



# Dépendances et normalisation

## Dépendances et normalisation

Fabien Coelho, Claire Medrala

Mines Paris – PSL

Décembre 2023



Deps & Norm



# Relation universelle...

# Problème : redondance des informations

■ relation unique : un tuple opère toutes les liaisons

titre	annee	pays	capitale	date	acteur	a-pays	realisateur	r-pays
Troy	2004	USA	Wash. DC	Q1 04	Bloom	USA	Petersen	Allemagne
Troy	2004	USA	Wash. DC	Q1 04	Pitt	USA	Petersen	Allemagne
Troy	2004	France	Paris	Q2 04	Bloom	USA	Petersen	Allemagne
Troy	2004	France	Paris	Q2 04	Pitt	USA	Petersen	Allemagne
Pirates	2003	USA	Wash. DC	Q2 03	Bloom	USA	Verbinski	USA
Pirates	2003	France	Paris	Q3 03	Bloom	USA	Verbinski	USA

acteur - a-pays, pays - capitale, titre - réalisateur

### Conséquences : risques d'incohérences

- mises à jour *changement de nationalité d'un réalisateur...*
- insertions *ajout d'un même acteur dans une autre nationalité*
- suppressions *disparition d'un acteur si aucun film*

Le film *Troy (2004)* est sorti en *France* (capitale *Paris*) en *Q2 04* avec l'acteur *Pitt* originaire des *USA* et réalisé par *Petersen* originaire d'*Allemagne*.

Deps & Norm



# Comment éviter la redondance ?

Deps & Norm

- séparation** des concepts, dans des relations différentes
- modélisation** approche ascendante avec E/A, UML...
  - à partir des entités, liées par des associations
- normalisation** approche descendante théorique
  - **relation universelle** initiale
  - analyse des dépendances fonctionnelles entre attributs
  - décomposition en relations jointes (coût joint, éco mém)
  - définition de formes normales avec bonnes propriétés



# Dépendances Fonctionnelles (DF)

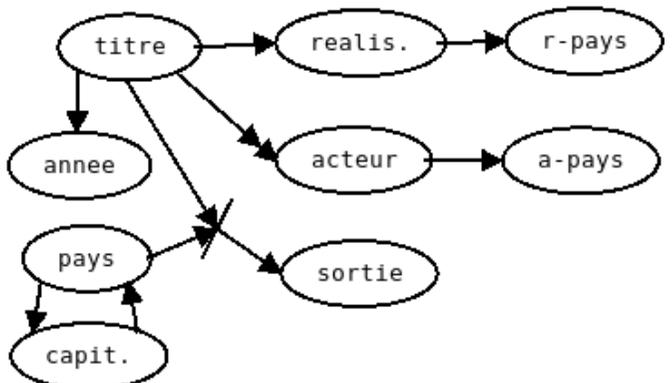
Deps & Norm

- dépendance fonctionnelle** d'une relation  $R$  :
  - soit deux sous-ensemble d'attributs  $S$  et  $D$  de  $R$
  - ils dépendent *fonctionnellement* l'un de l'autre, noté  $S \rightarrow D$  :
  - $\forall t_1 \in R, \forall t_2 \in R, \Pi_S(t_1) = \Pi_S(t_2) \Rightarrow \Pi_D(t_1) = \Pi_D(t_2)$
- exemples** titre pays  $\rightarrow$  sortie  
titre  $\rightarrow$  année réalisateur r-pays
- dépendance sémantique** réelle
  - pas simplement constatées sur un jeu de données !
  - donc ne peuvent pas être dérivées des données !
  - exemple, on a pas : année  $\rightarrow$  titre



# Exemple de dépendances (minimales) pour les films

Deps & Norm



- pas de DF entre un film et *un* acteur particulier  
mais DMV (dépendance multivaluée) entre un film et *ses* acteurs



# Propriétés des DF (axiomes d'Armstrong, 1974)

Deps & Norm

- réflexivité** auto-détermination des attributs
  - $X \subset S \Rightarrow S \rightarrow X$
  - e.g. titre année réalis.  $\rightarrow$  titre
- augmentation** ajout d'attributs
  - $\forall Z, X \rightarrow Y \Rightarrow XZ \rightarrow YZ$
  - e.g. titre  $\rightarrow$  année  $\Rightarrow$  titre nom  $\rightarrow$  année nom
- transitivité** des DF
  - $X \rightarrow Y \wedge Y \rightarrow Z \Rightarrow X \rightarrow Z$
  - e.g. titre  $\rightarrow$  pays  $\rightarrow$  capitale



