Conception de bases de données

Marc Plantevit

marc.plantevit@liris.cnrs.fr

Objectifs

Nous entendons par conception, dans ce cours, le fait de choisir une modélisation des données réelles à partir du cahier des charges des applications.

- il s'agit donc de déterminer l'ensemble des attributs, des relations et des contraintes qui constitueront le modèle.

Nous allons voir dans la suite :

- Quelles sont les propriété attendues d'une bonne modélisation en relationnel ;
- Comment les obtenir.

Rappels

Modéliser ?

Consiste à définir un monde abstrait qui coïncide avec les manifestations apparentes du monde réel.

Focus sur les données : représenter les données du monde réel dans une forme intelligible

- Support pour la communication lors de la phase de modélisation d'un système
- Support pour automatiser la production de code (schéma de base de données mais aussi interface de classes C++ ou Java)

Modéliser Vs Peindre

Modéliser ± art

Nous n'avons jamais la même perception du monde, rarement le même niveau de granularité dans la description des informations.

Exemple

- Gestion du département Informatique de l'UFR,
- Gestion des capteurs d'une centrale hydraulique, ...

Contexte

La modélisation des données trouve ses racines

- en Intelligence Artificielle, e.g. en représentation des connaissances
- en Bases de données,
- en Génie Logiciel

Pourquoi tant d'effort ?

- Fort intérêt en pratique, bagage de base d'un informaticien.
- Importance des données vis à vis du code $\Rightarrow$ en général, les données sont plus stables dans le temps que les codes qui les accèdent $\Rightarrow$ lié à l'évolution des besoins

Exemple

Soit $\mathcal{U} = \{id, nom, adresse, cnum, desc, note\}$ un univers décrivant des étudiants et des cours. Soient les deux schémas de BD suivants :

- $R1 = \{Donnees\}$ avec $schema(Donnees) = \mathcal{U}$ [1].
- $R2 = \{Etudiant, Cours, Affectation\}$ avec $schema(Etudiant) = \{id, nom, adresse\}$, $schema(Cours) = \{cnum, desc\}$, $schema(Affectation) = \{id, cnum, note\}$

1. NB : données de R1 $\approx$ données d'un tableur

Exemple

Soit $\mathcal{U} = \{id, nom, adresse, cnum, desc, note\}$ un univers décrivant des étudiants et des cours. Soient les deux schémas de BD suivants :

- $R1 = \{Donnees\}$ avec $schema(Donnees) = \mathcal{U}$ [1].
- $R2 = \{Etudiant, Cours, Affectation\}$ avec $schema(Etudiant) = \{id, nom, adresse\}$, $schema(Cours) = \{cnum, desc\}$, $schema(Affectation) = \{id, cnum, note\}$

Comment peut-on évaluer ces deux schémas de données ?

- Lequel est "meilleur" ?
- Pourquoi ?
- Selon quels critères ?

1. NB : données de R1 $\approx$ données d'un tableur

Exemple

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B

Exemple

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B

Quels sont les problèmes ?

- L'information est **redondante**.

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B

- L'information est **redondante**.

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B

- **Anomalie de modification :** Une modification sur une ligne peut nécessiter des modifications sur d'autres lignes.
  Ex. : Modif. Adresse de Paul.

- Certaines informations dépendent de l'existence d'autres informations.
  Ex. : Le cours 'CS24' de BD dépend de l'existence de Paul.
- $\Rightarrow$ **Anomalie de suppression**.

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B

- Valeurs manquantes (pas de valeurs nulles dans le cours).
  Ex : Soit '145, Evariste, Aubenas' un nouvel étudiant. On ne peut l'insérer que si l'on connait un de ses cours et sa note dans ce cours, à moins de permettre les valeurs nulles.
- $\Rightarrow$ **Anomalie d'insertion**.

Données	id	nom	adresse	cnum	desc	note
	124	Jean	Paris	F234	Philo I	A
	456	Emma	Lyon	F234	Philo I	B
	789	Paul	Marseille	M321	Analyse I	C
	124	Jean	Paris	M321	Analyse I	A
	789	Paul	Marseille	CS24	BD I	B
	145	Evariste	Aubenas	**? ? ?**	**? ? ?**	**? ? ?**

Comment formaliser tout ça ?

Le moyen simple qui permet d'éviter ces problèmes est l'étude des contraintes (e.g., les dépendances fonctionnelles, multi-valuées, etc.).

Quelques définitions

`http://en.wikipedia.org/wiki/Database_normalization`

Élémentaire (ou *minimale*) une DF $X \to Y$ est *élémentaire* ssi $\forall . X' \subsetneq X \Rightarrow X' \not\to Y$

Directe une DF $X \to Y$ est *directe* ssi $\nexists Z . X \to Z \wedge Z \not\to X \wedge Z \to Y$.

Clé un ensemble d'attributs X est *clé* ssi $\forall A \in R . X \to A$. On dit aussi que la *dépendance* $X \to R$ est *clé*.

Super clé un ensemble d'attributs X est *super clé* ssi $\exists K . K$ est clé $\wedge\, K \subseteq X$.

