asie

Figure 5: Classification et répartition des animaux dans une BD relationnelle avec trois tables.

de la BD. Après, on peut à tout moment changer le contenu des tables en insérant, en modifiant et en supprimant des rangées.

Dans le même article, E.F. Codd proposait d'utiliser une algèbre pour interroger les tables. L'algèbre proposée se compose de cinq opérateurs, parmi lesquels:

`JOIN` La jointure sert à joindre les rangées de deux tables. Les rangées à joindre sont celles qui ont la même valeur pour toute colonne commune aux deux tables.

`SELECT` La sélection sert à retenir les rangées qui vérifient une certaine condition, en supprimant les autres rangées.

`PROJECT` La projection sert à retenir certaines colonnes, en supprimant les autres.

Notons que le résultat de ces opérations est toujours une nouvelle table. La figure 6 montre que la question:

Quels animaux sont des mammifères?

peut être exprimée en algèbre par la requête suivante:

((**EO** `JOIN` **OC**)
`SELECT` **Classe**='Mammifères')
`PROJECT` **Espèce**

D'abord l'expression (**EO** `JOIN` **OC**) résulte en une table qui donne l'ordre et la classe de toute espèce dans la BD. Puis la sélection ne retient que les rangées qui portent sur les mammifères. Finalement, la projection ne retient que la colonne **Espèce**.

On peut maintenant vérifier que la requête:

((**EO** `JOIN` **VIT**)
`SELECT` **Ordre**='Carnivores' &
Continent='Asie')
`PROJECT` **Espèce**

donne tous les carnivores asiatiques. Cette requête est nettement plus simple que le programme équivalent pour la BD de type réseau montré par la figure 4. Notons que toute requête en algèbre relationnelle sera traduite en un programme efficace qui peut être exécuté par l'ordinateur. Mais contrairement aux BD de type réseau, ce programme reste caché aux utilisateurs; la traduction est effectuée automatiquement par le Système de Gestion de Bases de Données (SGBD). Pour cette raison, on dit que les

requêtes en modèle relationnel sont "assertionnelles": l'utilisateur définit les caractéristiques qui s'imposent au résultat. Le SGBD doit alors construire la stratégie de recherche.

Comme mentionné ci-dessus, au début des années 70, on considérait comme une curiosité intellectuelle l'idée de stocker les données dans des tables et d'interroger les tables de manière non-navigationnelle. Il faut comprendre que cette idée était révolutionnaire dans un temps où on était loin des interfaces conviviales pour interagir avec l'ordinateur. Ce scepticisme n'a cependant pas empêché E.F. Codd de poursuivre ses idées. Un premier prototype de Système de Gestion de Bases de Données Relationnelles (SGBDR) a été construit dans les laboratoires d'IBM. Depuis les années 80, cette technologie a mûri et a été adoptée par l'industrie. En 1987, le langage SQL, qui étend l'algèbre relationnelle, a été standardisé. A l'heure actuelle, les SGBDR sont présents dans toutes les compagnies et représentent une industrie de plusieurs milliards de dollars.

E.F. Codd a reçu le Prix Turing en 1981.

5 Le Web, une BD?

5.1 Un Manque de Structure

Aujourd'hui, une immense quantité de données se trouve sur le Web. Néanmoins, il n'est guère possible de parler d'une vraie BD parce que, d'une part, ces données sont peu ou pas structurées et, d'autre part, il n'existe pas de langage pour interroger le Web.

Le Web manque de structure. Il est construit à partir de "pages" écrites en langage HTML (*HyperText Markup Language*). En gros, ce langage permet (i) de spécifier à l'aide de balises comment une page doit être présentée sur l'écran de l'ordinateur et (ii) d'ajouter des liens vers d'autres pages. La figure 7 donne un exemple: le titre se trouve entre les balises `<H1>` et `</H1>` (*Header*); les balises `<OL>` et `</OL>` (*Ordered List*) délimitent le début et la fin d'une liste ordonnée; chaque article de la liste se trouve entre les balises `<LI>` et `</LI>` (*List Item*).

Il y a deux manières de chercher des informations sur le Web:

- Utiliser des moteurs de recherche tels que Google, Hotbot et Alta Vista. Ces moteurs sont comme l'index d'un livre: on saisit un mot clé et le moteur retourne toutes les pages contenant ce mot. Malheureusement, cette méthode de recherche manque de précision. Par exemple, un biologiste qui s'intéresse à la symbiose entre les pandas et les mouches peut demander toutes les pages contenant à la fois les mots "panda" et "mouche". Il ne lui sera pas possible d'éviter des pages non pertinentes telles que celle montrée par la figure 7.

