

# Évolution de schémas dans les entrepôts de données mise à jour de hiérarchies de dimension pour la personnalisation des analyses

Thèse présentée par Cécile FAVRE  
pour obtenir le titre de Docteur en Informatique  
sous la direction d'Omar Boussaid et de Fadila Bentayeb

Université de Lyon (Lumière Lyon 2) – EDIIS – Laboratoire ERIC

12 décembre 2007

# Contexte

## Contexte de travail

- Thèse CIFRE en collaboration avec LCL-Le Crédit Lyonnais



Problématiques de recherche

- Structuration,  
acquisition et stockage  
des demandes de  
marketing local

- Analyse

- Personnalisation des  
analyses

- Évolution de schéma  
dans les entrepôts de  
données

Terrain d'application

# Problèmes

## Stratégies de conception du schéma de l'entrepôt de données

- Guidées par les données
- Guidées par les besoins
- Hybrides

## Évolution du schéma de l'entrepôt de données

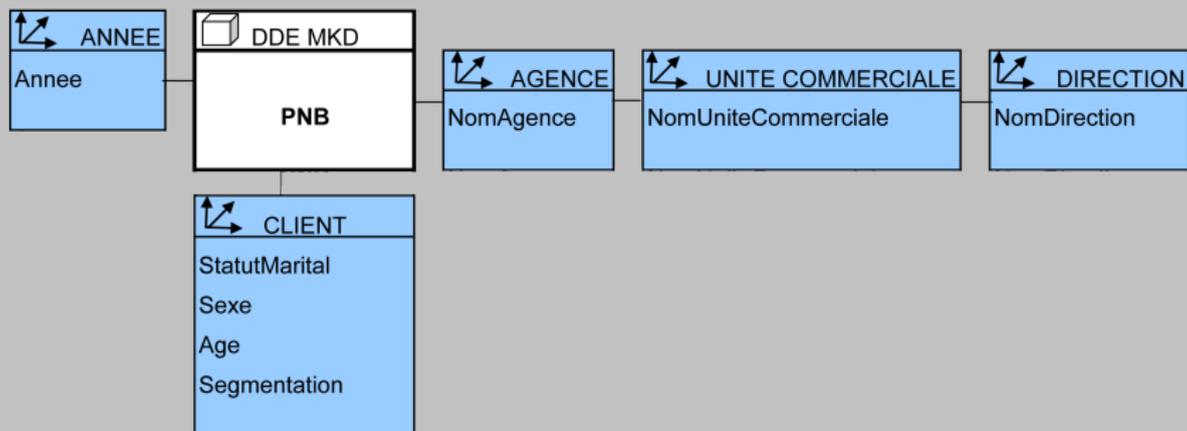
- Évolution des données
- Évolution des besoins

## Cas LCL

- Besoins d'analyse individuels émergents

# Exemple

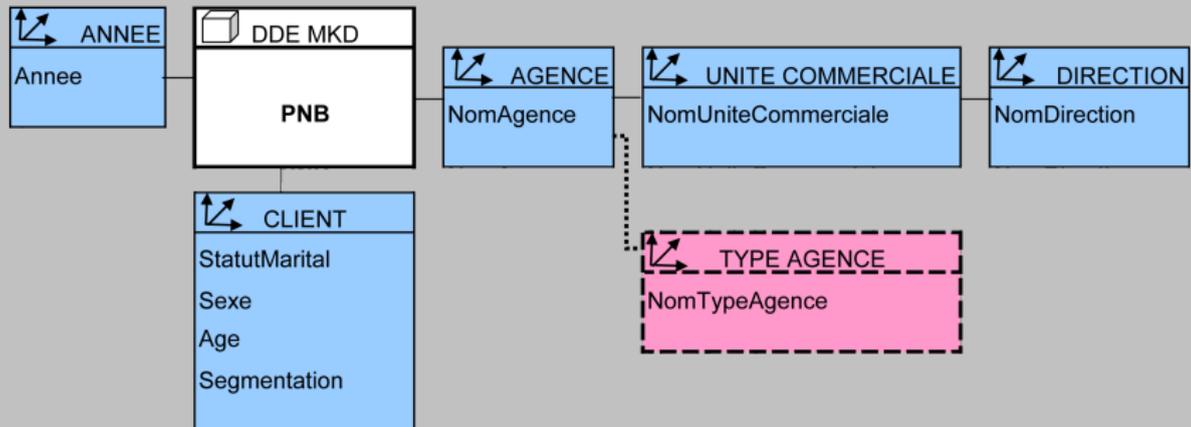
## Schéma d'entrepôt de données du LCL : LCL-DW



## Analyse du PNB

# Exemple

Besoin : analyse selon type d'agence



Analyse du PNB

# Proposition

## Personnalisation des analyses

- Objectif : nouvelles analyses personnalisées (nouveaux «roll-up» )
- Moyen : évolution de schéma de l'entrepôt de données (entrepôt de données évolutif)
- Idée : exploitation des connaissances utilisateurs