Clé candidate (ou *minimale*) un ensemble d'attributs X est *clé candidate* ssi la DF associée $X \to R$ est *élémentaire*.

Clé primaire c'est le *choix d'une clé* parmi les candidates.

Outline

Une anomalie de mise à jour à lieu lorsque, à la suite d'une modification de la base, des contraintes sémantiques valides se trouvent violées. Bien sûr, des mécanismes de contrôle sont intégrés aux SGBDR pour éviter ce genre de problèmes. Mais celà suppose :

- une perte de temps dans la gestion de la base, certains contrôles pouvant être assez lourds ;
- une implémentation rigoureuse de toutes les contraintes par le concepteur de la base. Sous Oracle, celà passe bien souvent par la mise en place de "Trigger" en PL/SQL.

Le compromis est alors atteint en faisant l'hypothèse suivante : **le concepteur n'implémente que les clés et les clés étrangères.** Le contrôle automatique de ces contraintes est peu coûteux par le SGBD, et leur implémentation est toujours intégrée dans les SGBDR. Ainsi, on considère que toute mise à jour respecte les clés.

Anomalie de m.a.j

R a une anomalie de mise à jour par rapport à F s'il est possible d'insérer ou de modifier un tuple t tel que :

- $r \cup t \models CLE(F)$, où $CLE(F)$ est l'ensemble des clés minimales induites par F.
- $r \cup t \not\models F$.

Anomalie de m.a.j

R a une anomalie de mise à jour par rapport à F s'il est possible d'insérer ou de modifier un tuple t tel que :

- $r \cup t \models CLE(F)$, où $CLE(F)$ est l'ensemble des clés minimales induites par F.
- $r \cup t \not\models F$.

Example

Supposons le schéma de relation $ETUDIANT(NUMETUD(A), NOM(B), VILLE(C), CP(D), DPT(E))$ muni de l'ensemble de DF $F = \{A \rightarrow BCD, CD \rightarrow E\}$

- La seule clé minimale de la relation est A (Toutes les autres clés sont des sur-ensembles de A).
- Supposons qu'un tuple soit inséré pour un nouvel étudiant, avec une ville et un CP déjà présent mais un département différent. La clé ne sera pas violée (pas de doublon sur A) mais la DF $CD \rightarrow E$ ne sera plus satisfaite. Puisqu'un tel cas de figure est possible, la relation $ETUDIANT$ possède une anomalie de mise à jour.

Suppression et Perte d'Information

Une suppression ne peut entraîner aucune anomalie proprement dite ; toutefois, elle peut engendrer une perte involontaire d'information, lié à la redondance dans les données.

- Par rapport à $ETUDIANT(NUMETUD(A), NOM(B), VILLE(C), CP(D), DPT(E))$ et $F = \{A \rightarrow BCD, CD \rightarrow E\}$

NUMETUD	NOM	VILLE	CP	DPT
1	Fagin	Lyon	69003	Rhône
2	Armstrong	Lyon	69001	Rhône
3	Bunneman	Clermont	63000	Puy-de-Dôme
4	Codd	Lyon	69001	Rhône

- si on supprime tous les étudiants de LYON ?

Suppression et Perte d'Information

Une suppression ne peut entraîner aucune anomalie proprement dite ; toutefois, elle peut engendrer une perte involontaire d'information, lié à la redondance dans les données.

- Par rapport à $ETUDIANT(NUMETUD(A), NOM(B), VILLE(C), CP(D), DPT(E))$ et $F = \{A \rightarrow BCD, CD \rightarrow E\}$

NUMETUD	NOM	VILLE	CP	DPT
1	Fagin	Lyon	69003	Rhône
2	Armstrong	Lyon	69001	Rhône
3	Bunneman	Clermont	63000	Puy-de-Dôme
4	Codd	Lyon	69001	Rhône

- si on supprime tous les étudiants de LYON ?
  - on perd le code postal et le département de Lyon ;
  - même si on souhaite garder cette information.

Redondances

- La notion de *redondance* est une autre façon de considérer les problèmes de mises à jour.
- Elle se définit sur les relations, alors les problèmes de mise à jour portent sur des schémas.

Definition

Une relation r sur R est redondante par rapport à un ensemble F de DF sur R si :

1. $r \models F$ et
2. il existe $X \rightarrow A \in F$ et $t_1, t_2 \in r$ tels que $t_1[XA] = t_2[XA]$.

Exemple

$ETUDIANT(NUMETUD(A), NOM(B), VILLE(C), CP(D), DPT(E))$ et $F = \{A \rightarrow BCD, CD \rightarrow E\}$:

NUMETUD	NOM	VILLE	CP	DPT
1	Fagin	Lyon	69003	Rhône
2	Armstrong	Lyon	69001	Rhône
3	Bunneman	Clermont	63000	Puys-de-Dôme
4	Codd	Lyon	69001	Rhône

Cette relation est bien correcte car elle respecte les DF. Néanmoins, elle est redondante car il existe un doublon sur $(Ville, CP)$: l'information du département de LYON 1er apparaît deux fois.

