

# LIF10 – FONDLEMENTS DES BASES DE DONNÉES

## TD2 – Introduction aux dépendances

Licence informatique – Automne 2017–2018

Les questions marquées du symbole (†) sont à préparer pour la séance

### Exercice 1 : modélisation avec les dépendances (†)

Soit  $R$  le schéma de bases de données suivant :

- $Films = \{IDFilm, Titre, Annee, IDStudio\}$  ;
- $Reprises = \{IDReprise, IDOriginal, Similarite\}$  ;
- $Studios = \{IDStudio, Nom, Adresse\}$ .

1. Trouver la dépendance fonctionnelle ou d'inclusion permettant de restreindre les extensions possibles de cette base, pour chacune des assertions suivantes :
  1. chaque film a un identifiant unique à partir duquel on connaît tous ses attributs ;
  2. la même année, deux films ne peuvent pas avoir le même titre ;
  3. chaque studio doit avoir effectivement participé à la réalisation d'un film ;
  4. un film peut être repris plusieurs fois, un film peut reprendre plusieurs films et pour chaque film repris par un autre, il y a un unique taux de similarité.
2. En supposant les contraintes précédentes satisfaites, peut-on également demander que plusieurs studios puissent participer à la réalisation d'un même film ? Justifier et proposer une solution au problème.

### Exercice 2 : préservation des dépendances par les requêtes

Soit  $r$  une instance de  $R$  qui satisfait à la dépendance  $R : X \rightarrow Y$  (soit  $r \models X \rightarrow Y$ ) et  $s$  une instance quelconque. Pour chaque expression ci-dessous, indiquer en le justifiant si elle est vraie.

1.  $\sigma_C(r) \models X \rightarrow Y$
2.  $r \cup s \models X \rightarrow Y$
3.  $r \setminus s \models X \rightarrow Y$
4.  $\pi_W(r) \models X \rightarrow Y$
5.  $r \times s \models X \rightarrow Y$
6.  $r \bowtie s \models X \rightarrow Y$

### Exercice 3 : axiomatisation des dépendances fonctionnelles

On rappelle les règles d'inférences suivantes pour les dépendances fonctionnelles.

$$\frac{Y \subseteq X}{X \rightarrow Y} \sigma_R \text{ (réflexivité)}$$

$$\frac{X \rightarrow Y \quad X \rightarrow Z}{X \rightarrow YZ} \sigma_C \text{ (composition)}$$

$$\frac{X \rightarrow Y}{WX \rightarrow WY} \sigma_A \text{ (augmentation)}$$

$$\frac{X \rightarrow YZ}{X \rightarrow Y} \sigma_D \text{ (décomposition)}$$

$$\frac{X \rightarrow Y \quad Y \rightarrow Z}{X \rightarrow Z} \sigma_T \text{ (transitivité)}$$

$$\frac{X \rightarrow Y \quad WY \rightarrow Z}{WX \rightarrow Z} \sigma_P \text{ (pseudo-transitivité)}$$

1. Donner une preuve que  $\{AB \rightarrow C, A \rightarrow D, CD \rightarrow EF\} \models AB \rightarrow F$  en utilisant le système  $\{\sigma_R, \sigma_A, \sigma_T\}$
2. La règle suivante est-elle correcte ?

$$\frac{XW \rightarrow Y \quad XY \rightarrow Z}{X \rightarrow (Z \setminus W)}$$

3. Montrer que toute preuve de  $F \models X \rightarrow Y$  utilisant la règle  $\sigma_P$  peut être transformée en une preuve n'utilisant que  $\sigma_A$  et  $\sigma_T$ .
4. Montrer que toute preuve de  $F \models X \rightarrow Y$  utilisant les règles  $\sigma_R, \sigma_A$  et  $\sigma_T$  peut être transformée en une preuve n'utilisant que  $\sigma_R$  et  $\sigma_P$ .
5. En déduire que le système  $\{\sigma_R, \sigma_P\}$  est correct et complet pour l'inférence des DFs.

### Exercice 4 : adéquation du système d'Armstrong

1. Démontrer que les règles du système d'Armstrong (réflexivité, transitivité et augmentation) sont justes en exploitant la définition de la satisfaction d'une dépendances.

### Exercice 5 : vérification des dépendances en SQL

1. Prouver que  $r \models X \rightarrow Y$  si et seulement si  $|\pi_X(r)| = |\pi_{XY}(r)|$ , en déduire une méthode qui permet de tester la satisfaction d'une dépendance fonctionnelle avec SQL. Commenter son efficacité par rapport à une méthode faisant intervenir seulement la sémantique des dépendances.
2. Prouver que  $r, s \models R[X] \subseteq S[Y]$  si et seulement si  $|\pi_X(r) \setminus \pi_Y(s)| = 0$ , en déduire une requête SQL qui permet de tester la satisfaction d'une dépendance d'inclusion.