  Une petite expérience: le lundi 5 mars 2001, le moteur de recherche Google (`http://www.google.com/`) trouvait 168 pages rédigées en français contenant les mots "panda" et "mouche". La page classée en tête parle des "Gîtes Panda au Parc naturel régional Normandie-Maine" où on sait "pêcher la truite fario à la mouche"...

- Naviguer de site en site, ce qui fait penser à la navigation dans les BD de type réseau. Il y a pourtant une différence importante: contrairement au Web, les BD de type réseau se conforment à une structure précise.

5.2 Traiter le Futur Web comme BD

Le défi est de mieux structurer et décrire le contenu des pages Web. Supposons que tous les biologistes du monde se mettent d'accord pour utiliser des balises standardisées de manière à décrire les animaux dans leurs pages Web. Voici

EO `JOIN` **OC**

Espèce	**Ordre**	**Classe**
Lion	Carnivores	Mammifères
Loup	Carnivores	Mammifères
Tigre	Carnivores	Mammifères
Hyène	Carnivores	Mammifères
Girafe	Artiodactyles	Mammifères
Hippopotame	Artiodactyles	Mammifères
Cobra	Serpents	Reptiles

(**EO** `JOIN` **OC**) `SELECT` **Classe**=Mammifères

Espèce	**Ordre**	**Classe**
Lion	Carnivores	Mammifères
Loup	Carnivores	Mammifères
Tigre	Carnivores	Mammifères
Hyène	Carnivores	Mammifères
Girafe	Artiodactyles	Mammifères
Hippopotame	Artiodactyles	Mammifères

((**EO** `JOIN` **OC**) `SELECT` **Classe**=Mammifères) `PROJECT` **Espèce**

Espèce
Lion
Loup
Tigre
Hyène
Girafe
Hippopotame

Figure 6: Pour répondre à la question *"Quels animaux sont des mammifères?"* on joint (`JOIN`) d'abord les tables **EO** et **OC** pour ensuite en retenir (`SELECT`) les mammifères. Finalement, on ne retient (`PROJECT`) que la colonne **Espèce**.

```
<H1> &Eacute;nigmes </H1> <OL>
<LI> Comment faire entrer quatre
&eacute;l&eacute;phants dans une
fiat panda?  </LI> <LI> Comment
un &eacute;l&eacute;phant se
mouche-t-il?  </LI> </OL>
```

Énigmes

1. Comment faire entrer quatre éléphants dans une fiat panda?
2. Comment un éléphant se mouche-t-il?

Figure 7: Les balises dans une page HTML (gauche) sont interprétées par le navigateur qui affiche la page (droite).

un exemple:

```
<ANIMAL>
  <ESPECE> Panda </ESPECE>
  <CLASS> Mammifères </CLASS>
  <NOURRITURE> bambou </NOURRITURE>
  <CONTINENT> Asie </CONTINENT>
</ANIMAL>
```

Par contre, le secteur automobile peut utiliser d'autres balises pour décrire les voitures:

```
<VOITURE>
  <MARQUE> Fiat </MARQUE>
  <MODELE> Panda </MODELE>
  <VITESSE en 'KMPH'> 140 </VITESSE>
</VOITURE>
```

Ces balises standardisées permettraient aux moteurs de recherche de faire la distinction entre un panda et une Fiat Panda. Un tel standard permettrait la construction de requêtes plus précises que celles s'appuyant uniquement sur des mots clés. Par exemple, la question:

Que mange le panda?

pourrait se traduire en une requête à un moteur de recherche:

```
<ANIMAL>
    <ESPECE> Panda </ESPECE>
    <NOURRITURE> ?  </NOURRITURE>
</ANIMAL>
```

Le moteur de recherche enverrait une liste contenant "bambou".

Est-ce que cette technologie est de la science-fiction? Pas du tout: ce nouveau langage pour structurer le Web existe déjà; il s'appelle XML (*eXtensible Markup Language*)! Quant aux recherches, plusieurs prototypes de langages pour interroger des pages XML ont déjà été proposés. Cette technologie permettra dans un futur proche d'interroger le Web de manière précise et non-navigationnelle, un peu comme les BD relationnelles. Il est fort possible que les recherches dans ce domaine aboutissent à un Prix Turing d'ici 10 ans. A suivre donc...

Références

[1] C. W. Bachman. The programmer as navigator. *Communications of the ACM*, 16(11):635–658, 1973.

[2] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.