BASES DE DONNEES

Eléments de conception et d'exploitation d'une Base de Données

Equipe Bases de Données, Université LYON 1, LIRIS.

LYON 1 - UFR Informatique - Laboratoire LIRIS

12 mars 2021

Intuition

- Concevoir une BD relationnelle, c'est décider des attributs, des relations, des contraintes
- le processus (EA => traduction relationnelle) ne fait pas tout
- De plus, il est important de savoir reconnaître des problèmes
  - De très nombreuses BD existantes sont mal conçues !

Les "fausses" BD relationnelles

- Une relation peut être représentée par un tableau. . .
- Mais l'inverse est faux !
- Un tableau avec des lignes en doublon n'est pas une relation
  - Une relation a **toujours** une clé
    - donc le doublon de tuple est impossible
- Les domaines des attributs sont toujours atomiques.
  - Eventuellement une chaîne de texte libre, juste à des fins de commentaire
  - Mais pas une liste de valeurs

Les "fausses" BD relationnelles : exemple

Étudiants	NUM	NOM	PRENOM	FORMATION	Commentaire
	28	Codd	Edgar, James	L3IF	Etudiant arrivé en cours d'année
	24	Mannila	Heikki	M2TIW, CAPES_INFO	Reprise d'études
	32	Armstrong	William	M1IF	
	53	Fagin	Ronald	M2TIW	
	107	Bunneman	Peter	Data Science, M2TIW	Demande de RV

TABLE – Base de données non relationnelle

- Plusieurs prénoms par personne (attribut multivalué)
- Plusieurs formations par étudiant : plusieurs connections dans le même tuple.
- Les mécanismes ne raisonnement ne marchent plus (Contraintes, requêtes)
- Le champ commentaire est une source de souplesse, pas interdit. . .
  - Mais champ très difficile à interroger, non formaté.

Le mode tableur...

Considérons la BD de la figure ci-dessous sous la forme d'une relation unique (on omet des attributs pour des raisons de place).

Étudiants	NUMETUD	NOMETUD	NUMENS	NOMENS	DATEENCADRE
	28	Codd	5050	Tarjan	2015
	28	Codd	3434	Papadimitriou	2020
	32	Armstrong	2123	Mannila	2019
	53	Fagin	5050	Tarjan	2005
	107	Bunneman	NULL	NULL	NULL

TABLE – Exemple de Base de Donnée redondante

Forte redondance !

- On répète plusieurs fois le nom des étudiants ou enseignants
  - Attentions aux mises à jour !
- Recours nécessaire aux valeurs NULL
- Sémantique obscure... Liste des étudiants ? Des enseignants ? Des encadrements ? Que veut dire cette relation ?

Comment détecter la redondance ?

La redondance est une conséquence du cahier des charges

- Un même étudiant a toujours le même nom. . .
- . . . et on identifie un étudiant avec son numéro.
- Donc pour un numéro donné, on ne peut avoir qu'un seul nom.
- On modélise cette info par une dépendance fonctionnelle (DF)
  - On note NUMETUD $\rightarrow$ NOMETUD (on lit la flèche "détermine")
  - En exprimant cette DF, on "formalise" des redondances éventuelles
  - Exercice : quelles autres DF imagine-t-on dans l'exemple ?

Définition des dépendances fonctionnelles (1)

Définition Syntaxique

- Expression $R : X \rightarrow Y$
- Avec X et Y des sous-ensembles de R

Conséquence dans une relation r

- On dit que r sur R satisfait $X \rightarrow Y$ (noté $r \models X \rightarrow Y$)
- SSI pour chaque valeur de X dans r, on trouve bien une seule valeur de Y.
- $\forall t_1, t_2 \in r, t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y]$
- Remarque : Si $Y \subseteq X$, la DF est triviale, car toujours satisfaite

Une DF **valide** dans un schéma R doit être satisfaite dans toutes les relations définies sur R.