## Autres propriétés des DF

Deps & Norm

**union** de deux dépendances fonctionnelles  
 $(X \rightarrow Y \wedge X \rightarrow Z) \Rightarrow X \rightarrow YZ$   
 titre pays  $\rightarrow$  réalisateur  $\wedge$  titre pays  $\rightarrow$  sortie  
 $\Rightarrow$  titre pays  $\rightarrow$  réalisateur sortie  
 preuve : augmentation et transitivité

**pseudo-transitivité** des DF  
 $(X \rightarrow Y \wedge WY \rightarrow Z) \Rightarrow WX \rightarrow Z$   
 capitale  $\rightarrow$  pays  $\wedge$  titre pays  $\rightarrow$  sortie  
 $\Rightarrow$  titre capitale  $\rightarrow$  sortie  
 preuve : augmentation et transitivité

9 / 24



## Autres propriétés (suite)

Deps & Norm

**décomposition** de DF  
 $\forall Z \subset Y, X \rightarrow Y \Rightarrow X \rightarrow Z$   
 titre pays  $\rightarrow$  réalisateur sortie  
 $\Rightarrow$  titre pays  $\rightarrow$  réalisateur  
 preuve : réflexivité et transitivité

### Conséquence de l'union et de la décomposition

- on peut toujours proposer des DF vers un singleton
- $X \rightarrow A_1 \dots A_n \quad X \rightarrow A_1, \dots, X \rightarrow A_n$

10 / 24



## DF élémentaire

Deps & Norm

**définition**  $S \rightarrow D$  avec  $S$  le plus petit pour  $D$   
 $S$  tel que  $D \cap S = \emptyset$  et  $\forall X | X \subset S \wedge X \rightarrow D \Rightarrow X = S$   
*plus petite clef d'un tuple, séparation des attributs*  
 $S$  attributs sources,  $D$  attributs déduits

**remarque** pas de solution unique à partir de  $D \cup S$   
 e.g. titre année pays sortie réalisateur capitale  
 titre pays  $\rightarrow$  année capitale réalisateur  
 titre capitale  $\rightarrow$  année pays réalisateur

11 / 24



## Fermeture transitive des DF d'une relation

Deps & Norm

**définition**  $F^+$  déduit de  $F$  par réflexivité, augmentation et transitivité...  
*construction itérative de toutes les DF possibles*

### La fermeture transitive définit toutes les DF vraies

- validité des DF dérivées par examen des trois règles
- complétude des DF dérivées...

12 / 24



## 6 Formes Normales hiérarchisées

Deps & Norm

- 1NF attributs atomiques
  - 2NF 1NF + attributs non clef en DF de la clef entière
  - 3NF 2NF + attributs non clef en dépendances directe des clefs
  - Boyce-Codd NF/3.5NF 3NF + tous les déterminants sont des clefs
  - 4NF BCNF + sans dépendances multivaluées non-triviales
  - 5NF 4NF + ...
- les NF mesurent divers degrés de redondance. ...
  - en pratique, on vise BCNF ou 4NF
  - moins si motivé par performances : éviter des jointures

13 / 24



## 1NF : première forme normale

Deps & Norm

- **attributs atomiques** (pas de tableaux ou d'ensembles)
- lié à une modélisation insuffisante (attribut vs entité)
- tableaux (Access, PostgreSQL) à éviter. ...

Cine(*titre*,acteurs. . .)

- requêtes non relationnelles avec tableaux. . . ou texte. . .  
*relation dont l'attribut est le nom d'un acteur ?*
- solution : Joue(*titre*,*nom*)

14 / 24



## 2NF : seconde forme normale

Deps & Norm

- 1NF + **attributs non clef en DF de la clef entière**  
*non clef : attribut n'appartenant à aucune des clefs*

Cine(*titre*,*nom*, a-pays) est 1NF

<i>titre</i>	<i>nom</i>	<i>a-pays</i>
Troy	Pitt	USA
Troy	Bloom	USA
Pirates	Bloom	USA

- problème : redondance du *a-pays* qui ne dépend que du *nom*
- solution : Film(*titre*,*nom*), Acteur(*nom*,*a-pays*)

15 / 24



## 3NF : troisième forme normale (Codd, 1970)

Deps & Norm

- 2NF + **attributs non clefs en dépendances directe des clefs**  
*pas de transitivité des DF pour les attributs non clefs !*