# Corrections

## Solution de l'exercice 1

1.  $Films : IDFilm \rightarrow Titre, Annee, IDStudio$  ;  
 2.  $Films : Titre, Annee \rightarrow IDFilm$  ou éventuellement  $Films : Titre, Annee \rightarrow IDFilm, IDStudio$  ;  
 3.  $Studios[IDStudio] \subseteq Films[IDStudio]$  ;  
 4.  $Reprises : IDReprise, IDOriginal \rightarrow Similarite$ , noter que la partie gauche contient deux attributs.
2. Ce n'est possible car  $Films : IDFilm \rightarrow IDStudio$  (et éventuellement car  $Films : Titre, Annee \rightarrow IDStudio$ ) : il n'y a qu'un unique studio associé à chaque film. Il faut donc choisir entre la dépendance précédente et  $Films : IDFilm \rightarrow Titre, Annee$ . Noter que cette solution pose un problème de normalisation et qu'un meilleur schéma serait de scinder  $Films$  en  $Films = \{IDFilm, Titre, Annee\}$  et  $Realise = \{IDFilm, IDStudio\}$ .

## Solution de l'exercice 2

1. vraie car comme  $\sigma_C(r) \subseteq r$  il ne peut pas y avoir plus de contre-exemples de  $X \rightarrow Y$  dans  $\sigma_C(r)$  que dans  $r$  qui n'en contient aucun ;
2. faux en général, même si  $s \models X \rightarrow Y$  car rien n'impose que que le graphe de la fonction  $X \rightarrow Y$  déterminé par  $r$  soit compatible avec celui de  $s$  : on peut avoir  $\langle x, y_0 \rangle \in r$  et  $\langle x, y_1 \rangle \in s$  avec  $y_0 \neq y_1$  ;
3. vraie, même justification que pour  $\sigma_C(r)$  ;
4. vraie, en précisant que  $X \subseteq W$  pour que l'expression soit sensée ;
5. vraie, (faut-il préciser que  $X \not\subseteq S$  et  $Y \not\subseteq S$  ?)
6. vraie (justification ?).

## Solution de l'exercice 3

1. L'arbre est le suivant :

$$\frac{\frac{\frac{AB \rightarrow C}{AB \rightarrow AC} \sigma_A \quad \frac{A \rightarrow D}{AC \rightarrow CD} \sigma_A}{AB \rightarrow CD} \sigma_T \quad \frac{CD \rightarrow EF}{AB \rightarrow EF} \sigma_T \quad \frac{F \subseteq EF}{EF \rightarrow F} \sigma_R}{AB \rightarrow F} \sigma_T$$

2. Non, on exhibe un contre-exemple avec l'instance suivante qui satisfait  $XW \rightarrow Y$  et  $XY \rightarrow Z$  mais pas  $X \rightarrow (Z \setminus W)$  :

| $W$   | $X$   | $Y$   | $Z$   |
|-------|-------|-------|-------|
| $w_0$ | $x_0$ | $y_0$ | $z_0$ |
| $w_1$ | $x_0$ | $y_1$ | $z_1$ |

3. Il s'agit de montrer que l'on peut prouver  $WX \rightarrow Z$  à partir de  $X \rightarrow Y$  et  $WY \rightarrow Z$  en utilisant uniquement  $\sigma_A$  et  $\sigma_T$  :

$$\frac{\frac{X \rightarrow Y}{WX \rightarrow WY} \sigma_A \quad WY \rightarrow Z}{WX \rightarrow Z} \sigma_T$$

4. Comme  $\sigma_R$  appartient aux deux ensembles, il suffit de montrer la transitivité et l'augmentation à l'aide de la réflexivité et la pseudo-transitivité seulement. La transitivité est en fait un cas dégénéré de la pseudo-transitivité avec  $W = \emptyset$  :

$$\frac{X \rightarrow Y \quad Y \rightarrow Z}{X \rightarrow Z} \sigma_P$$

Pour l'augmentation s'obtient en posant  $Z = WY$  dans la règle de pseudo-transitivité :

$$\frac{X \rightarrow Y \quad \frac{WY \subseteq WY}{WY \rightarrow WY} \sigma_R}{WX \rightarrow WY} \sigma_P$$

5. L'antépénultième question montre que le système  $\{\sigma_R, \sigma_P\}$  est correct. D'autre part, on sait que le système d'Armstrong  $\{\sigma_R, \sigma_A, \sigma_T\}$  est complet, la question précédente montre que on peut transformer toutes les preuves de ce système par des preuves ne faisant intervenir que  $\{\sigma_R, \sigma_P\}$  ce qui montre sa complétude.