# Plan

## Plan

- ① État de l'art
  - Personnalisation
  - Évolution de schéma dans les entrepôts de données
- ② Contributions
  - Entrepôt de données évolutif
    - Modèle formel d'entrepôt de données évolutif
    - Architecture globale
    - Modèle d'exécution
  - Performances
    - Évolution incrémentale de la charge
  - Validation logicielle
- ③ Conclusion et perspectives

# Personnalisation

## Contexte

- Placer l'utilisateur au centre d'un système

## Rôle de la personnalisation

En fonction de l'utilisateur, de son profil :

- RI : affiner les résultats de la recherche
- BD : affiner les requêtes
- IHM : adapter l'interface

# Personnalisation dans les ED

## Approches émergentes

- Affinement de requêtes (Bellatreche et al., 2005 et 2006)
- Personnalisation dans la navigation (Ravat et al., 2007)
- Gestion d'exceptions dans le processus d'agrégation (Espil et al., 2001 et 2002)

# Évolution de schéma dans les entrepôts de données (1/3)

## Deux stratégies pour la gestion de l'évolution de schéma

- Mise à jour de schéma
- Modélisation temporelle

## Évolution de schéma dans les entrepôts de données (2/3)

### Mise à jour de schéma

- Opérateurs d'évolution (Hurtado et al., 1999 ; Blaschka et al., 1999 ; Benitez-Guerrero, 2002)
- Enrichissement de hiérarchies de dimension (Mazon et al., 2006)
- Maintenance de vues matérialisées (Bellahsene, 2002)

# Évolution de schéma dans les entrepôts de données (3/3)

## Modélisation temporelle

Recours à des labels temporels :

- Instances (Bliujute et al., 1998)
- Liens d'agrégation (Mendelzon et al., 2000)
- Versions (Eder et al., 2000 ; Body et al., 2002 et 2003 ; Bebel et al., 2004 ; Morzy et al., 2004 ; Golfarelli et al., 2006 ; Ravat et al., 2006 ...)

# Étude comparative (1/2)

## Critères de comparaison

- Caractéristiques intrinsèques :
  - historisation des dimensions
  - cohérence des analyses
  - approche orientée utilisateurs
- Application :
  - nécessité d'implémenter la solution lors de la conception
  - complexité de la mise en œuvre (analyse, chargement)
- Performances :
  - stockage
  - temps de réponse aux analyses

# Étude comparative (2/2)

		Modélisation temporelle			Mise à jour modèle			
		Instances	Liens d'agrégation	Versions	Opérateurs d'évolution	Enrichissement hiérarchies	Maintenance de vues	Personnalisation des analyses
Caractéristiques intrinsèques	historisation des dimensions	+	+	+	-	-	-	-
	cohérence des analyses	+	+	+	-	+	-	+
	approche orientée utilisateurs	-	-	+ / -	-	-	-	+
Application	dès la conception	-	-	-	+	+	+	+
	complexité	-	-	-	+	+	+	+
Performances	stockage	-	-	-	+	+	+	+
	temps de réponse analyses	-	-	-	+	+	+	+

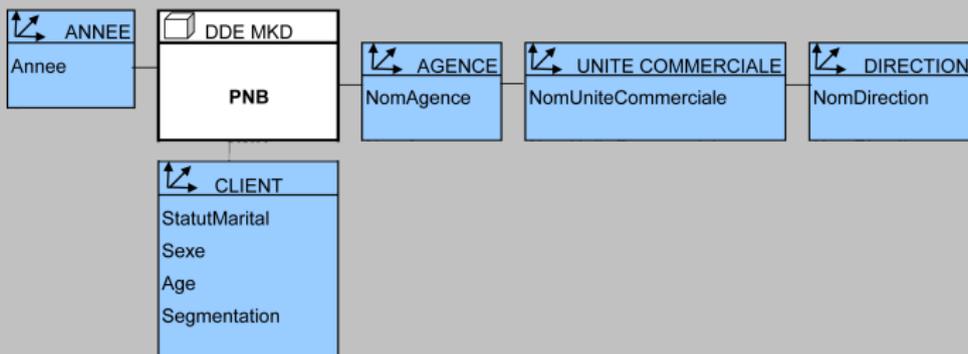
Tableau comparatif des approches

# Plan

## Plan

- ① État de l'art
  - Personnalisation
  - Évolution de schéma dans les entrepôts de données
- ② Contributions
  - Entrepôt de données évolutif
    - Modèle formel d'entrepôt de données évolutif
    - Architecture globale
    - Modèle d'exécution
  - Performances
    - Évolution incrémentale de la charge
  - Validation logicielle
- ③ Conclusion et perspectives

