Qualité et au delà du relationnel

BUT2, CM1, 2022-2023

Gilles Nachouki

1- Etude de la qualité dans les bases de données relationnelles : concevoir un schéma normalisé et connaître la qualité du schéma avec ses avantages et ses inconvénients
2- Aller au delà du modèle relationnel : les bases de données NoSQL.

Au sommaire:

- Bases de données relationnelles :
  1. Modèle de données
  2. Normalisation et qualité
- Bases de données non relationnelles :
  1. Modèles de données
  2. Introduction à Mongodb

Dépendances fonctionnelles

Soit la relation R(A,B,C), l´attribut B est dit fonctionnellement dépendant de l´attribut A si étant donné 2 n-uplets :
$< a1, b1, c1 > et < a2, b2, c2 > a1 = a2 \rightarrow b1 = b2$
On la note A $\rightarrow$ B

Propriétés des dépendances fonctionnelles

Soit une relation R (A,B,C,D)

1. réflexivité : $A \rightarrow A$
2. augmentation : si $A \rightarrow B$ alors $A, C \rightarrow B$
3. transitivité : si $A \rightarrow B$ et $B \rightarrow C$ alors $A \rightarrow C$
4. pseudo-transitivité : si $A \rightarrow B$ et $B, C \rightarrow D$ alors $A, C \rightarrow D$
5. union : si $A \rightarrow B$ et $A \rightarrow C$ alors $A \rightarrow B, C$
6. décomposition : si $A \rightarrow B, C$ alors $A \rightarrow B$ et $A \rightarrow C$.

Typologie des dépendances fonctionnelles

- dépendance triviale : une dépendance $X \rightarrow Y$ est triviale si $Y \subset X$,
- dépendance simple : une dépendance $X \rightarrow Y$ est dite simple si sa partie droite ne comporte qu'un seul attribut,
- dépendance élémentaire (complète): une dépendance $X \rightarrow Y$ est élémentaire si pour tout $X' \subset X$, la dépendance fonctionnelle $X' \rightarrow Y$ n'est pas vérifiée. En d'autres termes, Y ne dépend pas fonctionnellement d'une partie de X,
- dépendance directe : une dépendance $X \rightarrow Y$ est directe si Y ne dépend pas de X par transitivité (cad il existe Z tel que $X \rightarrow Z$ et $Z \rightarrow Y$).

Formes Normales

Les trois premières formes normales:

- 1NF: une relation est dite en 1NF si tout attribut a un domaine qui contient uniquement des valeurs atomique,
- 2NF : une relation est dite en 2NF si et seulement si elle est en 1NF et que tout attribut n'appartenant pas à la clé de la relation R ne dépend pas d'une partie de la clé de R (dépendances élémentaires).

Formes Normales

- 3NF: une relation est dite en 3NF si elle est en 2NF et si tout attribut n'appartenant pas à la clé de R ne dépend pas transitivement de la clé (dépendances directes).
  Remarque : la 3NF ne traite pas de façon correcte le cas d'une relation qui a 2 clés candidates ou plus telles que : (1) les deux clés candidates sont composées et (2) elles se chevauchent
- 3NFBC : une relation est en 3NFBC (Forme Normale de Boyce Codd) si elle est en troisième forme normale et si les seules sources de dépendances sont des clés de R

Formes Normales

On considère la relation R suivante :

A	B	C
1	2	4
2	2	4
3	2	4
3	5	6
5	5	7

- L'ensemble DF = { AB $\rightarrow$ C, C $\rightarrow$ B }
- La clé primaire primaire de la relation R est AB. AC est une clé candidate.
- {A,B } $\bigcap$ {A,C } $\neq \phi$.
- Cette relation est en 3NF mais pas en 3NFBC.

- Fermeture transitive de F (F^t): obtenue par application de la propriété de transitivité sur l'ensemble F de dépendances fonctionnelles.

Exemple : F = $\{A \rightarrow B, B \rightarrow C\}$

$F^t = \{A \rightarrow B, B \rightarrow C, A \rightarrow C\}$

- La fermeture d'un ensemble F de DFs (F^+) est obtenue en appliquant toutes les propriétés de dépendances fonctionnelles.

Exemple :
F = {A $\rightarrow$ B, B $\rightarrow$ A }

F^+ = {A $\rightarrow$ B, B $\rightarrow$ A, A $\rightarrow$ A, B $\rightarrow$ B, AB $\rightarrow$ A, AB $\rightarrow$ B, AB $\rightarrow$ AB }

La fermeture d'un ensemble X d'attributs de R dans F (X_F^+) constitue l'ensemble des attributs de R qui peuvent être obtenus à partir de X en appliquant les propriétés de dépendances fonctionnelles

Un ensemble minimal de dépendances de F permettant de générer toutes les dépendances de F^+:

1. $F^+(F_{min}) = F^+(F)$,
2. il n'existe pas F' / $F' \subset F_{min}$ et $F^+(F') = F^+(F_{min})$

Algorithmes

Dans cette section, on étudie quelques algorithmes indispensables pour concevoir une base de données normalisée en 3NF ou 3NFBC

- A1- Recherche de la fermeture d'un ensemble d'attributs X de R dans F
- A2- Recherche de la couverture minimale de F (F_{min})
- A3- Recherche d'un schéma en 3NF (algorithme de synthèse)
- A4- Recherche d'un schéma en 3NFBC (algorithme de décomposition)

Algorithme 1 : fermeture d'un ensemble X d'attributs de R dans F

X^+:=X
Répéter
Aux := X^+
Pour chaque DF Y $\rightarrow Z$ de F faire
si Y est inclus dans X^+ alors X^+ := X^+ $\cup$ Z
Jusqu'à Aux := X^+ ou X^+ = R

