Informatique Décisionnelle
Modélisation Multidimensionnelle

Med. AMNAI
Master spécialisé
BIG DATA Cloud Computing

Plan

Plan

Plan

Plan

Plan

Modélisation Entité/Association

- **Avantages**
  - Normalisation ;
    - Éliminer les redondances ;
    - Préserver la cohérence des données.
  - Optimisation des transactions ;
  - Réduction de l'espace de stockage ;
- **Inconvénients pour un utilisateur final**
  - Schéma trop complet : Contient des tables/champs inutiles pour l'analyse ;
  - Pas d'interface graphique capable de rendre utilisable le modèle E/A ;
  - Inadapté pour l'analyse.

Limites du modèle relationnel : OLTP vs OLAP

- **OLTP** : Requêtes simples "**qui, quoi**"
  - **ex.** les ventes de **X** ;
  - **jointures** : les ventes de **X** à quel prix de quel fournisseur.
- **OLAP** : besoin de données **agrégées**, synthétisées
  - nombre de ventes par vendeur, par région, par mois ;
  - nombre de ventes par vendeur, par fournisseur, par mois.
- **SQL** : Possibilité d'agréger les données (group by) :
  - **très coûteux** (parcourir toutes les tables) et il faut recalculer à chaque utilisation.

Nécessité d'une structure muti-dimensionnelle

- **Les BD relationnelles ne sont pas adaptées à l'OLAP car :**
  - **Pas les mêmes objectifs** ;
  - **Pas les mêmes données** ;
  - **Pas les mêmes traitements et requêtes.**
- **Il est donc nécessaire de disposer d'une structure de stockage adaptée à l'OLAP**, i.e. permettant de :
  - représenter les données dans plusieurs dimensions ;
  - manipuler les données facilement et efficacement.

Exemple d'un DW (Entrpôt de données)

- L'ED doit fournir le CA des ventes d'un produit, par date, client, magasin et vendeur, ainsi que toutes les sommations possibles de chiffre d'affaires dans une année donnée.

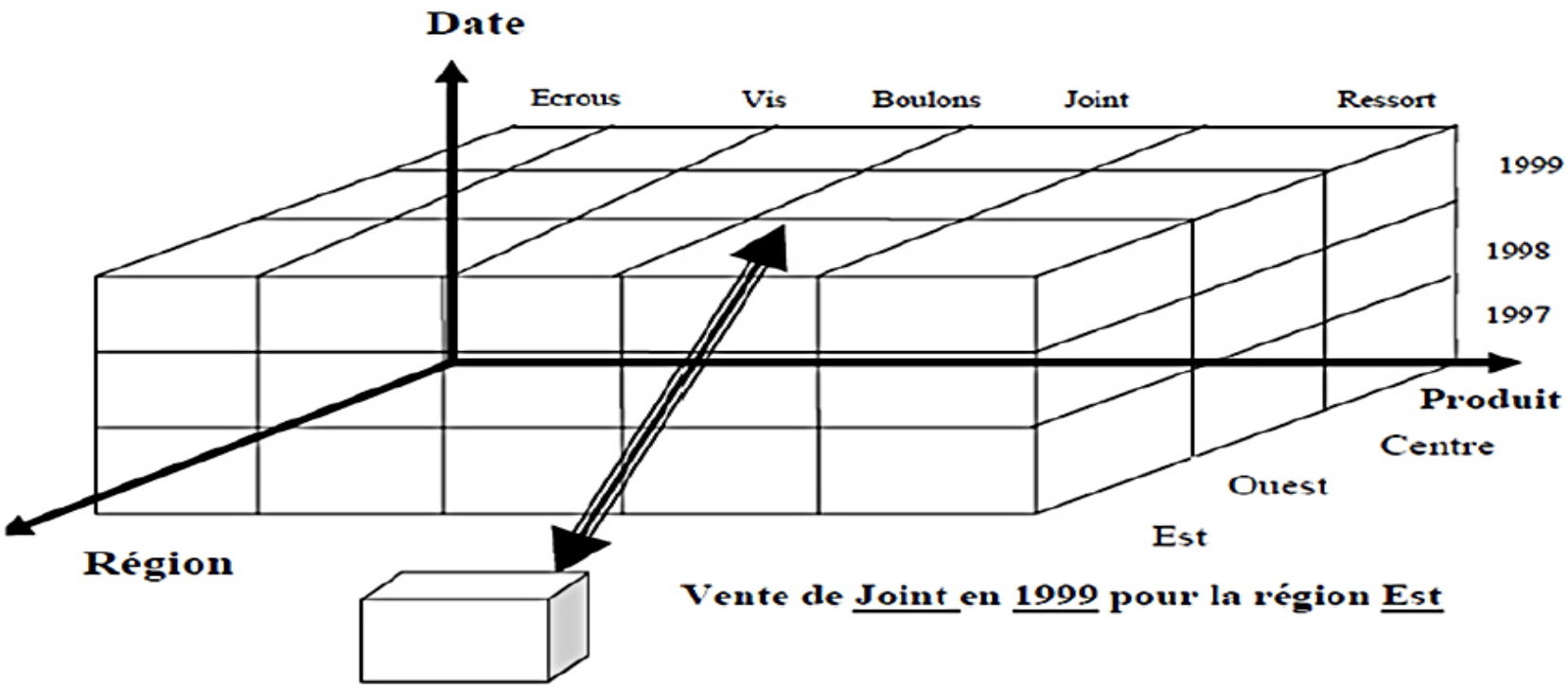
- Une vente est caractérisée par :