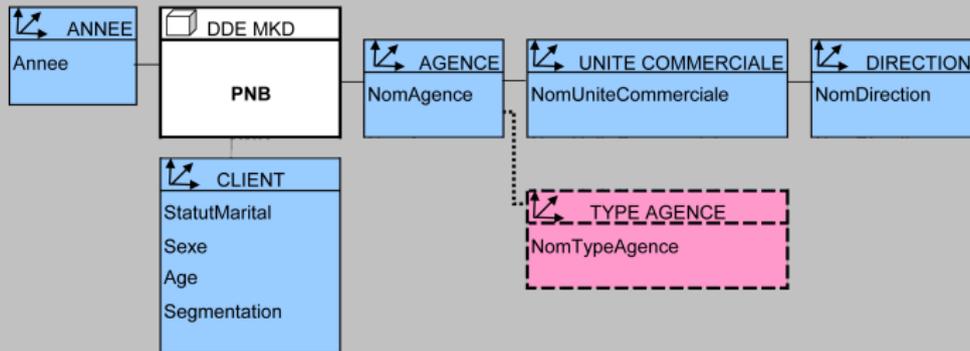
## Exemple illustratif LCL-DW



### Analyse du PNB

**AGENCE** ——— **UNITE COMMERCIALE** ——— **DIRECTION**

## Exemple illustratif LCL-DW



### Analyse du PNB



## Modèle R-DW (1/3)

### Modèle formel d'entrepôt de données évolutif

Modèle R-DW («Rule-based Data Warehouse»), basé sur des règles d'agrégation

- Partie fixe : table des faits et tables de dimension directement liées
- Partie évolutive : hiérarchies de dimension strictes

### Règles d'agrégation

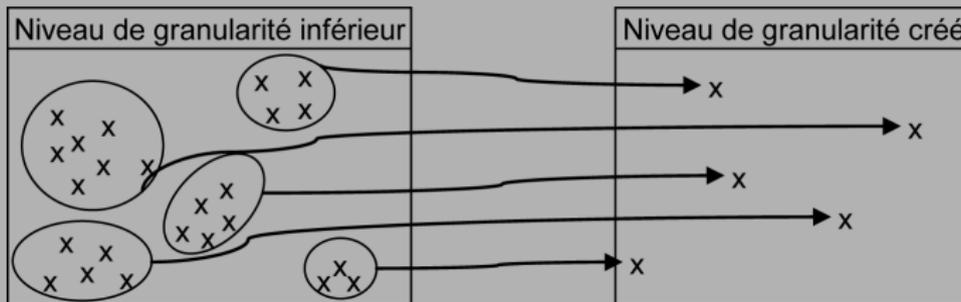
- Utilisation
  - Création de niveaux de hiérarchie
  - Représentation des connaissances utilisateurs
- Définition : *si cond1  $\wedge$  cond2  $\wedge$  ... alors att1 = val1  $\wedge$  ...*
- Contraintes : concept de partition

## Modèle R-DW (2/3)

### Règles d'agrégation

Un ensemble de règles pour définir un niveau de hiérarchie :

- Clauses «si» : conditions définissant les classes de la partition
- Clauses «alors» : définition des instances du niveau créé et du lien d'agrégation



# Modèle R-DW (3/3)

## Règles d'agrégation : exemple

AGENCE		TYPE AGENCE	
AgenceID	...	NomTypeAgence	CodeTypeAgence
01903		étudiants	ETU
01905		non résidents	NR
02256		classique	CLA
01929			
01000			
...			

r1 : si AgenceID IN {'01903','01905','02256'}

alors NomTypeAgence='étudiants' et CodeTypeAgence='ETU'

r2 : si AgenceID = '01929'

alors NomTypeAgence='non résidents' et CodeTypeAgence='NR'

r3 : si AgenceID NOT IN {'01903','01905','02256','01929'}

alors NomTypeAgence='étudiants' et CodeTypeAgence='ETU'

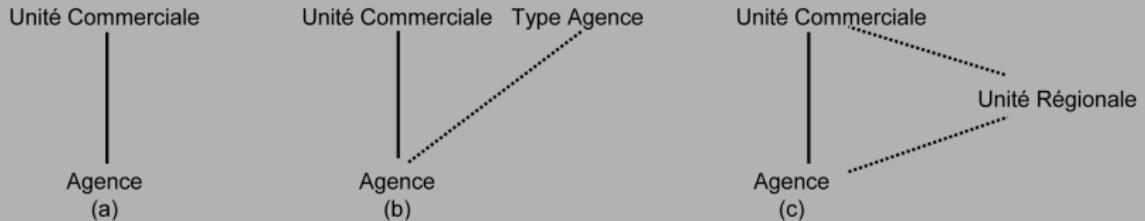
# Mise à jour de hiérarchies de dimension (1/3)

## Mise à jour des hiérarchies de dimension

- Création et suppression de niveaux de hiérarchie
- Propagation de ces mises à jour