Exemple :

Soit R = (A,B,C,D,E) et
F = {A $\rightarrow$ B, B $\rightarrow$ C, ABE $\rightarrow$ D, A $\rightarrow$ C}.
La fermeture de l'ensemble X = {B,E} dans F, $(BE)^+_F$ = {B,E, C }

Algorithme 2 : calcul de la couverture minimale (F_{min}) d'un ensemble de Dfs

Les étapes :

- Rendre les DF simples : $X \rightarrow$ A où X est un attribut simple ou composé, A est un attribut simple,
- Eliminer les attributs superflus: Soit (X $\rightarrow$ A) $\in$ F, on dit qu'un attribut B $\in$ X est superflu si on a: ((F - {X $\rightarrow$ A}) $\bigcup$ { X - {B} $\rightarrow$ A})$^+$ = F$^+$,
- Eliminer les DF redondantes : X $\rightarrow$ A est redondante ssi (F - (X $\rightarrow$ A))$^+$ = F$^+$.

Exemple : couverture minimale

soit R = (A,B,C,D,E) et F = {A $\rightarrow$ B, B $\rightarrow$ C, ABE $\rightarrow$ D, A $\rightarrow$ C }

- Première étape : On vérifie facilement que toutes les df sont simples dans F.
- Deuxième étape : éliminer les attributs superflus : on considère la dépendance suivante : ABE $\rightarrow$ D. Est-ce que l'attribut A est superflu dans cette dépendance? On cherche la fermeture de l'ensemble d'attributs X= {B,E} dans F. Est-ce que D $\in X_F^+$?. Si oui alors l'attriut A dans ABE $\rightarrow$ D est superflu. X_F^+ = {B,E, C } et D $\notin X_F^+$ donc A n'est pas superflu.
- Est-ce que l'attribut B est superflu? la réponse est oui
- Dernière étape : éliminer les dfs redondantes. Est-ce que la df A $\rightarrow$ C est redondante ? On cherche si C $\in A_F^+$ dans l'ensemble F - {A $\rightarrow$ C}

Algorithme 3 de synthèse

On part d'une relation et d'un ensemble F de dépendances fonctionnelles. Le principe de cet algorithme est le suivant :

- on remplace F par une couverture minimale F_{min},
- on regroupe les dfs ayant même membre gauche X ($X \rightarrow A$). Pour chaque membre gauche X, on définit un schéma de relation contenant tous les attributs intervenant dans ces dfs.
- si aucun des schémas définis à l'étape 2 ne contient de clé R, on ajoute un schéma de relation muni d'une clé de R.

Qualité de schéma

- Chaque schéma issu de cette décomposition est en 3NF;
- Cette décomposition est sans perte de données et sans perte de dépendances fonctionnelles.

Exemple : Algorithme de synthèse

Soit R = (A,B,C,D,E) et F = {$A \to B$, $A \to D$, $B \to C$, $E \to D$ }

- On part de la couverture minimale de F. On chercher la clé de R : AE.
- On regroupe les DF qui ont même membre gauche :
  R1(A,B,D) F1 = { $A \to B$, $A \to D$ }
  R2(B, C) F2 = { $B \to C$ }
  R3(E, D) F3 = { $E \to D$ }
- On ajoute R4(A,E) F = { }
  Cette décomposition est SPI et SPD.

Algorithme 4 de décomposition

On part d'une relation R et d'un ensemble F de dépendances fonctionnelles. Le principe de cet algorithme est le suivant :

- on remplace F par une couverture minimale F_{min},
- trouver dans R une dépendance non triviale $X \rightarrow Y$ telle que X ne contienne pas une clé de R
- remplacer R par R1($X \bigcup Y$, F1) et R2(U - Y,F2) où F1 et F2 sont deux ensembles de dépendances fonctionnelles obtenues à partir de F,
- pour tout Ri(Xi,Fi) et Rj(Xj,Fj) tels que $Xi \subseteq Xj$ supprimer Ri.

Qualité de schéma

- Chaque schéma issue de cette décomposition est en 3NFBC;
- Cette décomposition est sans perte de données mais elle ne garantie pas de retrouver toutes les dépendances fonctionnelles.

Exemple : Algorithme de décomposition

Soit R = (A,B,C,D,E) et F = {$A \rightarrow B$, $A \rightarrow D$, $B \rightarrow C$, $E \rightarrow D$ }

- On part de la couverture minimale de F. On chercher la clé de R AE.
- On décompose selon la DF suivante : $A \rightarrow B$. On obtient : R1(A,B), F1 = {$A \rightarrow B$ } et R2 (A,C,D,E) et F2 = {$A \rightarrow D$, $E \rightarrow D$ }. La clé de R2 est ACE. On décompose selon la DF $A \rightarrow D$. On obtient :
  R21(A,D), F21 = {$A \rightarrow D$ } et R22(A,C,E), F22 = { }.
  Cette décomposition est sans perte de données mais avec perte de DF.

Théorème de Heath

- Soit la relation R(X,Y,Z) et la dépendance fonctionnelle suivante : $X \rightarrow Y$ alors R est décomposable sans perte d'information (SPI) en deux relations R1(X,Y) et R2(X,Z). R égale à la jointure de sa projection sur {X,Y} et {X,Z},
- Toute relation a au moins une décomposition en troisième forme normale qui est SPI et sans perte de dépendances fonctionnelles (SPD)

Théorème d'ULLMAN

- la décomposition d'une relation est dite sans perte d'information ssi: $X = X1 \bigcup X2$ et $X1 \bigcap X2 \rightarrow X1 - X2$ ou $X2 - X1$,
- la décomposition est dite sans perte de dépendances ssi $F^+ = (F1 \bigcup F2)^+$