  ***Vente* (produit, client, magasin, date, pays, coût de vente, montant des ventes, Qté vendue)**

  - **Produit** : clé produit, Description (et libellés), . . .
  - **Client** : clé client, type client, . . .
  - **Magasin** : clé magasin, . . .
  - **Date** : clé temps, jour, semaine, mois, . . .

Représentation multidimensionnelle

Cube représentant le sujet des Ventes



Cocepts de schéma de modélisation

- Nouvelle méthode de conception autour des concepts métiers (Ne pas normaliser au maximum.)
- Notion de **cube** (en fonction des dimensions).
- Introduction de nouveaux types des tables :
  - Table de **faits (mesure)**.
  - Table de **dimensions**
- Introduction de nouveaux modèles :
  - **Modèle en étoile.**
  - **Modèle en flocon.**

Cube

- Modélisation multidimensionnelle des données facilitant l'analyse d'une quantité selon différentes dimensions :
  - Temps.
  - Localisation géographique ...
- Les calculs sont réalisés lors du chargement ou de la mise à jour du cube.

Manipulation du cube

- Opérateurs appliqués sur le cube sont algébriques (le résultat est un autre cube) et peuvent être combinés.
- Opérateurs :
  - **Slicing, Dicing (extraction)**.
  - Changement de la granularité d'une dimension
    - **Roll up** (agrégation d'une dimension => résumé).
    - **Drill down** (plus détaillées).

"Slicing" et "Dicing"

- **Slicing** : Sélection de tranches du cube par des prédicats selon une dimension
  - Filtrer une dimension selon une valeur.
  - Exemple : Slice (1998) : on ne retient que la partie du cube qui correspond à cette date.
- **Dicing** : extraction d'un sous-cube.
- **RQ : Slicing et Dicing Opérateurs sur le cube**

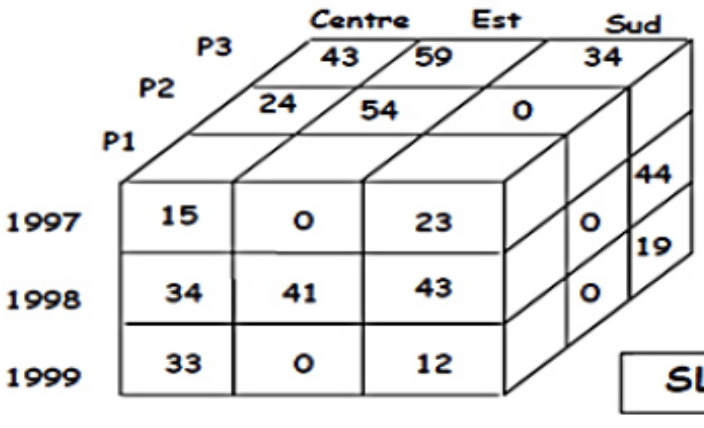
Exemple "Slicing"