Liens entre Anomalies de m.a.j et Redondances

On voit bien que les notions d'anomalie de mise à jour et de redondance sont très liées.

- Elles sont en fait équivalentes, selon le résultat suivant.

Theorem

Il y a équivalence entre :

- *R a une anomalie de mise à jour par rapport à F,*
- *Il existe une relation r sur R qui est redondante par rapport à F.*

Outline

Pour éviter les anomalies :

- Le principe de base est alors de décomposer les relations de telle sorte d'éviter d'avoir ces anomalies ;
- i.e. transformer une relation en plusieurs relations [a].

---

a. NB : Avec des schémas de relation à un seul attribut, il n'y a plus aucun problème de redondance !

Pour éviter les anomalies :

- Le principe de base est alors de décomposer les relations de telle sorte d'éviter d'avoir ces anomalies ;
- i.e. transformer une relation en plusieurs relations [a].

---

a. NB : Avec des schémas de relation à un seul attribut, il n'y a plus aucun problème de redondance !

Risques :

- Le risque en décomposant est de perdre de l'information.

Perte d'Information

- Il faut que toutes les informations de la base de donnée initiale puissent être retrouvées en effectuant des jointures sur les relations issues de la décomposition.
- Soit R un schéma de relation (c'est à dire un ensemble d'attributs), que l'on décompose en un schéma de base de données (un ensemble de relations) $\mathbf{R} = \{R_1, \ldots, R_n\}$.
  - **R** est *sans perte de jointures* par rapport à un ensemble F de dépendances fonctionnelles si, pour toute relation r sur R telle que $r \models F$ on a :

$$r = \pi_{R_1}(r) \bowtie \ldots \bowtie \pi_{R_n}(r)$$

Exemple :

Reprenons la relation de l'exemple précédent. Supposons que pour régler le problème de redondance, on découpe le schéma en deux de façon à obtenir les relations suivantes :

NUMETUD	NOM
1	Fagin
2	Armstrong
3	Bunneman
4	Codd

VILLE	CP	DPT
Lyon	69003	Rhône
Lyon	69001	Rhône
Clermont	63000	Puy-de-Dôme

Peut-on reconstruire r ?

Perte de DF

- Il ne faut pas que la décomposition ait "coupé" des dépendances fonctionnelles, ce qui conduirai à une perte inévitable de sémantique.
- On définit la notion de projection d'un ensemble de dépendances fonctionnelles :

Definition

projection d'un ensemble de DF Soit F un ensemble de DF sur R, et S un schéma de relation tel que $S \subseteq R$.

$$F[S] = \{X \to Y \mid X \to Y \in F \text{ et } XY \subseteq S\}$$

- La projection sur un schéma de bases de données est l'union des projection sur chaque relation du schéma.
- Soit R un schéma de relation et F un ensemble de DF sur R. Un schéma de relation $\mathbf{R}$ est une *décomposition qui préserve les dépendances* de R par rapport à F si :

$$F[\mathbf{R}]^+ = F^+$$

TODO

Soit la relation *Edition* définie sur le schéma
$R = \{isbn, titre, editeur, pays\}$ qui décrit des livres et leurs éditeurs et
$F = \{isbn \rightarrow titre, editeur, pays;\ editeur \rightarrow pays\}$
Soit r :

Edition	ISBN	TITRE	EDITEUR	PAYS
	2-212-09283-0	Bases de données - objet et relationnel	Eyrolles	France
	2-7117-8645-5	Fondements des bases de données	Vuibert	USA
	0-201-70872-8	Databases and Transaction Processing	Addison Wesley	USA
	2-212-09069-2	Internet/Intranet et bases de données	Eyrolles	France

- Exhibez des redondances et des exemples d'anomalies d'insertion, m.a.j., suppression ?
- Est-ce que la décomposition $Livre(ISBN, TITRE, EDITEUR)$ et $Edite(EDITEUR, PAYS)$ préserve des pertes d'information et de DF ?

Solution aux Anomalies

La solution à ces problèmes consiste à **normaliser la relation** en cause en la décomposant en plusieurs relations.

- Cette décomposition s'appuie sur les dépendances qui existent entre les attributs de la relation initiale :
  - dépendances fonctionnelles,
  - dépendances multivaluées.

Formes Normales

Permettent de spécifier formellement la notion intuitive de bons schémas.

- Pour les DFs, plusieurs formes normales (FN) existent : 1FN (moins restrictive) , 2FN, 3FN, FN de Boyce-Codd (plus restrictive).
- Pour les DMVs, on a la 4FN.

Quand ne pas normaliser ?

La normalisation n'est pas une obligation on peut (doit) s'en passer quand le jeu n'en vaut pas la chandelle :

- Pour retrouver ≪ toutes ≫ les données (originales), il faut calculer des jointures, qui peuvent être coûteuses :
  - elle sont généralement nombreuses car la décomposition est maximale,
  - leur calcul n'est pas toujours performant, en particulier si les index ne sont pas adaptés.
- C'est une étape difficile, et donc coûteuse en travail humain surtout pour obtenir des formes normales élevées
- On en a pas nécessairement besoin car la base n'a pas une très grande durée de vie.