## Mise à jour de hiérarchie de dimension (2/3)

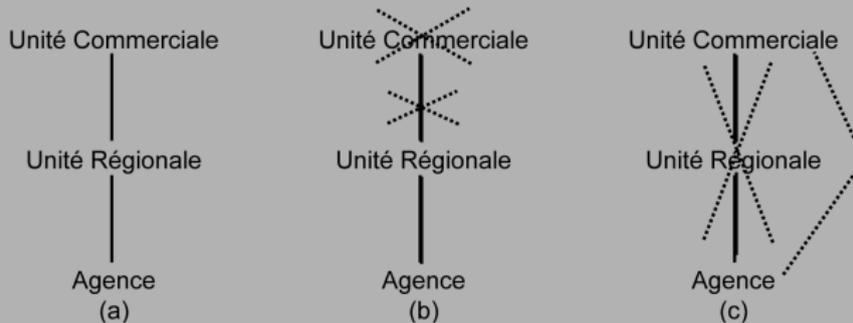
### Création de niveau de hiérarchie et propagation



Scénarios de création de niveau de hiérarchie

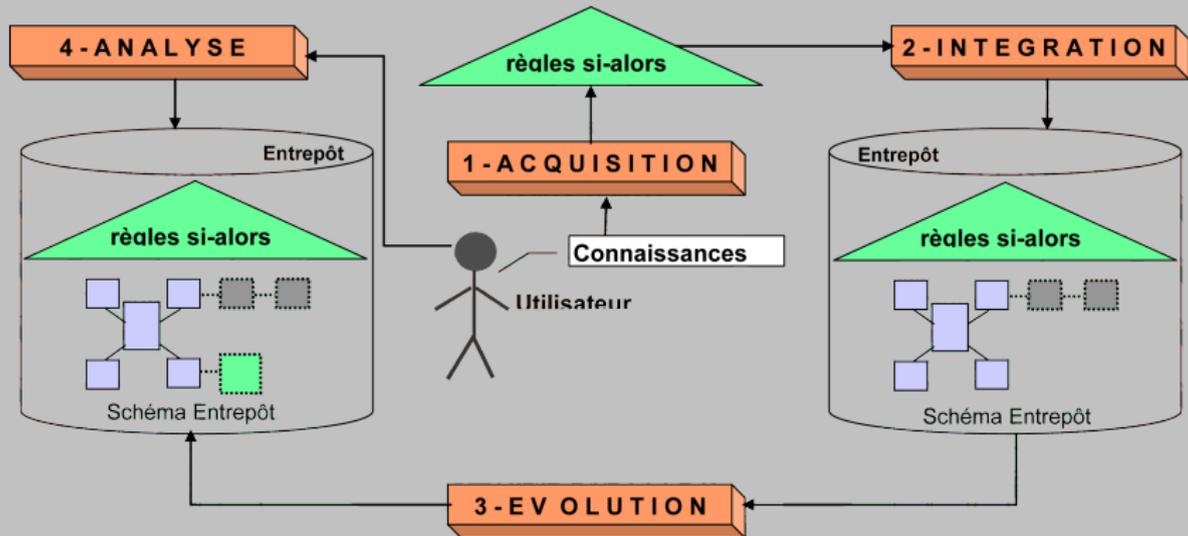
## Mise à jour de hiérarchie de dimension (3/3)

### Suppression de niveau de hiérarchie et propagation



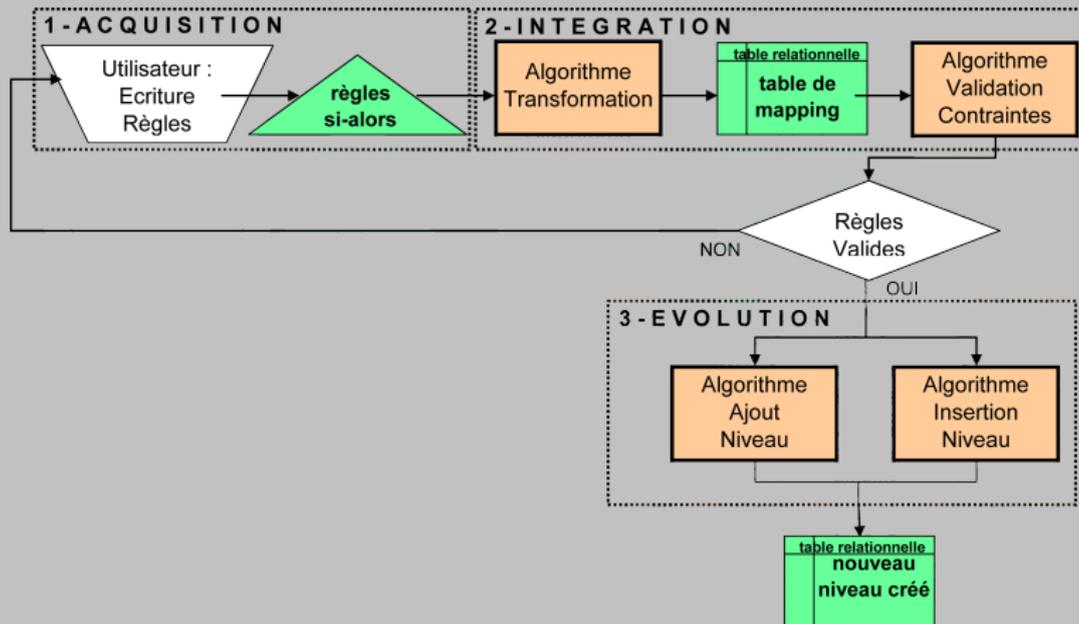
Scénarios de suppression de niveau de hiérarchie