Cine(*nom*,a-pays,capitale) est 2NF

<i>nom</i>	<i>a-pays</i>	<i>capitale</i>
Pitt	USA	Wash. DC
Bloom	USA	Wash. DC
Petersen	Allemagne	Berlin
Verbinski	USA	Wash. DC

- problème : lien *pays-capitale* répété pour chaque *nom*
- solution : Acteur(*nom*,*a-pays*), Pays(*pays*,*capitale*)

16 / 24



### 3.5/BCNF : forme normale de Boyce-Codd (Heath 1971...)



### Décomposition d'une relation

Deps & Norm

Deps & Norm

- 3NF + **tous les déterminants sont des clefs potentielles**

- motif  $AB \rightarrow C, C \rightarrow B$

– *cinéma d'un type le plus proche*

Proche(*personne,type,lieu*)

Proche(*personne,type,lieu*)

est 3NF

prénom	type	plus proche
Calvin	Cinéma	Le Club
Calvin	Multiplex	Alticiné
Hobbes	Cinéma	Le Club
Hobbes	Multiplex	Alticiné
Susie	Cinéma	Studio des Ursulines
Susie	Multiplex	MK2 Odéon

- **pas** solution : Proche(*personne,lieu*) + Lieu(*lieu, type*)
- solution partielle : Proche(*personne,type,lieu*) + Lieu + contrainte ?

17 / 24

- regroupement des attributs en sous relations  
schéma d'une relation  $R = \{A_1, \dots, A_n\}$   
vers un ensemble de schémas :  $R = R_1 \cup \dots \cup R_k$   
attributs communs entre les relations !

Jointure sans perte d'information vis à vis de dépendances

- $r$  relation de  $R$ ,  $R$  décomposé,  $r = \pi_{R_1}(r) \bowtie \dots \bowtie \pi_{R_k}(r)$

18 / 24



### Algorithme de Tsou et Fisher (1982) : DF vers BCNF



### Dépendances multivaluées (DMV)

Deps & Norm

Deps & Norm

- suppose d'avoir le graphe des DF...
- nécessite donc une modélisation précise et juste !

**lemme** toute relation a 2 attributs est BCNF, et si  $R$  non BCNF,  $\exists AB \in R : (R - AB) \rightarrow A$

**algorithme** décomposition récursive

- notées  $X \twoheadrightarrow Y$
- plusieurs valeurs possible d'un attribut dans une clef (1 pour n)
- ce n'est pas une *dépendance fonctionnelle* (1 pour 1)

Générique(*titre,acteur*)

titre	acteur
Troy	Pitt
Troy	Bloom
Pirates	Bloom

- exemple : film  $\twoheadrightarrow$  acteur
- l'ensemble des acteurs dépend du film

19 / 24

20 / 24



## 4NF : quatrième forme normale (Fagin, 1977)

Deps & Norm

- BCNF + **sans dépendances multivaluées non-triviales**

triviale :  $X \rightarrow Y \Rightarrow X$  est une clef

Générique(*titre,acteur,réalisateur*)

est BCNF

<i>titre</i>	<i>acteur</i>	<i>realisateur</i>
Troy	Pitt	Petersen
Troy	Bloom	Petersen
Pirates	Bloom	Verbinski

- problèmes : relations indépendantes liés dans une même relation  
*pas de film sans acteur, réalisateur répété...*  
pas de lien direct entre réalisateur et acteur
- solution : *Joue(titre,acteur)*, *Réalise(titre,réalisateur)*

21 / 24



## 5NF : cinquième forme normale ?

Deps & Norm

- pas de décomposition possible d'une relation ?
- similaire à 4NF, mais décomposition en 3 relations ?
- exemples un peu tordus...

### Autres types de dépendances

- jointure* Rissanen 1979
- algébriques* Yannakakis et Papadimitriou 1980
- généralisées* Sadri et Ullman 1981
- implicationnelles* Fagin 1982

22 / 24



## Conseil de lecture

Deps & Norm

- Jeffrey D. Ullman
- Principles of Database and Knowledge-Base Systems
- Volume 1
- Chapitre 7
- Computer Science Press, 1988

23 / 24



## Dénormalisation ?

Deps & Norm

- normalisation réduit les redondances  
facilite insertion, mise à jour et effacement
- mais impact sur les performances de consultation  
rétablissement des liens par jointures  
optimisation : utilisation d'index...
- problème de l'évolution du schéma  
ajout a posteriori de concepts  
en général, pas de reprise de la modélisation
- dénormalisation parfois utile pour performances  
systèmes distribués, SQL ou NoSQL...

24 / 24