

Bases de données



Principes

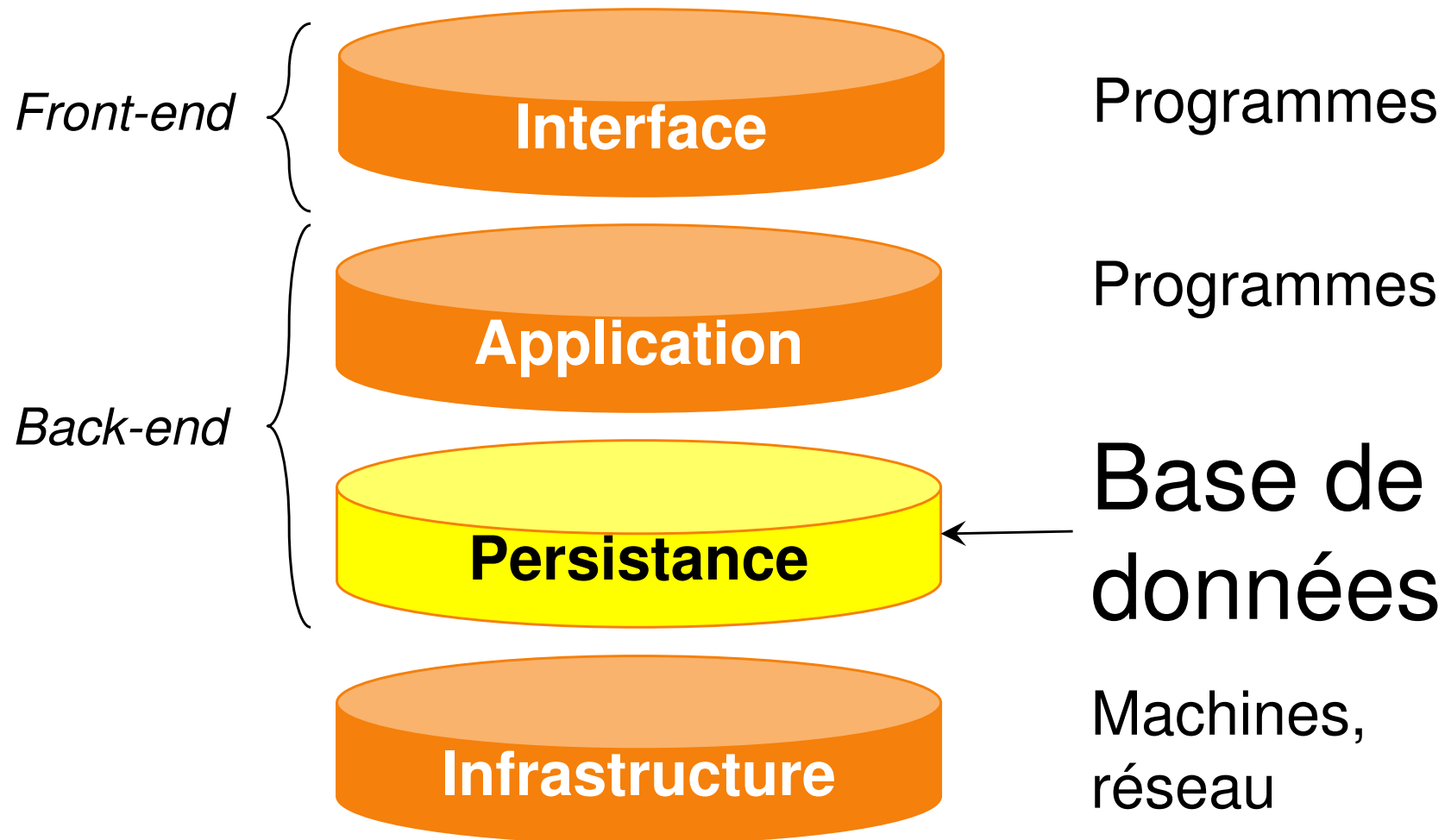
Bases de données pré-informatiques



- Notions de fiche, de fichier, de fichage, de champ
- Structuration et métadonnées
- Volume et standardisation
- Une certaine versatilité des utilisations
- Utilisation pilotée par l'indexation



Les BdD au sein de l'activité informatique



Le stockage des données

- Anciens modèles (BdD réseau, BdD hiérarchiques)
- Modèle relationnel (Oracle, SQL Server, MySQL...)

- Relationnel normalisé

*Type d'utilisation :
opérationnelle (OLTP)*

- Relationnel dénormalisé

- Stockage vertical

*Type d'utilisation :
analytique (OLAP)*

- Multidimensionnel (hypercube)

- Nouveaux modèles ou « NoSQL » (bases documentaires, paires clefs/valeurs, etc.)