# Architecture



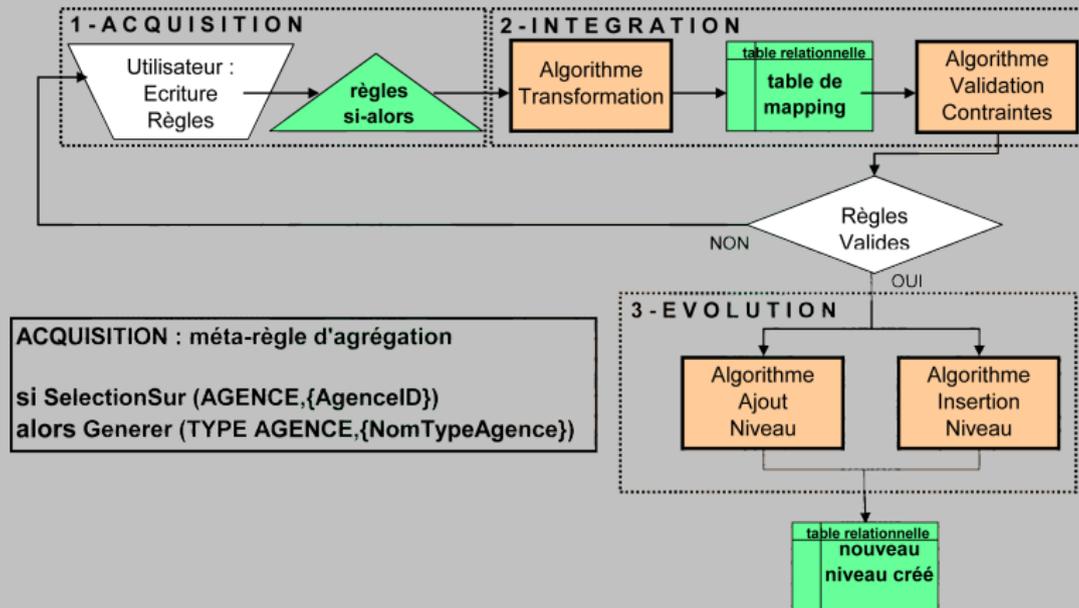
Architecture d'évolution de schéma guidée par les utilisateurs

# Modèle d'exécution



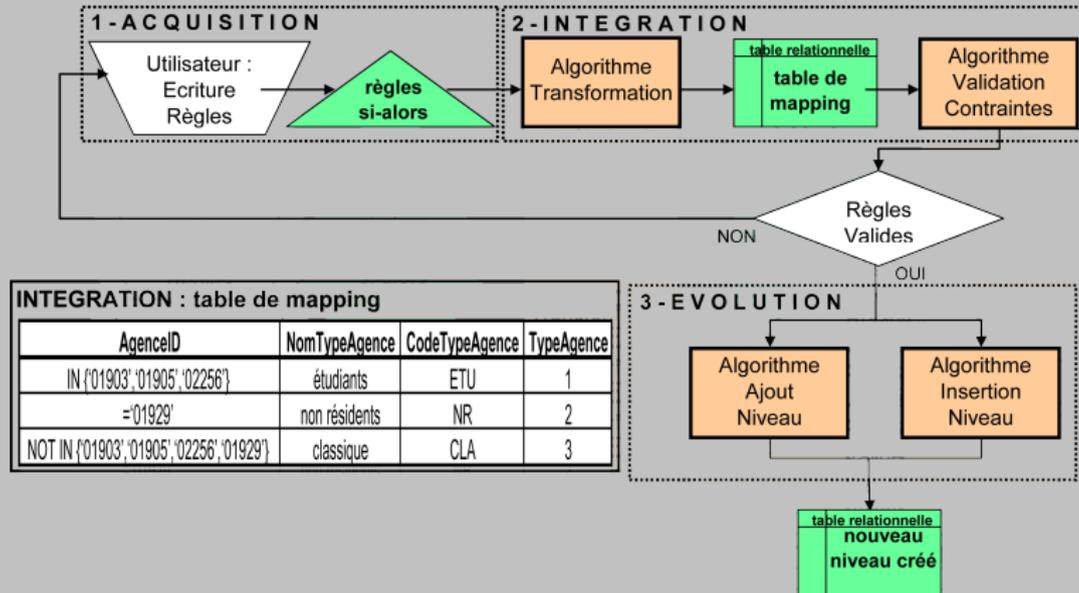
Modèle d'exécution dans le contexte ROLAP