Outline

Introduction

- Les formes normales sont des propriétés que doivent vérifier les schémas pour éviter les anomalies de mises à jour.
- Une forme normale s'applique à un schéma de relation, en fonction d'un certain ensemble de contraintes d'une classe donnée.
- Concernant les DF, l'idée générale est de n'avoir que les clés à vérifier, et d'éliminer au maximum des DF qui ne définissent pas des clés.
- Dans la suite, soit R un schéma de relations et F un ensemble de DF définies sur R.

1FN

Rappel : en relationnel, on est toujours en première forme normale, soit toutes les valeurs des attributs sont atomiques.

Exemple

- Pas en 1FN :

ISBN	Auteurs
2-212-09283-0	Gardarin
2-7117-8645-5	Abiteboul, Hull, Vianu

1FN

Rappel : en relationnel, on est toujours en première forme normale, soit toutes les valeurs des attributs sont atomiques.

Exemple

- Pas en 1FN :

ISBN	Auteurs
2-212-09283-0	Gardarin
2-7117-8645-5	Abiteboul, Hull, Vianu

⇒ En 1FN :

ISBN	Auteurs
2-212-09283-0	Gardarin
2-7117-8645-5	Abiteboul
2-7117-8645-5	Hull
2-7117-8645-5	Vianu

2FN

Definition

2FN R est en 2FN par rapport à F si, pour chaque DF $X \rightarrow A$ de F, l'une des deux conditions suivantes est remplies :

- A appartient à une clé de R,
- X n'est pas un sous-ensemble propre d'une clé de R

Une relation est en 2e forme normale si elle est en 1FN et si chaque attribut non clé dépend totalement et non partiellement de la clé primaire.

Donc, aucun attribut (en dehors de ceux qui forment les clés), ne sont déterminés par des sous-ensembles des clés.

Exemple

$$F = \{isbn, bib \rightarrow nb\_ex; isbn \rightarrow titre\}$$

- TODO : 2FN respectée ?

Exemple

$$F = \{isbn, bib \rightarrow nb\_ex; isbn \rightarrow titre\}$$

- TODO : 2FN respectée ?

ISBN	Titre	Bib	Nb_Ex
2-212-09283-0	BD ...	UCBL1	5
2-7117-8645-5	Fondements des BD	UCBL1	10
2-212-09283-0	BD ...	INSA-L	3

décomposée en

ISBN	Bib	Nb_Ex
2-212-09283-0	UCBL1	5
2-7117-8645-5	UCBL1	10
2-212-09283-0	INSA-L	3

ISBN	Titre
2-212-09283-0	BD ...
2-7117-8645-5	Fondements des BD

- **TODO :** vérifiez que cette décomposition est sans perte et préserve les dépendances fonctionnelles.

Insuffisance de la 2FN

$F = \{isbn \rightarrow titre, editeur \ ; \ editeur \rightarrow pays\}$ En 2FN

ISBN	Titre	Editeur	Pays
2-212-09283-0	BD ...	Eyrolles	France
2-7117-8645-5	Fondements ...	Vuibert	USA
2-212-0969-2	Internet/Intranet ...	Eyrolles	France

Insuffisance de la 2FN

$F = \{isbn \rightarrow titre, editeur \ ; \ editeur \rightarrow pays\}$ En 2FN

ISBN	Titre	Editeur	Pays
2-212-09283-0	BD ...	Eyrolles	France
2-7117-8645-5	Fondements ...	Vuibert	USA
2-212-0969-2	Internet/Intranet ...	Eyrolles	France

Il reste des redondances, notamment avec la df Editeur $\rightarrow$ Pays

3FN

Definition

3FN R est en troisième forme normale par rapport à F si, pour chaque DF $X \rightarrow A$ de F, l'une des deux conditions suivantes est remplies :

- X est une clé (ou une superclé [a]),
- A appartient à une clé de R.

a. sur-ensemble d'une clé.

- La 2e partie de la règle est importante car elle dit qu'un constituant d'une clé *candidate* peut dépendre :
  - soit d'un constituant d'une clé candidate ,
  - soit d'un constituant non clé.
- **Ce qui peut être source de redondances.**

Exemple

$$isbn, bib \rightarrow nb\_ex \ ; \ isbn \rightarrow titre$$

ISBN	Titre	Bib	Nb_Ex
2-212-09283-0	BD ...	UCBL1	5
2-7117-8645-5	Fondements des BD	UCBL1	10
2-212-09283-0	BD ...	INSA-L	3

Pas en 3NF (ni 2NF) et décomposée en

ISBN	Bib	Nb_Ex
2-212-09283-0	UCBL1	5
2-7117-8645-5	UCBL1	10
2-212-09283-0	INSA-L	3

$$isbn, bib \rightarrow nb\_ex$$

ISBN	Titre
2-212-09283-0	BD ...
2-7117-8645-5	Fondements des BD

$$isbn \rightarrow titre$$

- **Cette décomposition est sans perte et préserve les dépendances fonctionnelles.**