Types de données et longueur

Exemple avec MySQL :

Nombre entier entre -2 milliards et +2 milliards : 4 octets

Nombre en virgule flottante, 7 chiffres significatifs : 4 octets

Nombre en virgule flottante, 65 chiffres significatifs : 8 octets

Date simple : 3 octets

Date avec heures, minutes et secondes : 5 octets

Date avec précision au millièème de seconde : 8 octets

Texte de taille variable limité à 10 caractères : 11 octets

Texte de taille variable limité à 255 caractères : 256 octets

Principes de modélisation

Présentation du modèle relationnel

- Modèle proposé en 1970 par Edgar « Ted » Codd, d'IBM. En 1976, le langage SEQUEL (renommé **SQL** en 1982) permet de décrire la plupart des opérations d'une BDD relationnelle, encore théorique. Le premier **SGBD** relationnel est finalement Oracle, publié en 1979 par un concurrent d'IBM.
- Fondé sur les **mathématiques ensemblistes** et la **logique des prédicats**.
- Postule l'**indépendance entre niveau logique et niveau physique** : le modèle relationnel décrit l'aspect logique du stockage et de l'exploitation des données ; l'implémentation physique (enregistrement des données sur un support physique) est indifférente.
- Les données sont stockées dans des **tables**, liées entre elles par des **relations** (*relationships*).
- La modélisation a un double objectif : assurer l'**intégrité** des données et minimiser la taille occupée par les données. Un point essentiel pour ces deux objectifs est de lutter contre la **redondance** (*redundancy*).
- **Normalisation** : série de règles permettant d'améliorer la modélisation, notamment en éliminant toute redondance ; il existe plusieurs niveaux de normalisation : *1NF* à *5NF*, *BCNF*, *DKNF*.

Organisation d'une table


Exemple d'une table de livres : chaque ligne correspond à un livre, dont les différents attributs (titre, titre original...) sont stockés dans les colonnes correspondantes.

Colonne (*Column*) : parfois appelée, à tort, champ (*field*)

Ligne (*Row*) :
également appelée
enregistrement
(*record*) ou tuple

**Les colonnes ont un nom
et un type (numérique,
texte, date, etc.). Leur
ordre est indifférent.**

**L'ordre des lignes est
indifférent ; les lignes
sont généralement
distinguées par
l'utilisation d'une clé...**



Titre	TitreVO	CodeType	CodeGenre
Main Courante		NOUVS	Pol
Zapping		NOUVS	Pol
			SF
L'homme dont toutes les dents	The man whose teeth were	ROM	SF
Le gorille se mange froid		ROM	Pol
Les Rose-Croix		RECH	SH
Initiation, rites, sociétés secrètes		ARTIC	SH
Dictionnaire des religions		DICO	SH
Glamorama	Glamorama	ROM	Pol
L'art de la guerre par l'exemple		DICO	SH
Ciao, Bella, ciao !		ROM	Pol
Les Hauts Vents		ROM	Pol
Guerres	The wars	ROM	Litt
Les lions du Panshir	Lions down with lions	ROM	Pol
Les saveurs du thé		MAN	Cu
Cinq leçons sur la psychanalyse	Uebung Psychoanalyse, fünf	RECH	SH
Aguigui Mouna		BIO	Cu
Intermezzo		TH	Litt
Milosevic, la diagonale du fou		ESS	SH

Modélisation d'une relation un-à-plusieurs

Exemple : un auteur par livre, plusieurs livres par auteur

Répartition des infos entre tables

Mélanger dans la même table des entités du monde réel qui ne sont pas liés par une relation 1 à 1 provoque des redondances.

Titre	TitreVO	Nom	Prénom
Milosevic, la diagonale du fou		Hartmann	Florence
Dans le nu de la vie		Hatzfeld	Jean
Blaireau se cache	Hunting badger	Hillerman	Tony
Les clowns sacrés	Sacred Clowns	Hillerman	Tony
L'historien engagé		Hobsbawm	Eric
Vivre fatigue		Izzo	Jean-Claude
Une folie meurtrière	A mind to murder	James	Phyllis Dorothy
Meurtre dans un fauteuil	The black tower	James	P.D.
Ubu roi		Jarry	Alfred
Le pauvre nouveau est arrivé !		Jonquet	Thierry
L'impérialisme macédonien et l'hel		Jouquet	Pierre

Les redondances gâchent de l'espace de stockage (ex. 1, Hillerman). Surtout, elles sont la porte ouverte aux discordances (ex. 2, James).