Définition des dépendances fonctionnelles (2)

Les DF englobent la notion de clé

- X est une clé de R ssi $X \rightarrow R$ est valide dans R.
- Un doublon sur X impliquerait un doublon de tuple => impossible
- Donc les deux définitions sont équivalentes :
  - $X \rightarrow R$ est valide
  - Les doubons sur X sont interdits

Exemple dans la relation 2

- DF valides : $NUMETUD \rightarrow NOMETUD$, $NUMENS \rightarrow NOMENS$, $\{NUMETUD, NUMENS\} \rightarrow ANNEE$
- $\{NUMETUD, NUMENS\}$ est une clé car on déduit que $\{NUMETUD, NUMENS\} \rightarrow \{ANNEE, NOMETUD, NOMENS\}$

Lien entre les DF et la redondance

- Toute $DF\ X \rightarrow Y$ crée de la redondance si elle est valide
- Car en cas de répétition de X, il faudra répéter Y
- Sauf si X est clé : la répétition est alors impossible

Que faire alors si $X \rightarrow Y$ est valide ?

- Bien sûr on n'abandonne pas cette contrainte pour autant !
- On va chercher à "découper" la relation, pour isoler XY de façon à ce que X devienne une clé.
- Cela s'appelle normaliser une relation

Les niveaux de normalisation

- Forme Normale de Boyce-Codd (FNBC)
  - Pour toute DF valide $R : X \rightarrow Y$, X est une clé de R
  - Plus aucune redondance générée par les DFs
- Troisième Forme Normale (3FN)
  - Accepte certaines DF $X \rightarrow A$ avec X non clé
  - Uniquement si A attribut d'une clé minimale
- Deuxième Forme Normale : pour l'Histoire...
- Première Forme Normale
  - Toutes les DF sont acceptées
  - Mais les domaines de défition des attributs sont atomiques
  - Sinon : la définition des DF perd son sens. . .

les niveaux de normalisation (suite)

FNBC $\subseteq$ 3FN $\subseteq$ 1FN

- Toute relation qui n'est pas en 3FN PEUT être normalisée
- On "tolère" parfois des 3FN... pour de bonnes et mauvaises raisons
- Même en FNBC, il y d'autres sources de redondance que les DF...
  - Les Dépendances de Jointures

Dépendance de jointure : intuition

"A l'université, les salles peuvent être utilisées par plusieurs formations ; la maintenance des salles est répartie sur plusieurs techniciens."

salles	SALLE (S)	FORMATION (F)	TECHNICIEN (T)
	C1	L3IF	SMITH
	C1	M2TIW	JAMES
	C2	L3IF	SCOTT

Il n'y a aucune DF dans le cahier des charges : FNBC

- Le lien entre F et T est "involontaire" contrairement à ce qu'on peut croire ici
- "La salle C1 est maintenue par SMITH et JAMES, mais indépendamment de la formation"
- Il manque des combinaisons !

Dépendance de jointure : intuition

"A l'université, les salles peuvent être utilisées par plusieurs formations ; la maintenance des salles est répartie sur plusieurs techniciens."
Donc : Si un technicien est affecté à une salle, il se trouve de fait "en lien" avec toutes les formations aussi affectées à cette salle.

salles	SALLE (S)	FORMATION (F)	TECHNICIEN (T)
	C1	L3IF	SMITH
	C1	M2TIW	JAMES
	C1	M2TIW	SMITH
	C1	L3IF	JAMES
	C2	L3IF	SCOTT

On DOIT rajouter les combinaisons manquantes

- Pour coller au cahier des charges
- Cette contrainte s'appelle une "dépendance de jointure"
- Elle impose donc de la redondance !

Définition des Dépendance de Jointure (DJ)

Syntaxe

- Expression $\bowtie [X_1, ...X_n]$
- Où $X_1, ..., X_n$ sont des ensembles d'attributs dont l'union donne le schéma de relation R
- Chaque X_i est appelé "composante de jointure" de la Dépendance.