Exemple (2)

$F = \{isbn \rightarrow titre, editeur \ ; \ editeur \rightarrow pays\}$

En 2FN, pas en 3FN

ISBN	Titre	Editeur	Pays
2-212-09283-0	BD ...	Eyrolles	France
2-7117-8645-5	Fondements ...	Vuibert	USA
2-212-0969-2	Internet/Intranet ...	Eyrolles	France

décomposée en

ISBN	Titre	Editeur
2-212-09283-0	BD ...	Eyrolles
2-7117-8645-5	Fondements ...	Vuibert
2-212-0969-2	Internet/Intranet ...	Eyrolles

Editeur	Pays
Eyrolles	France
Vuibert	USA

$isbn \rightarrow titre, editeur$

$editeur \rightarrow pays$

- **Cette décomposition est sans perte et préserve les dépendances fonctionnelles.**

Insuffisance de la 3NF

$rue, ville \rightarrow cp; cp \rightarrow ville$

en 3FN car :
{rue, ville} est une super clé ;
ville appartient à une clé candidate.

clés candidates :
{rue, ville}
{rue, cp}

Rue	CP	Ville
Rue J. Capelle	69100	Villeurbanne
Rue de la Doua	69100	Villeurbanne
Rue de la République	69001	Lyon
Rue de Baleine	69001	Lyon

Il reste des redondances : $cp \rightarrow ville \Rightarrow$ **(...,69100,Villeurbanne)**

FNBC

La forme normale de Boyce-Codd impose que toutes les parties gauches des DF sont des clés.

Definition

FNBC R est en forme normale de Boyce-Codd par rapport à F si, pour chaque DF $X \rightarrow A$ de F, X est une superclé de R.

Remarque : La forme normale de Boyce-Codd implique la 3ème forme normale.

Une forme idéale

La FNBC est, en ce qui concerne les DF, la forme idéale d'un schéma de bases de données. En effet, les trois propriétés suivantes sont équivalentes :

- R est en FNBC par rapport à F ;
- R n'a pas de problème de redondances par rapport à F ;
- R n'a pas de problème de mise à jour par rapport à F ;

- La forme normale de Boyce Codd est la forme idéale relativement aux dépendances fonctionnelles,
- La 3FN est toujours possible, quelque soit les DF considérées.
- La 2FN n'a donc qu'un intérêt "historique".
- En revanche, il existe des situations où la FNBC n'est pas possible (peut ne pas préserver les dépendances fonctionnelles).

Exemple

$$rue, ville \rightarrow cp; cp \rightarrow ville$$

Rue	CP	Ville
Rue J. Capelle	69100	Villeurbanne
Rue de la Doua	69100	Villeurbanne
Rue de la République	69001	Lyon
Rue de Baleine	69001	Lyon

CP	Ville
69100	Villeurbanne
69100	Villeurbanne
69001	Lyon
69001	Lyon

Rue	CP
Rue J. Capelle	69100
Rue de la Doua	69100
Rue de la République	69001
Rue de Baleine	69001

Cette décomposition est sans perte **mais ne préserve pas les dépendances fonctionnelles**.

Décomposition en 3FN et en FNBC

Il est démontré que :

- Toute relation a au moins une décomposition en 3FN qui préserve les dépendances fonctionnelles et qui est sans perte.
- Toute relation a au moins une décomposition en FNBC qui est sans perte mais qui peut ne pas préserver les dépendances fonctionnelles.

Que faire pour les pertes de DF en FNBC ?

$$rue, ville \rightarrow cp; cp \rightarrow ville$$

Quelle que soit la décomposition sans perte réalisée, elle n'est pas en FNBC (essayer toutes les décompositions imaginables à partir de ces trois attributs)
Dans un tel cas, il faut soit :

- Accepter de ne pas être en FNBC, et donc accepter d'avoir une anomalie de mise à jour.
  - Se contenter de déclarer les clés.
  - Mais il faudra implémenter les DF (Trigger sous un SGBD, ou au niveau de l'application) qui ne sont pas des clés.
- Accepter de relâcher des contraintes, c'est à dire enlever quelques DF et rendre l'application "plus souple".

Création d'attributs

- On peut aussi rajouter des attributs et des DF. Dans l'exemple, on peut rajouter un attribut qui "regroupe" les couples Rue/Ville. On obtiendrait :

$$(Rue, Ville, Rue/Ville, CP)$$

ainsi que les DF

$$(Rue, Ville \rightarrow Rue/Ville)$$

$$(Rue/Ville \rightarrow Rue, Ville, CP)$$

et

$$CP \rightarrow Ville$$

La décomposition en FNBC est alors possible, et aucune sémantique n'est perdue.

Outline

- Les dépendances fonctionnelles nous ont permis jusque-là de mettre en évidence une forme de redondance, que les formes normales cherchent à faire disparaître.
- Mais des cas de redondance ne sont pas capturés par les DF.

