Bases de données

Skander Zannad et Judicaël Courant

Lycée La Martinière-Monplaisir

2014-03-13

1 Motivations

Gestion de gros volumes de données.

Recherche sur ces données.

2 Exemple : IMDB

2.1 Qu'est-ce ?

Base de données sur les films, les réalisateurs et les acteurs.

Accessible en ligne (`http://www.imbd.com/stats`).

$\approx 60 \times 10^6$ visiteurs par mois (`http://fr.wikipedia.org/wiki/IMDB`).

2.2 Taille

- $\approx 2,8 \times 10^6$ titres ;
- $\approx 5,8 \times 10^6$ personnes (dont $1,5 \times 10^6$ acteurs) ;

(source : `http://www.imdb.com/stats`, 2014-03-19)

NB : Au regard des standards actuels IMDB est

- une base de données de taille moyenne ;
- avec un nombre de consultations moyen.

3 Et si on faisait notre IMDB ?

(oui, c'est un peu ambitieux, on va simplifier un peu)

Premières questions :

- comment modéliser ce problème ?
- comment représenter les données à stocker ?
- *comment faire une interface web ?*

4 Modèle conceptuel des données

4.1 Généralités

Besoin de modéliser le problème qu'on aborde *avant* de programmer.

Tâche très difficile. Point fondamental à retenir :

- nécessite une collaboration spécialistes du domaine/informaticiens [1] ;
- en cas d'hésitation/d'ambiguïté, les informaticiens doivent **refuser** de choisir.

1. Et en général, besoin de collaborations entre informacitiens de spécialités différentes.

4.2 Le modèle entité-association

(*Entity-Relationship model*)

Modèle Entité-association :

- façon de modéliser les données à traiter ;
- est un modèle *conceptuel* de données (MCD) ;
- *conceptuel* : par opposition à *implantation* (modèle *physique*, MPD).

Appelé ainsi car il distingue :

- Les entités (objets d'intérêt) ;
- Les associations (liens) entre ces entités.

4.3 Entités

Ici :

(i) Les films ;

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

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.

Données à garder sur les entités :

Pour les films Titre, date de sortie ;

Pour les personnes Nom, prénom, date de naissance.

4.4 Associations

Ici :

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

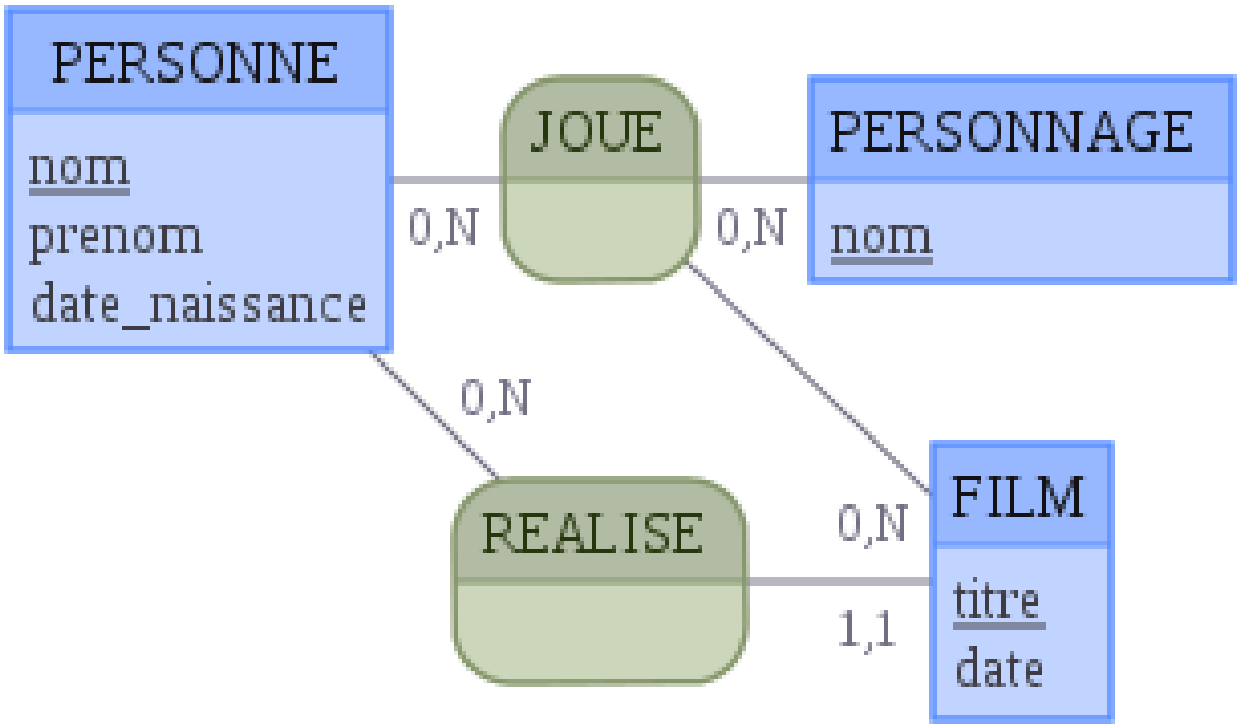
(ii) «a réalisé» (Clint Eastwood a réalisé *Invictus* et *Gran Torino*).

