Livraison 3 – normalisation I

Pascal Ostermann – pasccal@orange.fr

12 avril 2021

Redondances

En modélisant une base de données, le principal écueil à éviter est la redondance. Supposons en effet qu'une même donnée ait été représentée plusieurs fois… ou pour nous fixer les idées, que dans le schéma d'école qui court depuis la première livraison, on dispose comme attributs de la note d'un étudiant à un cours, mais aussi de moyennes – calculables à partir des notes. Lorsqu'on modifie une note, il faut également modifier les moyennes. C'est au mieux relativement inefficace, puisqu'il s'agit de modifier plusieurs valeurs au lieu d'une seule. Mais un utilisateur ignorant des subtilités du schéma peut même omettre de modifier les moyennes, et rendre la base de données incohérente. Ce n'est pas trop grave ici : il suffit d'y calculer les nouvelles moyennes. Mais il est des redondances qui ne se gèrent pas aussi aisément.

Réécrivons par exemple le schéma total de cette même école sous la forme d'une unique relation :

$$Ecole(num\_et, nom\_et, prenom\_et, titre\_cours, note, date\_seance, salle\_seance)$$

FIGURE 1 – Encore le même schéma, mal normalisé

Quoique techniquement possible, une telle relation « globale » n'est généralement pas acceptable, et on en voit la raison. Par exemple la note d'un étudiant à un cours est reproduite autant de fois qu'il en a suivi de séances… et s'il y a incohérence entre ces notes, impossible de déterminer laquelle est la bonne. On dit formellement que cette relation n'est pas en deuxième forme normale ; mais il me faut d'abord quelques définitions.

Dépendances fonctionnelles

Soit X et Y deux ensembles d'attributs d'une relation R. La **dépendance fonctionnelle** (DF) X → Y est vraie dans R si et seulement si

$$\forall s, t \in R\,(s.X = t.X \supset s.Y = t.Y)$$

Propriétés

— si Y $\subset$ X, alors X $\rightarrow$ Y (DF **triviale**)
— si X $\rightarrow$ Y et Y $\rightarrow$ Z, alors X $\rightarrow$ Z
— X $\rightarrow$ YZ si et seulement si X $\rightarrow$ Y et X $\rightarrow$ Z

Clefs et superclefs

Un ensemble d'attributs X est dit **superclef** de la relation R ssi X $\rightarrow$ R (R abus de langage pour l'ensemble des attributs de R)

X est **clef** de la relation R ssi c'est une superclef minimale pour l'inclusion

Clef primaire, foreign key

Il est évident que toute relation a au moins une clef. Chaque fois que l'on définira une relation – qui deviendra informatiquement une table de hachage –, on en choisira une pour servir de clef d'accès, la **clef primaire**. À partir de maintenant, la clef primaire sera systématiquement soulignée dans nos relations. Et j'exige que vous fassiez de même en TD et à l'examen.

Lorsqu'on fera référence au contenu d'une relation R dans une autre relation S, ce sera via la clef primaire K de R. À l'intérieur de S, K est parfois appelée une **foreign key**, expression que je refuse de traduire car trompeuse : une foreign key dans S n'est en rien une clef de S. Notez bien que lorsque qu'on modifie une clef primaire, il faut également la modifier partout où elle apparaît en tant que foreign key, et modifier également tous les n-uplets concernés. Il convient donc de bien choisir la clef primaire afin de ne jamais avoir à faire cette manipulation. [1]

Décomposition sans perte d'information (SPI)

La décomposition d'une relation R en S et T est dite **sans perte d'information** ssi

$$R = S \bowtie T$$

Théorème de décomposition

Soient X et Y deux ensembles *disjoints* d'attributs d'une relation R. Si X $\rightarrow$ Y est vraie dans R, alors R se décompose SPI en $\prod_{XY}(R)$ et $\prod_{\complement Y}(R)$.

1. En particulier, dès qu'il s'agit d'êtres humains, il convient de les identifier par un numéro de client, d'employé ou d'étudiant AVANT de découvrir qu'il existe des homonymes. Ce numéro ne peut en France être le numéro de sécu, qui ne peut légalement (lois informatique et liberté) être employé que par l'INSEE et la Sécurité Sociale.

Remarque importante

$\complement Y$ est ici une notation pour le complémentaire de Y dans l'ensemble des attributs de R. Et j'insiste sur le fait qu'il s'agit du complémentaire de Y, la seule *partie droite de la DF !* La partie gauche, commune aux deux résultats de la décomposition, est ce qui nous permet de faire la jointure naturelle entre ces deux relations, et donc de démontrer le théorème de décomposition.

Forme normale de Boyce-Codd (FNBC)

La relation R est en forme normale de Boyce-Codd si et seulement si les seules DFs $X \rightarrow Y$ qui y sont vraies sont d'une des formes
— $Y \subset X$ (DF triviale) ou
— X est superclef

Propriété

Donné un ensemble de dépendances fonctionnelles, toute relation peut se décomposer SPI en un ensemble de relations en FNBC.

Méthode de normalisation

La propriété précédente est facile à démontrer : il suffit d'appliquer le théorème de décomposition chaque fois qu'on rencontre une DF qui viole la condition de Boyce-Codd. C'est la logique qui sous-tend la **méthode de normalisation** qui se déroule en trois temps :

1. Établir une liste d'attributs permettant de rendre compte de l'univers du discours. Elle sera plus tard considérée comme une relation, dite ≪ globale ≫.
2. Établir la liste des dépendances fonctionnelles valides sur ces attributs.
3. Décomposer la relation globale à l'aide du théorème de décomposition, afin d'obtenir des relations normalisées.