Supposons l'énoncé suivant : "les étudiants suivent des parcours, et sont inscrits dans des transversales indépendantes du parcours. Chaque étudiant peut être inscrit à plusieurs parcours". Soit la modélisation suivante :

$$R(\mathit{etudiants}, \mathit{parcours}, \mathit{transversale})$$

On ne peut dégager aucune DF dans ce schéma, donc la seule clé est la combinaison des trois attributs.

$\Rightarrow$ redondance

Dépendance multivaluée

Il y a une *dépendance multivaluée* entre un constituant X et un constituant Y d'une relation $R(X, Y, Z)$ si :

- pour toute extension de R, à chaque valeur de X il correspond toujours le même ensemble de valeurs de Y et que cet ensemble de valeurs ne dépend pas des valeurs de Z.

On dit que X **multidétermine** Y et l'on note : $X->->Y$.

Exemple

Soit par exemple la relation : *livre(isbn, titre, auteur)* :

- Si un livre peut avoir plusieurs auteurs, la relation livre possède la dépendance multivaluée :
  - $isbn->->auteur$
- Soit un livre d'isbn i et d'auteurs a_1, a_2 et a_3. Si le triplet (i, t, a_1) apparaît dans une extension de la relation livre, alors les triplets (i, t, a_2) et (i, t, a_3) doivent y apparaître aussi.

Une dépendance multivaluée $X->->Y$ d'une relation R est dite **non triviale** si :

- Y n'est pas un sous-ensemble de X,
- $X \cup Y$ n'inclut pas tous les attributs de R.

Une dépendance fonctionnelle est un cas particulier de dépendance multivaluée.

4ème Forme Normale

Définition 1

Une relation R est en 4FN, si pour chaque dépendance multivaluée $X->->Y$ non triviale, X est une super-clé de R.

Définition alternative

- R est en troisième forme normale.
- Les seules dépendances multivaluées sont du type $X->->R-X$. Ou encore, R ne doit pas être décomposable en deux relations sans perte de jointure.

Remarque : La 4FN implique la FNBC puisqu'une dépendance fonctionnelle est un cas particulier de dépendance multivaluée.

Exemple

$$isbn- > - > auteur;\ isbn- > - > mot\_cle$$

ISBN	AUTEUR	MOT CLE
2-7117-8645-5	Abiteboul	BD
2-7117-8645-5	Hull	BD
2-7117-8645-5	Vianu	BD
2-7117-8645-5	Abiteboul	Relationnel
2-7117-8645-5	Hull	Relationnel
2-7117-8645-5	Vianu	Relationnel

Pas en 4FN

Exemple

$$isbn- > - > auteur;\ isbn- > - > mot\_cle$$

ISBN	AUTEUR	MOT CLE
2-7117-8645-5	Abiteboul	BD
2-7117-8645-5	Hull	BD
2-7117-8645-5	Vianu	BD
2-7117-8645-5	Abiteboul	Relationnel
2-7117-8645-5	Hull	Relationnel
2-7117-8645-5	Vianu	Relationnel

Pas en 4FN

ISBN	AUTEUR
2-7117-8645-5	Abiteboul
2-7117-8645-5	Hull
2-7117-8645-5	Vianu

ISBN	MOT CLE
2-7117-8645-5	BD
2-7117-8645-5	Relationnel

Cette décomposition est sans perte et préserve les dépendances fonctionnelles.

Fin de la sixième séance.

Inclusion des formes normales

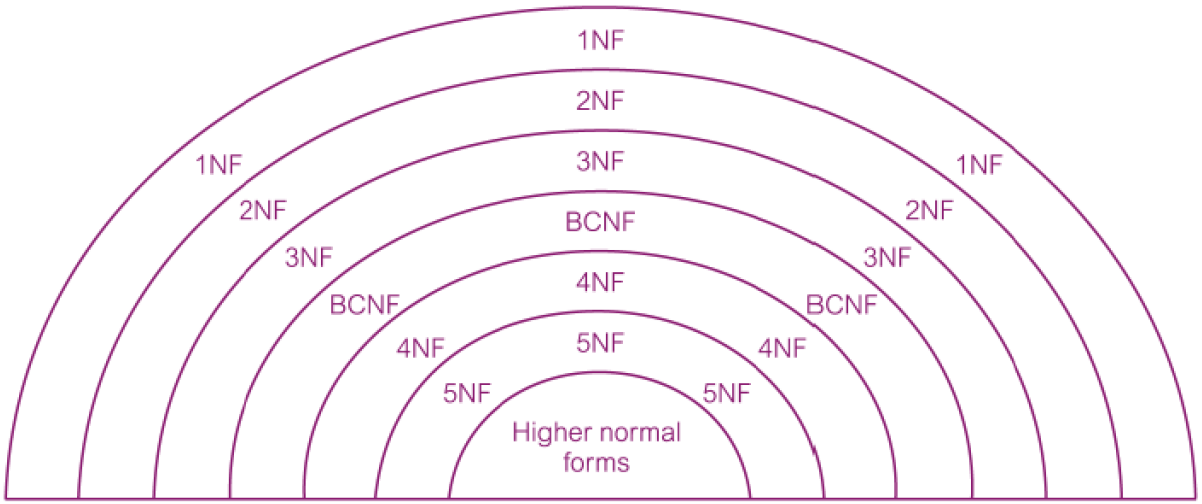

Figure 13.7
Diagrammatic illustration of the relationship between the normal forms.