Informations à prendre en compte sur les relations :

- tout film a été réalisé par une personne ;
- tout film a été réalisé par au plus une personne (?) ;
- 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).

4.5 Diagramme entité/association

Utilisé pour représenter cette modélisation :



4.6 Cardinalités

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

Considérons l'ensemble R des (p, f) où :

- p est une personne
- f est un film
- p a réalisé f.

(R : modélise mathématiquement l'association «REALISE»)

Dans R :

- 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».

Même principe pour la relation «JOUE».

5 Modèle logique des données

5.1 Tables

Comment représenter ces entités et associations ?

On passe par la notion de table. On peut représenter les entités par des tables.

Par exemple, 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

5.2 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 ],
 ...
]
...
```

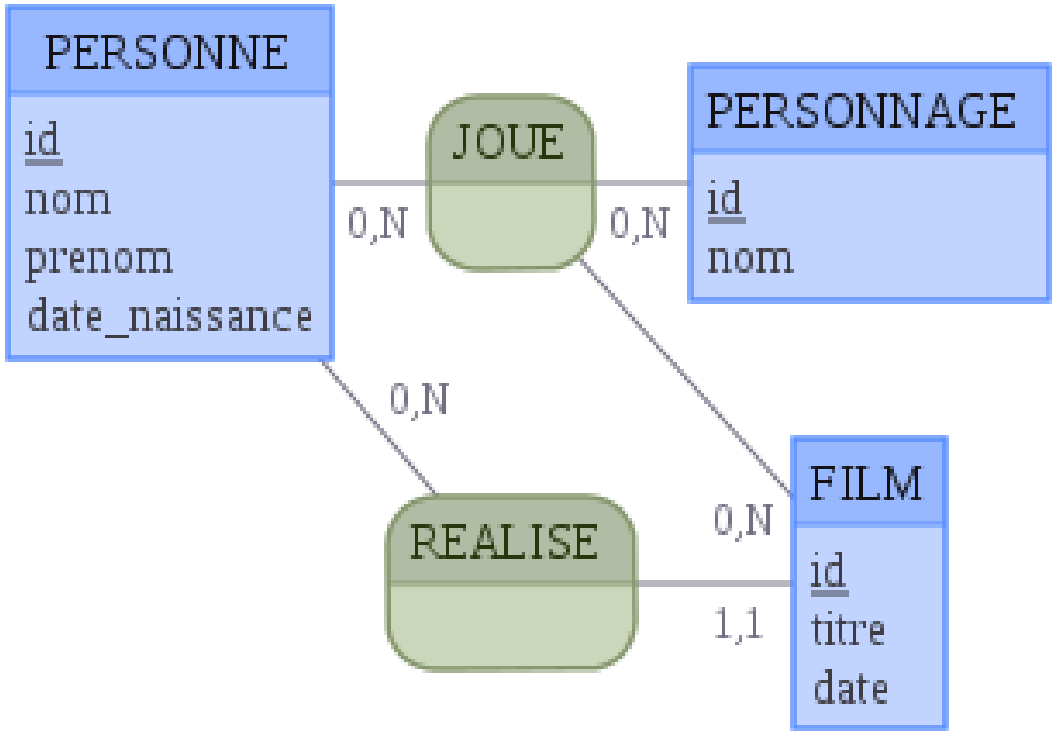
Ce modèle à base de tables, appelé *modèle logique des données* est plus proche de l'implantation que le modèle conceptuel. Il reste cependant à faire quelques choix avant de pouvoir vraiment passer à une implantation.

5.3 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 ?

Solution classique (en gestion) : attribuer un numéro unique (de dossier, de personne, . . .)


PERSONNE
id
nom
prenom
date_naissance
JOUE
PERSONNAGE
id
nom
0,N
0,N
0,N
REALISE
FILM
0,N
1,1
id
titre
date

Ce qui donne les tables :

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» devient alors :

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 celle pour l'association «REALISE» :

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

NB : 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

6 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.