Données d'une activité de Vente

Produit	Région	Vente	Période: 97
P1	Centre	15	
P2	Centre	24	
P3	Centre	43	
P2	Est	54	
P3	Est	59	
P1	Sud	23	
P3	Sud	34	

Produit	Région	Vente	Période: 98
P1	Centre	34	
P2	Centre	25	
P3	Centre	37	
P1	Est	41	
P3	Est	26	
P1	Sud	43	
P3	Sud	44	

Produit	Région	Vente	Période: 99
P1	Centre	33	
P2	Centre	15	
P3	Centre	26	
P2	Est	21	
P3	Est	35	
P1	Sud	12	
P3	Sud	19	



SLICE(1998)

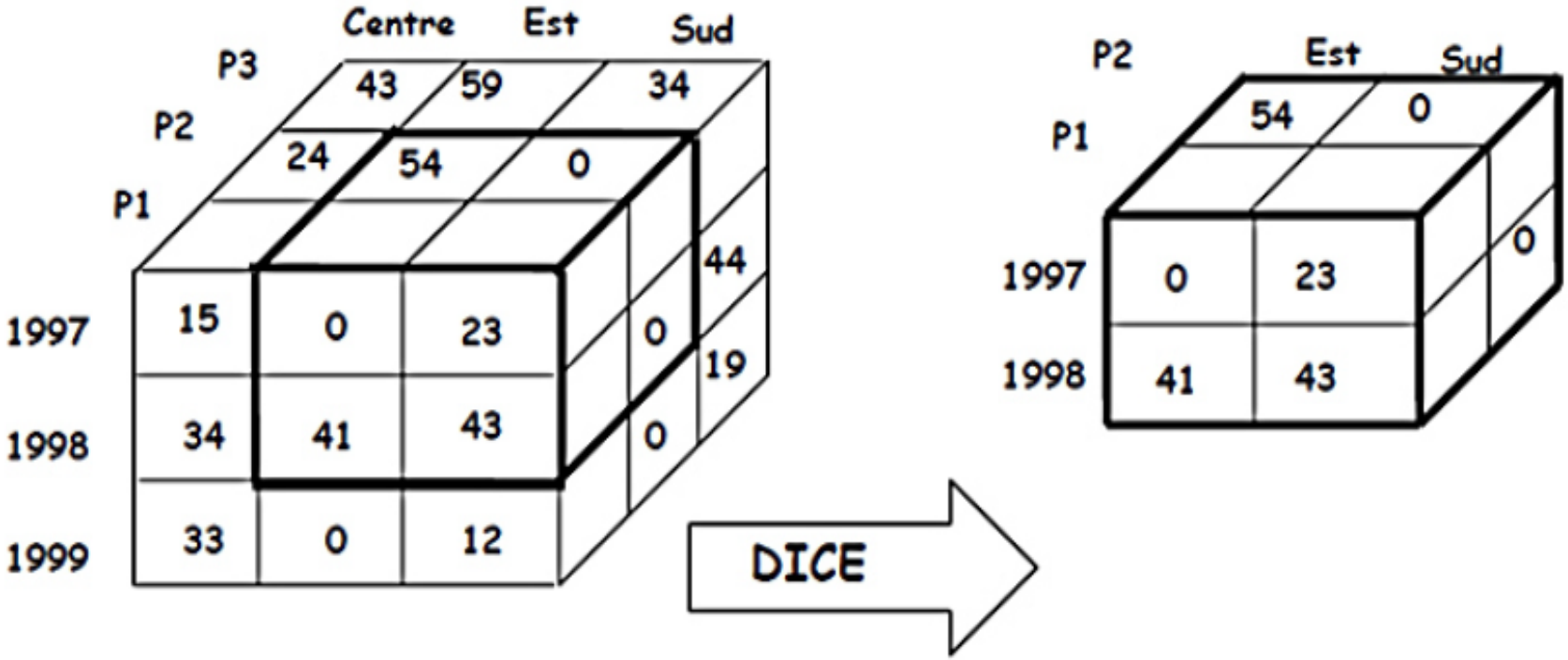
Année 1998	P1	P2	P3
Centre	34	25	37
Est	41	0	26
Sud	43	0	44