Dépendance multivaluée triviale

Une dépendance multivaluée $X \twoheadrightarrow Y$ d'une relation R est dite triviale si :

- soit $Y \subseteq X$: comme pour les DFs,
- soit $X \cup Y = R$: la MVD n'impose l'existence d'aucun tuple.

Propriétés des dépendances multivaluées

- les MVDs triviales sont toujours satisfaites.
- la définition de la satisfaction étant symétrique,
  si $r \models X \twoheadrightarrow Y$ alors $r \models X \twoheadrightarrow R \setminus XY$ également.
- si $r \models X \rightarrow Y$ alors $r \models X \twoheadrightarrow Y$ également
  toutes les DFs sont des MVD, mais la réciproque est fausse.

Axiomatisation des dépendances multivaluées

- Réflexivité

$$\frac{Y \subseteq X}{X \twoheadrightarrow Y}$$

- Augmentation

$$\frac{X \twoheadrightarrow Y}{WX \twoheadrightarrow WY}$$

- Transitivité

$$\frac{X \twoheadrightarrow Y \qquad Y \twoheadrightarrow Z}{X \twoheadrightarrow Z}$$

- Complémentation

$$\frac{X \twoheadrightarrow Y}{X \twoheadrightarrow R \setminus XY}$$

- Union

$$\frac{X \twoheadrightarrow Y \qquad X \twoheadrightarrow Z}{X \twoheadrightarrow YZ}$$

- Décomposition

$$\frac{X \twoheadrightarrow Y \qquad X \twoheadrightarrow Z}{X \twoheadrightarrow Y \cap Z, X \twoheadrightarrow Y \setminus Z, X \twoheadrightarrow Z \setminus Y}$$

- Sous-ensemble

$$\frac{X \twoheadrightarrow Y \qquad W \twoheadrightarrow Z \qquad Y \cap W = \emptyset}{X \twoheadrightarrow Y \cap Z, X \twoheadrightarrow Y \setminus Z}$$

Réflexivité, complémentation, augmentation et transitivité forment un système correct et complet pour l'inférence des MVD

Axiomatisation MVD et DF ensembles

- Généralisation

$$\frac{X \to Y}{X \twoheadrightarrow Y}$$

- Pseudo-transitivité mixée

$$\frac{X \twoheadrightarrow Y, Z \subseteq Y \qquad W \cap Y = \emptyset, W \to Z}{X \to Z}$$

Ajoutés à réflexivité, complémentation, augmentation et transitivité, ces règles forment un système correct et complet pour l'inférence des MVD et de DFs prises ensembles.

Outline

Un algorithme de normalisation :

Entrée : Une relation universelle (l'ensemble de tous les attributs du problème) et un ensemble de contraintes : DF, DI, DMV.

Sortie : Construit un schéma de BD normalisé, en 3FN ou en BCNF.

Un algorithme de normalisation :

Entrée : Une relation universelle (l'ensemble de tous les attributs du problème) et un ensemble de contraintes : DF, DI, DMV.

Sortie : Construit un schéma de BD normalisé, en 3FN ou en BCNF.

Deux grandes catégories d'algorithmes de normalisation :

- les algorithmes *de décomposition* ;
- les algorithmes *de synthèse* ;

Dans les deux cas on va s'appuyer sur un ensemble canonique de dépendances et utiliser le théorème de décomposition.

Quelques propriétés

Rappel

$R(XYZ)$ est décomposable sans perte d'information sur XY et XZ si

$$R = \Pi_{XY}(R) \bowtie \Pi_{XZ}(R)$$

Quelques propriétés

Rappel

$R(XYZ)$ est décomposable sans perte d'information sur XY et XZ si

$$R = \Pi_{XY}(R) \bowtie \Pi_{XZ}(R)$$

On a des théorèmes :

Theorem

Soit $R(XYZ)$ (X,Y, Z disjoints) une relation. Si R vérifie $X \rightarrow Y$ alors la décomposition $R_1(XY); R_2(XZ)$ est SPI.

Theorem

Soit $R(XYZ)$ (X,Y, Z disjoints) une relation satisfaisant un ensemble F de DF. La décomposition sur $R_1(XY); R_2(X, Z)$ est SPI $\Leftrightarrow$ $F \models X \rightarrow Y$ ou $F \models X \rightarrow Z$.

Calcul d'une couverture canonique

Pour décomposer selon F, on va utiliser un ensemble F' qui soit :

- Couverture de F : $F^+ = F'^+$,
- Minimal : on ne peut pas retirer de DF en préservant toujours la couverture,
  - On a vu un algorithme pour obtenir les deux points précédents.
- Sans attributs redondants, ni à droite ni à gauche,
- Regroupé : il n'y a pas deux DF avec la même partie gauche.