# Modèle d'exécution



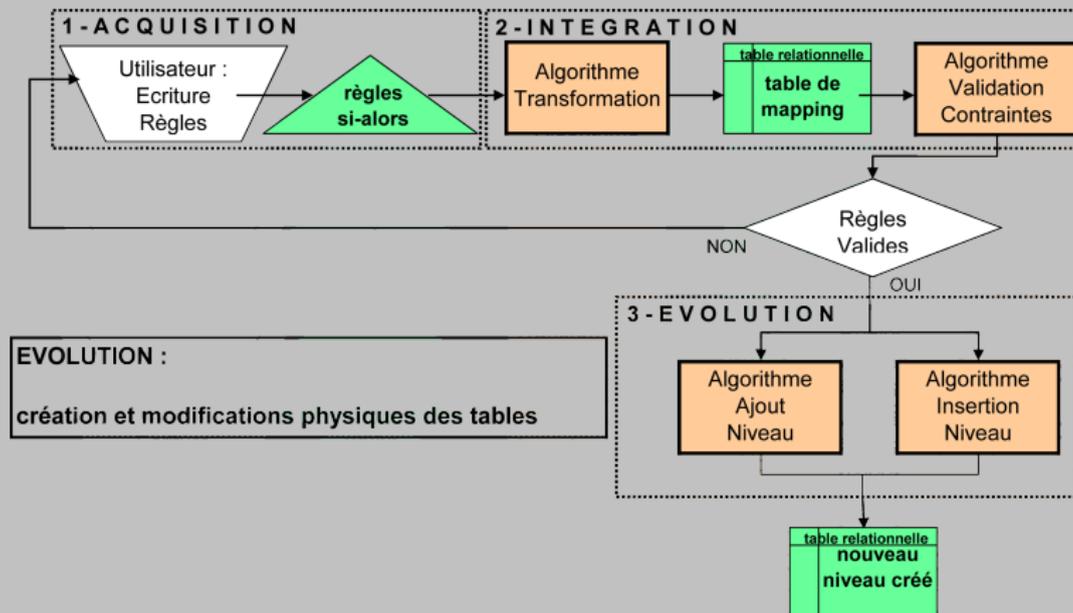
Modèle d'exécution dans le contexte ROLAP

# Modèle d'exécution



Modèle d'exécution dans le contexte ROLAP

# Modèle d'exécution



Modèle d'exécution dans le contexte ROLAP

# Résultat

Exemple : roll up de AGENCE vers TYPE AGENCE :

```
SELECT NomTypeAgence, NomAgence, SUM(PNB)
FROM DDE_MKD, AGENCE, TYPE_AGENCE
WHERE DDE_MKD.AgenceID=AGENCE.AgenceID AND
AGENCE.TypeAgenceID=TYPE_AGENCE.TypeAgenceID
GROUP BY ROLLUP (NomAgence, NomTypeAgence)
```

NomTypeAgence	NomAgence	PNB (euros)
NULL	NULL	23000
<b>classique</b>	<b>NULL</b>	<b>12000</b>
classique	Annonay	500
classique	Lyon Cordeliers	3000
classique	...	...
<b>étudiants</b>	<b>NULL</b>	<b>3000</b>
étudiants	La Doua	1500
étudiants	...	...
<b>non résidents</b>	<b>NULL</b>	<b>8000</b>
non résidents	Lyon NR	4000
non résidents	...	...

# Plan

## Plan

- ① État de l'art
  - Personnalisation
  - Évolution de schéma dans les entrepôts de données
- ② Contributions
  - Entrepôt de données évolutif
    - Modèle formel d'entrepôt de données évolutif
    - Architecture globale
    - Modèle d'exécution
  - Performances
    - Évolution incrémentale de la charge
  - Validation logicielle
- ③ Conclusion et perspectives

# Évolution incrémentale de la charge

## Contexte

- Évaluation des performances grâce à une charge

## Utilisation de charge

- Adapter les performances du système en fonction de son utilisation (Gallo, 2002) : sélection des structures d'optimisation basée sur une charge (Aouiche et al., 2006)

## Problème

- Cohérence de la charge après l'évolution de schéma ?

## Proposition

- Évolution incrémentale de la charge

## État de l'art (1/2)

### Évolution de charge

- Changement dans les probabilités des requêtes (Lawrence et al., 2006)
- Ajout de nouvelles requêtes (Theodoratos et al., 2000)

=> Supposition : requêtes existantes toujours correctes

## État de l'art (2/2)

### Évolution de requêtes

- Bases de données :
  - Problème : applications liées aux bases de données qui évoluent
  - Modèle à base de graphe annoté par l'administrateur (Papastefanatos et al., 2005 et 2006)
- Entrepôts de données :
  - Non considéré jusque là comme un problème à part entière
  - Traitement à travers la maintenance des vues (Bellahsene, 2002)
  - Problème : utilisation multiple des requêtes dans une architecture décisionnelle (états d'analyse, tests de performances, ...)