### Solution de l'exercice 4

- Les démonstrations sont assez directes. Je pense qu'il ne faut en corriger qu'une, la deuxième par exemple. Soit  $R$  un schéma de relation et  $X, Y, Z$  trois sous-ensembles de  $R$ .
  - réflexivité : soient  $t_1, t_2 \in r$  où  $r$  une relation quelconque. Supposons  $t_1[X] = t_2[X]$ . Si  $Y \subseteq X$ , alors  $t_1[Y] = t_2[Y]$  et  $r \models X \rightarrow Y$ .
  - transitivité : c'est essentiellement la transitivité de l'implication logique. Soit  $r$  une relation quelconque sur  $r$  telle que  $r \models \{X \rightarrow Y; Y \rightarrow Z\}$ . Soient  $t_1, t_2 \in r$  deux tuples quelconques tels que  $t_1[X] = t_2[X]$ . Puisque  $r \models X \rightarrow Y$  on a  $t_1[Y] = t_2[Y]$ . Puisque  $r \models Y \rightarrow Z$  on a  $t_1[Z] = t_2[Z]$ .
  - augmentation : on suppose  $r \models \{X \rightarrow Y\}$  et  $t_1[WX] = t_2[WX]$ . On a donc  $t_1[W] = t_2[W]$  d'un part et  $t_1[X] = t_2[X]$  d'autre part. Comme  $r \models \{X \rightarrow Y\}$  on déduit que  $t_1[Y] = t_2[Y]$  et ainsi  $t_1[WY] = t_2[WY]$ .

### Solution de l'exercice 5

- Pour la direction *seulement si*, on a une injection évidente de  $\pi_X(r)$  dans  $\pi_{XY}(r)$  : à chaque  $t \in \pi_X(r)$  il existe au moins un tuple  $t_0 \in r$  avec  $t_0[X] = t$  et on fait correspondre ce  $t_0[XY] \in \pi_{XY}(r)$  à  $t$  (on a donc  $|\pi_X(r)| \leq |\pi_{XY}(r)|$ ). Supposons que  $r \models X \rightarrow Y$  alors pour chaque choix de  $t_0 \in r$  tel que  $t_0[X] = t$  il existe exactement un unique  $t_0[XY]$  pour chaque  $t$ , hors  $t_0[XY] \in \pi_{XY}(r)$  et la fonction précédemment définie précédemment est surjective. On vient de construire une bijection entre  $\pi_{XY}(r)$  et  $\pi_X(r)$  les deux ensembles ont ainsi même cardinalité.  
 Pour la direction *si*, on prouve la contraposée. Supposons que  $r \not\models X \rightarrow Y$ , autrement dit  $\exists t_0, t_1 \in r$  tels que  $x_0 = t_0[X] = t_1[X]$  mais  $y_0 = t_0[Y] \neq t_1[Y] = y_1$ . Dès lors,  $\langle x_0 \rangle \in \pi_X(r)$  et  $\{\langle x_0, y_0 \rangle, \langle x_0, y_1 \rangle\} \subseteq \pi_{XY}(r)$ , on ne peut donc pas construire de bijection entre  $\pi_X(r)$  et  $\pi_{XY}(r)$  ce qui implique que les cardinalités de  $\pi_X(r)$  et  $\pi_{XY}(r)$  sont différentes.  
 Pour SQL, il faut comparer le résultats des requêtes suivantes qui calculent les cardinalités des projections (éventuellement en SQL avec un MINUS ou un FROM DUAL) :

```
SELECT COUNT(*)
FROM (SELECT DISTINCT X
      FROM r)
```

```
SELECT COUNT(*)
FROM (SELECT DISTINCT X, Y
      FROM r)
```

La propriété démontrée est intéressante car elle remplace un éventuel produit cartésien  $r \times r$  sur lequel on teste  $\forall t_0, t_1. t_0[X] = t_1[X] \Rightarrow t_0[Y] = t_1[Y]$  par deux parcours linéaires de  $r$ .

- On procède par équivalence successives. Par définition,  $r, s \models R[X] \subseteq S[Y]$  est équivalent à  $\pi_X(r) \subseteq \pi_Y(s)$  ce qui à son tour équivaut à  $\pi_X(r) \setminus \pi_Y(s) = \emptyset$  qui équivaut à  $|\pi_X(r) \setminus \pi_Y(s)| = 0$ . Pour SQL, on va traduire assez directement  $|\pi_X(r) \setminus \pi_Y(s)|$  et tester si le résultat est 0 :

```
SELECT COUNT(*)
FROM (SELECT DISTINCT X FROM r
      MINUS
      SELECT DISTINCT Y FROM s)
```