Réduction du nombre d'attribut pour un ensemble de DF

$Min := F$
/* Réduction des parties gauches */
1 **for** $X \to Y \in Min$ **do**
2 $\quad W := X$
3 $\quad$ **for** $A \in X$ **do**
4 $\quad\quad$ **if** $Min \models (W - A) \to X$ **then** $W := W - \{A\}$
5 $\quad Min := (Min - \{X \to Y\}) \cup \{W \to Y\}$
/* Réduction des parties droites */
6 **for** $X \to Y \in Min$ **do**
7 $\quad W := Y$
8 $\quad$ **for** $A \in Y$ **do**
9 $\quad\quad G := (Min - \{X \to Y\}) \cup \{X \to (W - A)\}$
10 $\quad\quad$ **if** $G \models X \to Y$ **then** $W := W - \{A\}$
11 $\quad Min := (Min - \{X \to Y\}) \cup \{X \to W\}$
12 **return** Min

Exemple de réduction

Soit l'ensemble de DFs Σ [a] :

$$\Sigma = AB \rightarrow ABCDF;\ B \rightarrow BCD;\ DE \rightarrow F;\ E \rightarrow D$$

- Réduction des parties gauches :

$$Min = AB \rightarrow ABCDF;\ B \rightarrow BCD;\ E \rightarrow F;\ E \rightarrow D$$

- Réduction des parties droites :

$$Min = AB \rightarrow F;\ B \rightarrow CD;\ E \rightarrow F;\ E \rightarrow D$$

a. Attention, sur cet exemple, Σ n'est pas une couverture minimum.

Algorithme de décomposition en 4FN

Soit U la relation à décomposer et D l'ensemble des dépendances existantes entre les attributs de U :

Principe général :

- $S := \{U\}$
- Tant qu'il existe dans S une relation R qui n'est pas en 4FN :
  - On cherche dans D une dépendance $X->->Y$ telle que $R(X, Y, Z)$ et X n'est pas une clé de R.
  - On ajoute à Y l'ensemble Z' des attributs de Z fonctionnellement déterminés par X, produisant la dépendance $X->->YZ'$.
  - On remplace R dans S par les deux relations $R_1(X, Y \cup Z')$ et $R_2(X, Z \setminus Z')$.

Cet algorithme est sans perte d'information mais pas toujours sans perte de dépendances.

Algorithme de Synthèse

Entrée : Une relation universelle et un ensemble de contraintes.
On commence par répertorier tous les attributs et construire ainsi la relation universelle de notre application. Puis on dresse l'inventaire des contraintes DF, DI, DMV.

Principe général

- Construire une couverture minimum de F, et réduire les parties gauches et droites au maximum.
- Générer une relations XY pour chaque DF $X \rightarrow Y$;
- Générer une relations XY' pour chaque DMV $X->>Y$ avec $F \models Y' \rightarrow Y$;
- On supprime les schémas de relation qui ne sont pas maximaux par inclusion.
- S'il y a perte de jointure, alors on rajoute une relation composée d'une clé de F.

A partir de cette base, de nombreuses variantes incluent des heuristiques pour diminuer la redondance en sortie.
Propriétés : l'algorithme finit toujours, en donnant la meilleure forme normale (FNBC si elle existe, 4FN sinon)

On gère une liste de **projets**. Chaque **projet** a un **responsable**, un **ensemble d'employés**, et utilise certains **produits** en une **quantité** donnée. Pour un **produit** et un **projet**, plusieurs **fournisseurs** à des **cout** différents sont concernés. Un **fournisseur** a plusieurs **adresses**. Finalement, un **employé** a une **date d'embauche** et un **salaire**.

Exemple 1

Les attributs sont donc
(*Projets*, *produits*, *fournisseur*, *adresse*, *cout*, *responsable*, *employés*, *salaire*, *dateEmbauche*).
Les DF sont :

$$Projet, produit \rightarrow quantite;\ \ Projet, fournisseur, produit \rightarrow coût$$

$$projet \rightarrow responsable;\ \ employe \rightarrow salaire, dateEmbauche$$

et on a deux MVD :

$$Fournisseur ->> Location; projet ->> employe, salaire, dateEmbauche$$

Exemple Suite

Au tableau.

TODO

$$U = \{ISBN, TITRE, AUTEUR, KW, EDITEUR, PAYS, BIB, N\_EX\}$$

Les contraintes sont :

$$ISBN \rightarrow TITRE;\ ISBN \rightarrow EDITEUR;\ EDITEUR \rightarrow PAYS$$

$$ISBN, BIB \rightarrow N\_EX$$

$$ISBN- > - > AUTEUR;\ ISBN- > - > KW$$

Normaliser ce schéma.

Bilan

	3FN	FNBC	4FN
Elimination de la redondance due aux DF	Pas toutes	Oui	Oui
Elimination de la redondance due aux DMV	Non	Non	Oui
Préservation des DF	Oui	Pas toujours	Pas toujours
Préservation des DMV	Pas toujours	Pas toujours	Pas toujours

Fin de la septième séance.