Conséquence dans une relation r

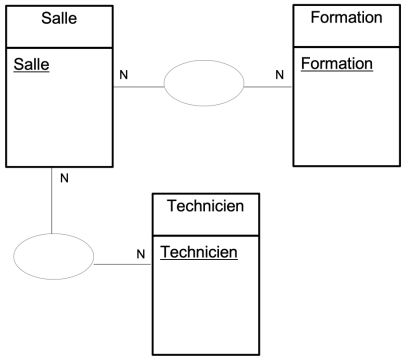
- Si on décompose la relation en n relations $r_1, ..., r_n$
- En projettant r sur chaque composante de jointure ($r_i = \Pi_{X_i}(r)$)
- Puis si on fait la jointure $r_i \bowtie ... \bowtie r_n$
- alors on retrouve exactement r.

Définition des Dépendance de Jointure (suite)

Remarques complémentaires

- Si une de ses composantes de jointure est égale à R, alors la Dépendance de Jointure est triviale
- Si une dépendance de jointure est de la forme $\bowtie [XY, XZ]$:
  - On la nomme "dépendance multivaluée" et on peut la noter $X \rightarrow\rightarrow Y$ ('X multidétermine Y indépendemment du reste de la relation')
  - Si $X \rightarrow Y$ est satisfaite, alors $X \rightarrow\rightarrow Y$ est satisfaite.
    - DONC les DJ sont une "généralisation" des DF.

Exemple



salles	SALLE (S)	FORMATION (F)	TECHNICIEN (T)
	C1	L3IF	SMITH
	C1	M2TIW	JAMES
	C1	M2TIW	SMITH
	C1	L3IF	JAMES
	C2	L3IF	SCOTT

Affectation	SALLE (S)	FORMATION (F)
	C1	L3IF
	C1	M2TIW
	C2	L3IF

Nettoyage	SALLE (S)	TECHNICIEN (T)
	C1	SMITH
	C1	JAMES
	C2	SCOTT

Ici, "Salle" respecte la DJ $\bowtie [SF; ST]$

- Car salles = Affectation $\bowtie$ Nettoyage
- Les liens entre F et T ne sont pas une info pertinente, simple conséquence des deux associations
- Cette DJ est une dépendance multivaluée $S \rightarrow\rightarrow F$ (équivalente à $S \rightarrow\rightarrow T$)

Exemple de DJ de taille 3

Revenons à la relation d'origine.

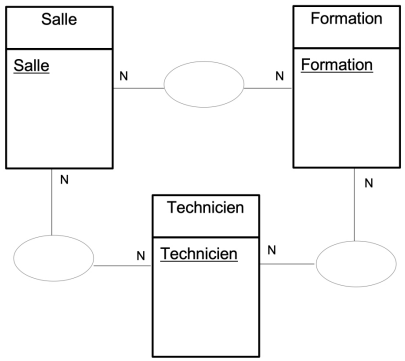
Considérons un autre cahier des charges

Si une salle affectée à une formation pour laquelle un technicien est habilité entre dans le périmètre de ce technicien, alors ce technicien est effectivement affecté à cette salle pour cette formation.

salles	SALLE (S)	FORMATION (F)	TECHNICIEN (T)
	C1	L3IF	SMITH
	C1	M2TIW	JAMES
	C2	L3IF	SCOTT

Ici, aucune dépendance multivaluée dans le cahier des charges... Pourtant, cette relation est bien redondante avec ce cahier des charges car elle peut être décomposée en trois relations. Il faut une DJ de taille 3 pour capturer cette redondance.

Exemple de DJ de taille 3



salles	SALLE (S)	FORMATION (F)	TECHNICIEN (T)
	C1	L3IF	SMITH
	C1	M2TIW	JAMES
	C2	L3IF	SCOTT

Ici, "Salle" respecte la DJ $\bowtie [SF; ST; FT]$

- Car salles = SF $\bowtie$ ST $\bowtie$ FT (à essayer en exercice)
- Les liens SF, ST et FT sont pertinents
- Les combinaisons SFT sont une simple conséquence, une "fausse" association ternaire...
- On peut donc décomposer en trois relations, "Salles" sera retrouvée par jointure.