Répartition des infos entre tables (2)

Chaque entité du monde réel doit être modélisée par une table spécifique. Un lien de type « un à plusieurs » (*one-to-many*) se modélise par la référence à une clé :

Table LIVRES

Titre	TitreVO	IDAuteur
Milosevic, la diagonale du fou		35
Dans le nu de la vie		64
Blaireau se cache	Hunting badger	125
Les clowns sacrés	Sacred Clowns	125
L'historien engagé		33
Vivre fatigue		70
Une folie meurtrière	A mind to murder	101
Meurtre dans un fauteuil	The black tower	101
Ubu roi		8
Le pauvre nouveau est arrivé !		121
L'impérialisme macédonien et l'he		102

Table AUTEURS

IDAuteur	Nom	Prénom
35	Hartmann	Florence
64	Hatzfeld	Jean
125	Hillerman	Tony
33	Hobsbawm	Eric
70	Izzo	Jean-Claude
101	James	Phyllis Doroth
8	Jarry	Alfred
121	Jonquet	Thierry
102	Jouquet	Pierre

Dans cet exemple, on a doté la table AUTEURS d'une clé sous forme de numéro arbitraire. Chaque livre contient la référence à son auteur sous forme, et deux livres peuvent avoir le même auteur sans redondance.

Comme il n'y a pas de redondance, une correction sur un auteur n'a besoin d'être effectuée qu'une seule fois, quel que soit le nombre de livres de cet auteur.

Les clés

- **Définition** : une clé est une colonne ou une combinaison de colonnes permettant d'identifier sans ambiguïté chaque ligne d'une table.
- La qualité essentielle d'une clé est l'**unicité** (*unicity*), c'àd que chacune de ses valeurs ne peut apparaître qu'une fois sur table dont elle est la clé. Les qualités souhaitables sont la **concision** et la **stabilité**.
- Les diverses clés possibles d'une table sont nommées **clés candidates** (*candidate keys*). La clé choisie par le concepteur de la BDD est nommée **clé primaire** (*primary key*) de la table.
- Une colonne (ou une combinaison de colonnes) faisant référence à la clé d'une autre table (par ex., l'IDauteur de la table LIVRES) est une **clé étrangère**. On dit que **la clé étrangère référence la clé primaire**.
- On parle de **clé naturelle** (*natural key*) si la clé reflète des informations existant dans le monde réelle (par ex., la combinaison Titre-IDauteur sur la table LIVRES), et de **clé artificielle** (*artificial key*) si la clé n'existe pas dans le monde réel. Les clés artificielles sont généralement **arbitraires** (*surrogate*), c'àd sans aucune signification.
- Le plus souvent, on choisit comme clé primaire une clé artificielle arbitraire, sous forme de **numéro automatique** (*autonumber*)

Modélisation d'une relation plusieurs-à-plusieurs

Exemple : plusieurs auteurs par livre, plusieurs livres par auteur

Solutions à éviter...

Nous avons vu qu'une relation un-à-plusieurs entre tables se traduit par l'utilisation côté "plusieurs" d'une clé étrangère référençant la clé primaire de la table côté "un". Comment faire avec une relation plusieurs-à-plusieurs ?

Titre	TitreVO	IDAuteur
Le choix de Gorbatchev		69
Le journalisme sans peine		96
Le journalisme sans peine		42
Les contes du whisky		65

Doublonner le livre revient évidemment à introduire une redondance fâcheuse...

Titre	TitreVO	IDAuteur1	IDAuteur2
Le choix de Gorbatchev		69	
Le journalisme sans peine		96	42
Les contes du whisky		65	

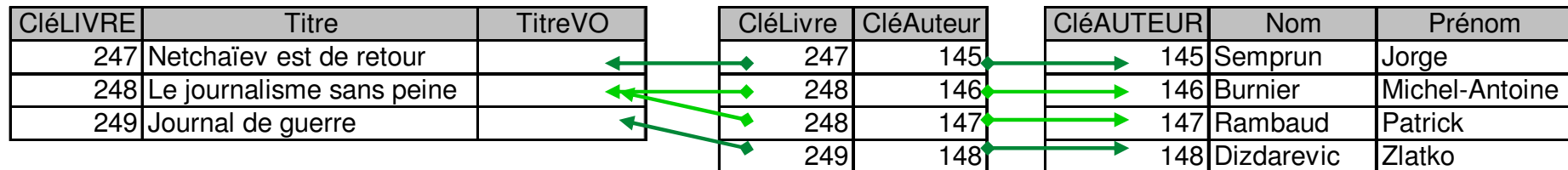
Créer deux colonnes IDAuteur permet de modéliser une relation 1-2, mais pas 1-n...