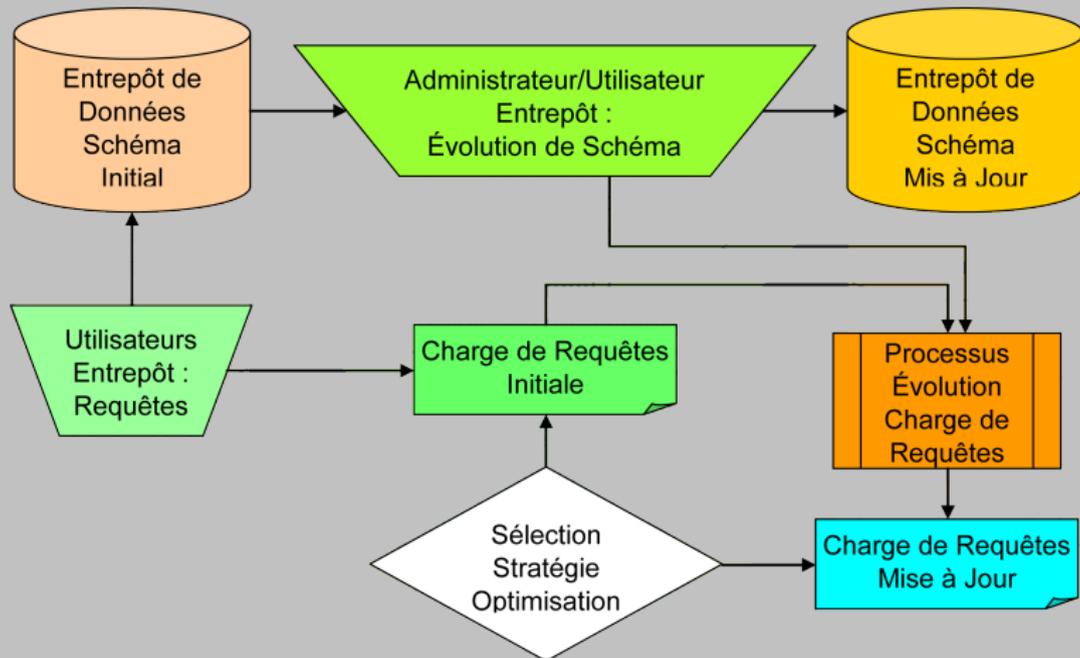
# Proposition

## Évolution incrémentale de la charge

Suite aux évolutions du schéma :

- Maintenir cohérentes les requêtes existantes
- Créer de nouvelles requêtes

# Architecture

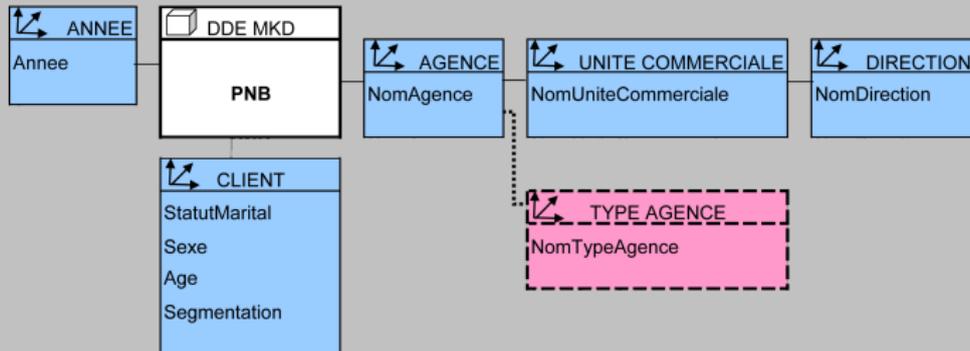


Principe de l'évolution incrémentale d'une charge

# Typologie des changements

Rôle dans le schéma	Type	Opération	Requête à modifier	Requête à supprimer	Requête à créer
Dimension et Niveau	Table	Création			OUI
Dimension et Niveau	Table	Suppression	OUI	OUI	
Dimension et Niveau	Table	Mise à jour	OUI		
Fait	Table	Modification	OUI		
Mesure	Attribut	Ajout			OUI
Mesure	Attribut	Suppression	OUI	OUI	
Mesure	Attribut	Modification	OUI		
Descripteur Dimension	Attribut	Ajout			OUI
Descripteur Dimension	Attribut	Suppression	OUI	OUI	
Descripteur Dimension	Attribut	Modification	OUI		

