Bases de données

7 juillet 2016

1 Introduction

On s'intéresse au problème de la gestion d'un gros volumes de données, ainsi qu'à celui de la recherche sur ces données.

1.1 Exemple : IMDB

C'est une base de données sur les films, les réalisateurs et les acteurs, qui est accessible en ligne (`http://www.imdb.com/stats`). Elle a environ 6×10^7 visiteurs par mois (`http://fr.wikipedia.org/wiki/IMDB`).

Quelques données chiffrées : cette base de donnée répertorie

— $\approx 3,7 \times 10^6$ titres ;

— $\approx 7,1 \times 10^6$ personnes (dont $\approx 1,8 \times 10^6$ acteurs).

(source : `http://www.imdb.com/stats`, 2016-03-07)

Au regard des standards actuels IMDB est

— une base de données de taille moyenne ;

— avec un nombre de consultations moyen.

1.2 Et si on faisait notre IMDB ?

C'est un peu ambitieux, on va simplifier un peu.

On commence par se poser les questions suivantes.

— Comment modéliser ce problème ?

— Comment représenter les données à stocker ?

— *Comment faire une interface web ?*

2 Modèle conceptuel des données

On a besoin de modéliser le problème qu'on aborde *avant* de commencer à programmer. C'est une tâche très difficile. On peut retenir deux points fondamentaux.

— Cela nécessite une collaboration entre des spécialistes du domaine et des informaticiens [1].

— En cas d'hésitation ou d'ambiguïté, les informaticiens doivent **refuser** de choisir.

2.1 Le modèle entité-association

De l'anglais *Entity-Relationship model*.
C'est

— une façon de modéliser les données à traiter ;

— un modèle *conceptuel* de données (MCD).

Le mot *conceptuel* doit s'entendre par opposition à l'*implantation* concrète de la base de données (qui donne le modèle *physique* de données, MPD).
On l'appelle ainsi car il distingue :

— les entités (objets d'intérêt) ;

— des associations (liens) entre ces entités.

2.2 Entités

Ici, dans notre exemple de base de donnée cinématographique, nous choisissons comme entités :

(i) les films ;

(ii) les personnes (acteurs, réalisateurs, scénaristes, etc.).

Remarque 2.2.1. On pourrait rajouter : les entreprises (producteurs), les livres (dont sont tirés certains scénarios), les pays (où ont eu lieu le tournage, du producteur, où sont distribués les films), les langues (des films), les versions d'un même film (langues, montages), *etc*.

Pour chaque entité, on considère les données suivantes.

Pour les films Titre, date de sortie ;

Pour les personnes Nom, prénom, date de naissance.

Remarque 2.2.2. Là encore, c'est un choix, on aurait pu allonger cette liste.

1. En général, cela demande une collaboration entre informacitiens de spécialités différentes.

2.3 Associations

Ici, on ne considère dans notre exemple que deux types de relation :

(i) «joue dans» (Clint Eastwood joue Blondin dans *Le bon, la brute et le truand* et Walt Kowalski dans *Gran Torino*) ;

(ii) «a réalisé» (Clint Eastwood a réalisé *Invictus* et *Gran Torino*).

Quelques informations pertinentes sont à prendre en compte sur ces relations.

— Tout film a été réalisé par une personne.

— Tout film a été réalisé par au plus une personne (ce n'est pas forcément vrai pour tout film !).

— Toute personne peut avoir réalisé 0, 1 ou plusieurs films.

— Toute personne peut avoir joué dans 0, 1 ou plusieurs films.

— Tout film a 0, 1 ou plusieurs acteurs.

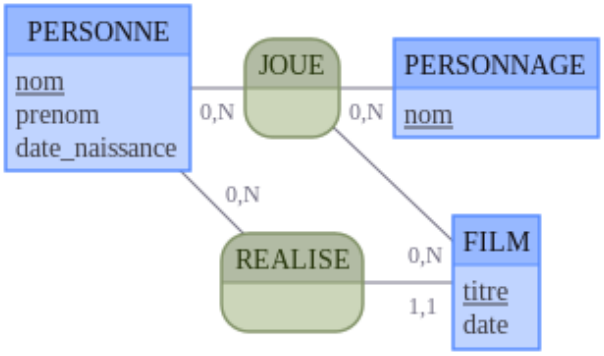
— Lorsqu'une personne joue dans un film, il est intéressant de savoir quel est son rôle (elle peut alors jouer un ou plusieurs rôles).

Clairement, nous sommes ici dans une étape de modélisation, qui possède donc des limites.

On remarquera que nous pourrions avoir intérêt à introduire une entité supplémentaire : les personnages. Pour chaque personnage, on ne considèrera que son nom.

2.4 Diagramme entité/association

On représente la modélisation précédente du lien entre les entités et les associations de notre projet de base de données cinématographique par le diagramme suivant.



On porte sur le diagramme des indications pour préciser comment fonctionnent les relations.

Considérons la relation R des (p, f) où :

— p est une personne ;

— f est un film ;

— p a réalisé f.

Remarque 2.4.1. La relation R modélise mathématiquement l'association «REALISE»

Dans R, par notre modélisation, on a les propriétés suivantes.

— Un même film apparaît au moins une fois et au plus une fois (tout film a un unique réalisateur), d'où le «$1, 1$» sur le trait reliant «REALISE» à «FILM» sur le diagramme.

— Une même personne peut apparaître 0, 1 ou plusieurs fois, d'où le «$0, N$» sur le trait reliant «REALISE» à «PERSONNE».