Un exemple : centre de recherche

Un centre de recherche, divisé en laboratoires, est réparti dans plusieurs bâtiments d'un campus.

Un bâtiment est identifié par son nom, et une salle par le nom de son bâtiment et un numéro. Une salle est affectée à un unique laboratoire, en tant que lieu collectif (salle machines, local café, etc...) ou en tant que bureau d'un ou plusieurs employés. Dans ce dernier cas, il lui est associé un unique numéro de téléphone. Toutes les salles d'un même laboratoire sont dans le même bâtiment, mais un bâtiment peut accueillir plusieurs laboratoires.

À chaque employé est associé un numéro, ses nom et prénom, une adresse E-mail, le laboratoire auquel il est affecté, son numéro de téléphone au travail, et – pour ceux qui l'acceptent – un numéro de téléphone personnel.

On conservera également les différents articles signés par les chercheurs. Un article sera identifié par son titre, sa date de publication, et le nom de la revue où il a été publié – ou de la conférence où il a été présenté. Un chercheur a pu signer (ou co-signer) un nombre quelconque d'articles. À chaque article est associé un unique domaine (qui n'a rien à voir avec un nom de laboratoire...), un résumé, et un ensemble de mots-clefs. Dans un proche avenir, on prévoit de stocker le texte même de l'article, sous forme de fichier LaTeX.

Liste des attributs

nom_bât, num_salle, nom_labo, num_tél, num_emp, nom_emp, prénom_emp, tél_perso, titre_art, date_art, nom_revue, texte_art, mot_clef [2]

Dépendances fonctionnelles

1. nom_bât, num_salle → nom_labo, num_tél
2. nom_labo → nom_bât
3. num_emp → nom_emp, prénom_emp, tél_perso, nom_bât, num_salle
4. titre_art, date_art, nom_revue → texte_art

Décomposition en FNBC

Le seul résultat indiscutable de cette décomposition est d'obtenir les relations
Salle(nom_bât, num_salle, nom_labo, num_tél)
Article(titre_art, date_art, nom_revue, texte_art)
Emp(num, nom, prénom, tél_perso, bât_bur, num_bur)
et un « reste » R(num_emp, titre_art, date_art, nom_revue, mot_clef)
Les relations Emp et Article sont parfaites, mais R devra être discutée au cours suivant – elle fait intervenir une dépendance multivaluée –; et Salle n'est pas en forme normale de Boyce-Codd à cause de la DF 2 : elle peut se décomposer SPI en
Labo(nom, bâtiment)
Attribution(num_salle, nom_labo, num_tél)
La relation Attribution n'est pas satisfaisante : la clef primaire ne peut y être le téléphone (ne serait-ce que parce que toutes les salles n'ont pas le téléphone) et ne peut être que la paire très peu intuitive num_salle, nom_labo. Par ailleurs, la dépendance fonctionnelle qui permettait d'exprimer la « bonne » clef nom_bât num_salle de la relation Salle ne peut plus s'exprimer dans Labo et Attribution :

2. Chaque fois que je pose cet énoncé ou que j'use de la formulation « ensemble de bidules » certains étudiants créent un attribut « ensemble de bidules. » Je répète qu'un « ensemble » ou une « liste » ne peuvent être des attributs.

on dit qu'il y a perte de dépendances. Mieux vaut ici s'abstenir de cette dernière décomposition, en garder Salle, qui est en troisième forme normale. [3]

Troisième forme normale (3FN)

La relation R est en troisième forme normale si et seulement si les seules DFs X $\rightarrow$ A (A est un attribut !) qui y sont vraies sont d'une des formes
— A $\in$ X (DF triviale) ou
— X est superclef ou
— A fait partie d'une clef

Propriétés

— Si R est en FNBC, alors elle est en 3FN.
— Toute relation peut se décomposer SPI et sans perte de dépendances en des relations en troisième forme normale.

Première et deuxième formes normales

Puisqu'il y a une troisième forme normale, vous devinez qu'il y en a aussi une première et une deuxième. Ces définitions sont plus ou moins obsolètes.

À l'origine du modèle relationnel, un attribut pouvait prendre un ensemble de valeurs : la première forme normale l'interdit et exige que ces valeurs soient atomiques : on retombe donc sur notre définition moderne du relationnel. On peut cependant trouver de la littérature sur le relationnel *non first normal form*, que je ne traduis car j'aime trop les acronymes anglais, NFNF ou NF2. Ces modèles permettent d'écrire dans une « case » du tableau un ensemble de valeurs, voire toute une relation. Ils sont plus ou moins équivalents aux modèles orientés-objets...

Quand une DF X $\rightarrow$ A viole la condition de la troisième forme normale, il y a deux cas : soit X fait partie d'une clef K, soit il ne fait partie d'aucune clef, ce qui donne un des schémas suivants

$$\begin{array}{l} X \rightarrow A \\ \subset \quad \subset \\ K \rightarrow R \end{array} \qquad\qquad\qquad K \rightarrow X \rightarrow A$$

Dans le premier cas, une partie de la clef détermine quelque chose : on dit qu'il y a DF *partielle* ; dans le second, il y a une DF *transitive*. Une relation est dite en deuxième forme normale ssi elle ne contient pas de DF partielles.

3. Une autre raison de garder Salle est le peu de sens de la redondance que l'on supprime en prenant Labo et Attribution. De fait, dans Salle, la seule redondance est celle de la répétition de l'information « tel labo est dans tel bâtiment » de sorte qu'il n'y a problème que lorsque qu'on veut modifier cette info : il faut alors la modifier pour toutes les salles. Mais dans le monde réel, une telle mise-à-jour correspond à un déménagement du labo : il ne va pas garder la même répartition de salles dans son nouveau bâtiment. Ce problème n'existe pas vraiment.