Niveaux de normalisation supplémentaires

- un schéma R est en 4ème Forme Normale ssi, pour toute DJ non triviale de la forme $\bowtie [XY, XZ]$ [1]
  - X est clé.
- un schéma R est en 5ème Forme Normale ssi, pour toute DJ non triviale $\bowtie [X_1, \ldots, X_n]$
  - R est en FNBC, et
  - au moins un des ensembles X_i est une clé de R

Inclusion des formes normales

Chaque forme normale est un rafinement de la précédente : 5FN$\subseteq$ 4FN$\subseteq$FNBC $\subseteq$3FN $\subseteq$ 1FN

1. Appelée aussi dépendance multivaluée $X \rightarrow\rightarrow Y$ ou $X \rightarrow\rightarrow Z$

Synthèse

- Savoir repérer des relations problématiques
  - Repérer des redondances liées à des DFs
    - Comprendre la définition d'une DF $X \rightarrow Y$
    - Comment elle s'exprime dans le cahier des charges
    - "Si je fixe X, j'aurai toujours le même Y"
  - Repérer une redondance liée à une DJ
    - Comprendre la définition d'une DJ $\bowtie [XY, XZ]$
    - Comment elle s'exprime dans le cahier des charges
    - "Si je fixe X, j'aurai plusieurs Y, mais indépendamment des valeurs de Z"
- Savoir définir le niveau d'anomalie (forme normale)
- Corriger des problèmes en s'appuyant sur E/A

Pourquoi des BD sont dénormalisées ?

Une bonne conception a des coûts

- Pour acquérir les compétences nécessaires, peu fréquentes
- En temps de travail pour une étude fine
- En temps d'exécution des requêtes (jointures)

Pourtant incontournable en BD de production !

- C'est est à dire des BD avec de nombreuses mises à jour
- Beaucoup de moyens de régler le problème des performances
  - Index, requêtes pré-calculées, clusters de table, partitionnement

Optionnel pour des BD en lecture seule

- BD qui "intègrent" d'autres BD, données exportées
- Attention toutefois à ne pas perdre la documentation sémantique

Méthode : repérer la forme normale d'une relation

1 - Repérer les DF qui s'appliquent dans la relation

- C'est est le cahier des charges qui les donne
- Si la relation est peuplée de tuples, leur observation peut aider

2 - Trouver les clés minimales

- A partir des DF
- Rappel : X est clé de R ssi X détermine chaque attribut de R
- Remarque : un attribut qui n'est jamais en partie droite d'une DF est forcément dans toutes les clés minimales.
- Attention : la validité des DF est transitive.
  - $X \rightarrow Y \cap Y \rightarrow Z \Leftrightarrow X \rightarrow Z$
- Une méthode plus formelle sera vue plus loin.

Méthode : repérer la forme normale d'une relation(suite)

3 - Tester la forme normale pour les DF

- Il faut repérer les DF $X \rightarrow A$ telles que X n'est pas clé.
  - Il n'y en a pas ? On est donc en FNBC.
  - Il y en a ? Parmi elles, on cherche celles dont A n'est pas un élément d'une clé
    - Il n'y en a pas ? On est donc en 3FN
    - Il y en a ? On est en 1FN

Méthode : repérer la forme normale d'une relation (suite)

4 - valable uniquement si R est en FNBC

- Si toutes les clés minimales sont de taille 1, alors R est en 5FN
- S'il existe au moins un attribut qui ne participe à aucune clé minimale, alors R est en 5FN
- Sinon, essayer de repérer une DJ dans le cahier des charges qui pourrait s'appliquer à la relation.
- S'il n'y a pas de DJ, alors la relation est en 5FN.
- S'il y en a
  - Si aucune n'est une dépendance multivaluée : on est en 4FN
  - Si une des DJ est une dépendance multivaluée : on est en FNBC.

Méthode : comment corriger un schéma redondant ?

- Comprendre la situation sur un diagramme E/A
- Traduire ce diagramme en relationnel
- Le résultat est toujours en 5FN
- MAIS
  - L'ajout des contraintes peut regénérer des problèmes
    - Contraintes X, T, XT, I
    - Contraintes du cahier des charges non modélisées en E/A
- En cas de persistence de problèmes, l'analyse doit être poussée...
- CF suite du cours.