On utilise le même principe pour renseigner le diagramme de la relation «JOUE».

3 Modèle logique des données

3.1 Tables

On représente ces entités et ces associations par des tables.
Par exemple, voici une table pour les personnes.

nom	prénom	date_naissance
Kubrick	Stanley	1928
Spielberg	Steven	1946
Eastwood	Clint	1930
Cumberbatch	Benedict	1976
Freeman	Martin	1971
Leone	Sergio	1929
McGuigan	Paul	1963
Sellers	Peter	1925

Une pour les films.

titre	date
Gran Torino	2008
The good, the Bad and the Ugly	1966
Study in Pink	2010
Schindler's List	1993
Dr Strangelove	1964
Invictus	2009

Une pour les personnages.

nom
Walt Kowalski
Blondie
Shelock Holmes
Dr John Watson
Dr Strangelove
Group Capt. Lionel Mandrake
President Merkin Muffley

Une pour l'association «JOUE».

nom	prenom	titre	nom (de personnage)
Eastwood	Clint	The good, the Bad and the Ugly	Blondie
Eastwood	Clint	Gran Torino	Walt Kowalski
Cumberbatch	Benedict	Study in Pink	Sherlock Holmes
Freeman	Martin	Study in Pink	Dr John Watson
Selers	Peters	Dr Strangelove	Dr Strangelove
Selers	Peters	Dr Strangelove	Group Capt. Lionel Mandrake
Selers	Peters	Dr Strangelove	President Merkin Muffley

Une pour l'association «REALISE».

titre	nom (réalisateur)	prénom (réalisateur)
Gran Torino	Eastwood	Clint
The good, the Bad and the Ugly	Leone	Sergio
Study in Pink	McGuigan	Paul
Schindler's List	Spielberg	Steven
Dr Strangelove	Kubrick	Stanley
Invictus	Eastwood	Clint

3.2 Vers une implantation ?

On peut considérer ces tables comme des ensembles de n-uplets ($n = 3$ pour la table «PERSONNE», $n = 2$ pour «FILMS», $n = 4$ pour «JOUE», $n = 3$ pour «REALISE»).

Ces ensembles de n-uplets peuvent être implantés en python par des listes de listes :

```
FILMS = [
 [ 'Gran Torino', 2008 ],
 [ 'The good ...', 1966 ],
 ...
]
...
```

Ce modèle à base de tables, appelé *modèle logique des données* est plus proche de l'implantation que le modèle conceptuel. Il reste cependant à faire quelques choix avant de pouvoir vraiment passer à une implantation.

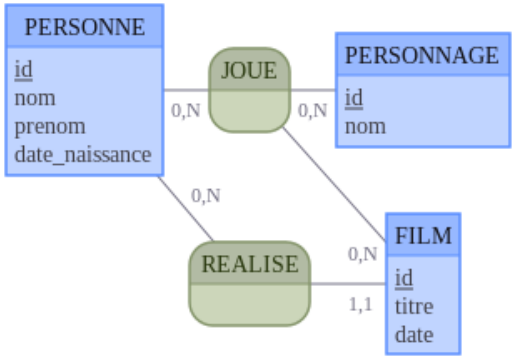
3.3 Une erreur de conception

Dans la table «JOUE», on a décidé de ne mettre que le nom et le prénom de l'acteur car on considère que cela suffit à représenter l'acteur.

C'est en général une très très mauvaise idée : que se passe t-il si deux personnes ont le même nom et prénom ?

Une solution classique (en gestion) : attribuer un numéro unique (de dossier, de personne, ...), *l'identifiant*.

On modifie alors le diagramme entité/association comme suit.



Les tables d'entités sont alors modifiées comme suit.

id	nom	prénom	date_naissance
1	Kubrick	Stanley	1928
2	Spielberg	Steven	1946
3	Eastwood	Clint	1930
4	Cumberbatch	Benedict	1976
5	Freeman	Martin	1971
6	Leone	Sergio	1929
7	McGuigan	Paul	1963
8	Sellers	Peter	1925

id	titre	date
1	Gran Torino	2008
2	The good, the Bad and the Ugly	1966
3	Study in Pink	2010
4	Schindler's List	1993
5	Dr Strangelove	1964
6	Invictus	2009

id	nom
1	Walt Kowalski
2	Blondie
3	Shelock Holmes
4	Dr John Watson
5	Dr Strangelove
6	Group Capt. Lionel Mandrake
7	President Merkin Muffley

L'association «JOUE» est donc modifiée de cette manière.

idacteur	idfilm	idpersonnage
3	2	2
3	1	1
4	3	3
5	3	4
8	5	5
8	5	6
8	5	7

Et voici celle pour l'association «REALISE».

idfilm	idrealisateur
1	3
2	6
3	7
4	2
5	1
6	3

Remarque 3.3.1. Tout film a un réalisateur, ce qui conduit à supprimer la table REALISE et à ajouter un champ réalisateur à la table films.

id	titre	date	idrealisateur
1	Gran Torino	2008	3
2	The good, the Bad and the Ugly	1966	6
3	Study in Pink	2010	7
4	Schindler's List	1993	2
5	Dr Strangelove	1964	1
6	Invictus	2009	3

4 Conclusion

On a vu :

Modèle conceptuel de données utilisation du modèle entité association ;

Modèle logique de données utilisation de tables (modèle relationnel) ;

Passage du MCD au MLD.

Reste à voir comment on implante ce MLD.