Titre	TitreVO	IDAuteur
Le choix de Gorbatchev		69
Le journalisme sans peine		96,42
Les contes du whisky		65

Il existe des solutions techniques pour faire fonctionner ceci, mais c'est contraire au principe d'atomicité.

Solution normalisée

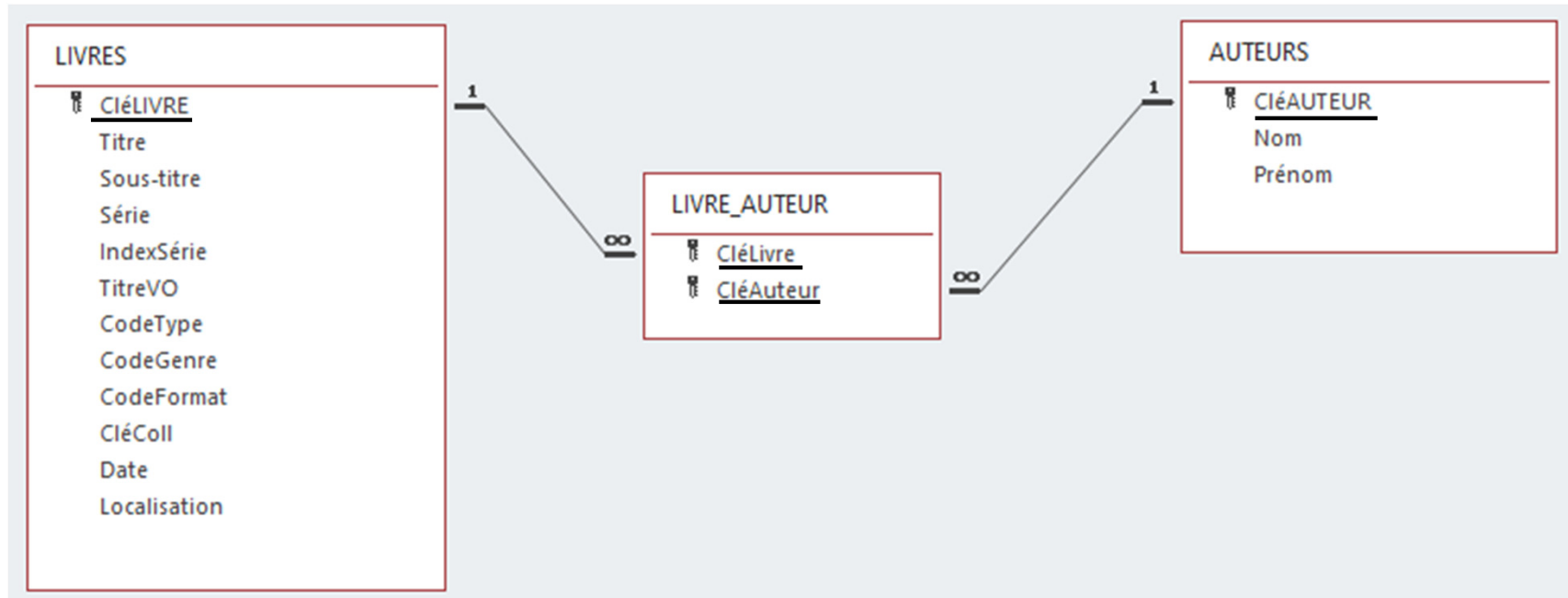
Une relation plusieurs-à-plusieurs entre deux tables se normalise grâce à une troisième table, qui joue le rôle de "charnière". Chacune des deux tables d'origine est liée à la charnière par une relations un-à-plusieurs.



Les avantages de cette solution sont nombreux :

- pas de redondance
- pas de limite artificielle sur le nombre d'auteurs
- jointures classiques de type *clé étrangère* = *clé primaire*

Modèle physique de données



Le modèle physique représente l'organisation réelle de la base de données.

On y retrouve :

- Les tables
- Les colonnes
- Les clefs primaires
- Les relations

Synthèse

Concepts à retenir

Principes

- Indépendance des niveaux logique et physique
- Cohérence
- Normalisation
 - Atomicité
 - Cardinalité
 - Non-redondance

Opérations

- Projection (choix des colonnes)
- Filtrage (conditions sur les lignes)
- Tri (ordre de présentation des lignes)
- Calcul
- Dédoublonnage (DISTINCT)
- Regroupement / Agrégation

Objets

- Objets de base de données :
 - Tables
 - Colonnes (fields)
 - Lignes (rows / records / tuples)
 - Numéro auto (séquenceur)
 - Clé primaire (PK)
 - Clé étrangère (FK)
- Objets hors base de données :
 - Requêtes
 - Formulaires