# Exemple



## Charge enrichie

Nouvelle requête :

```
SELECT NomTypeAgence, NomAgence, SUM(PNB)
FROM DDE_MKD, AGENCE, TYPE_AGENCE
WHERE DDE_MKD.AgenceID=AGENCE.AgenceID AND
AGENCE.TypeAgenceID=TYPE_AGENCE.TypeAgenceID
GROUP BY ROLLUP (NomAgence, NomTypeAgence)
```

# Plan

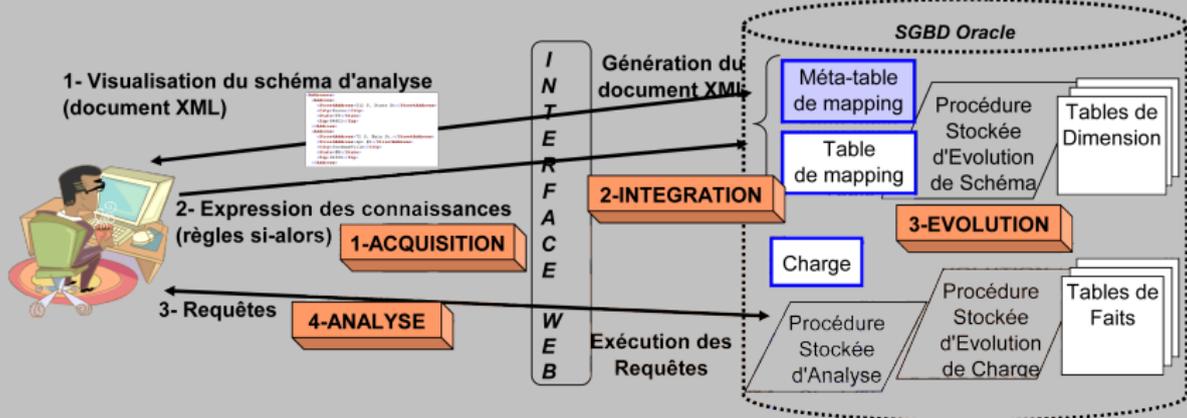
## Plan

- ① État de l'art
  - Personnalisation
  - Évolution de schéma dans les entrepôts de données
- ② Contributions
  - Entrepôt de données évolutif
    - Modèle formel d'entrepôt de données évolutif
    - Architecture globale
    - Modèle d'exécution
  - Performances
    - Évolution incrémentale de la charge
  - Validation logicielle
- ③ Conclusion et perspectives

# Validation logicielle

## Plateforme WEDriK (data Warehouse Evolution Driven by Knowledge)

- Entrepôt stocké sous Oracle 10g
- Interfaces Web centrées utilisateur
- Modules acquisition, intégration, évolution, analyse



# Conclusion

## Contributions

Personnalisation des analyses dans les entrepôts de données :

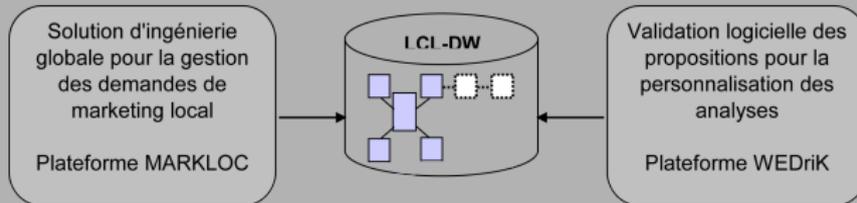
- Modèle formel d'entrepôt évolutif R-DW
- Architecture d'entrepôt de données évolutif centrée utilisateur
- Modèle d'exécution ROLAP de R-DW

Vers l'évaluation de performances des entrepôts de données évolutifs :

- Mise à jour incrémentale de charge

# Conclusion

## Contributions



# Perspectives

## Évolution des règles

- Gérer l'évolution des règles en répercussion de celle des données, des règles de gestion, etc.

## Découverte automatique de règles d'agrégation

- Utilisation d'un algorithme d'apprentissage non supervisé

## Partage des nouvelles possibilités d'analyse

- Processus d'annotation

## Gestion des performances dans un contexte évolutif

- Utilisation de l'évolution incrémentale de la charge pour l'étude de performances

# Valorisation scientifique

<b>Articles dans une revue internationale :</b>	<b>2</b>
ICAIE 2008, IJBIS 2007	
<b>Chapitre dans un ouvrage d'audience internationale :</b>	<b>1</b>
Encyclopedia of Database Technologies and Applications, 2008	
<b>Publications dans une conférence internationale :</b>	<b>5</b>
DaWaK 07, ICEIS 07, CE 06, BNCOD 06, ISMIS 05	
<b>Publications internationales :</b>	<b>8</b>
<b>Publications dans une conférence nationale :</b>	<b>6</b>
EDA 07, INFORSID 07, EGC 07, EDA 06, EDA 05, EGC 05	
<b>Ateliers avec comité de lecture :</b>	<b>3</b>
ASD 06, FDC 06, FDC 05	
<b>Publications nationales :</b>	<b>9</b>
<b>Total :</b>	<b>17</b>