SLICE

L'opération SLICE sur un Cube

Exemple "Dicing"

Extraction d'un sous cube

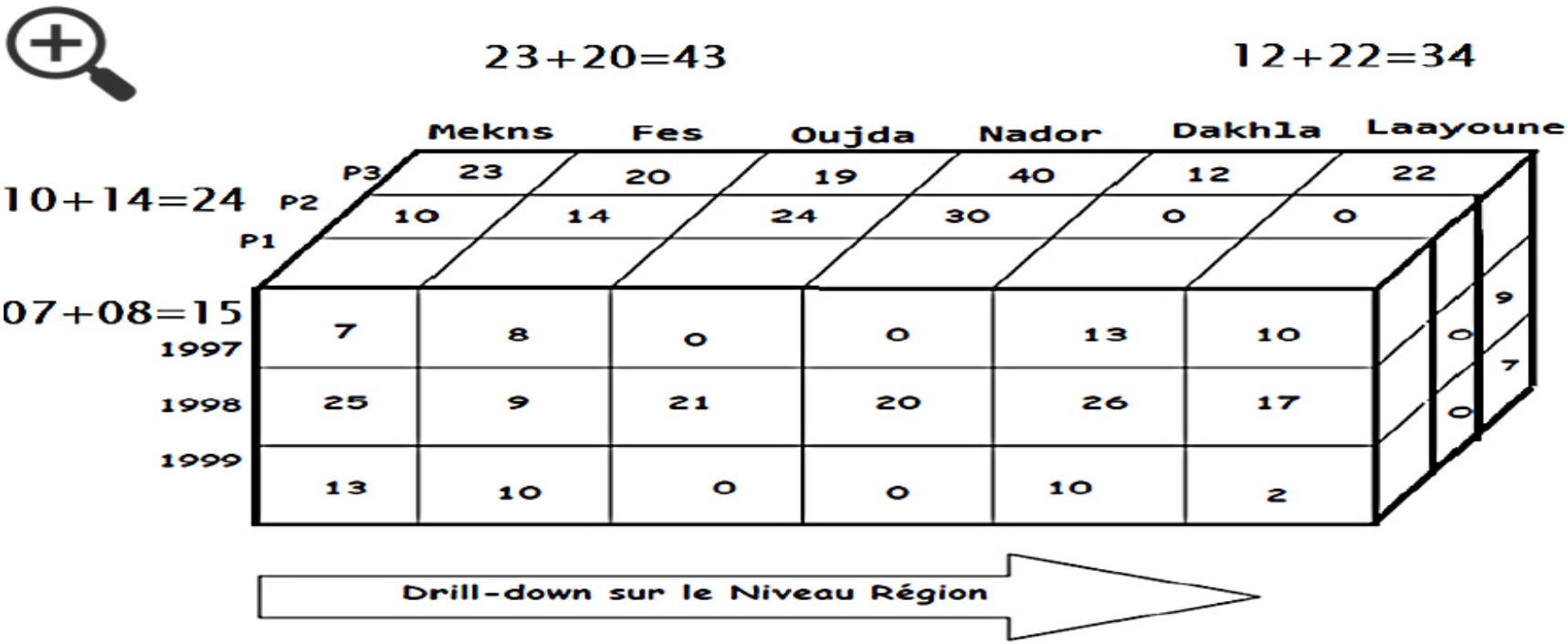


L'opération DICE sur un Cube

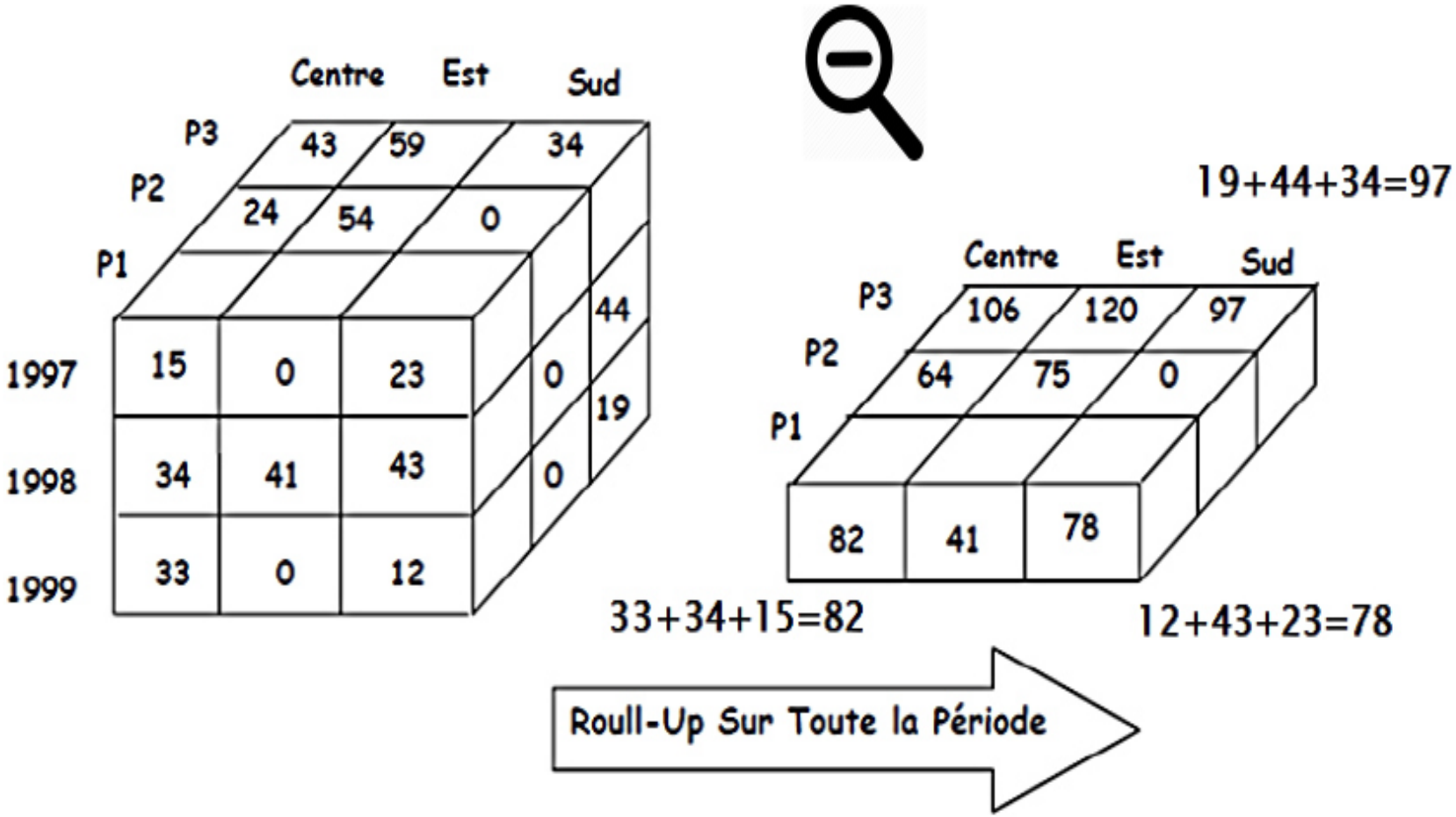
"Drill-Down" et "Roll-Up"

- Les opérations agissant sur la granularité d'observation des données caractérisent la hiérarchie de **navigation entre les différents niveaux**.
- **Roll-up** ou **forage vers le haut** : consiste à représenter les données du cube à un niveau de granularité supérieur conformément à la hiérarchie définie sur la dimension.
  - Utilisation de la fonction d'agrégation (somme, moyenne, etc) spécifée pour la mesure et la dimension.
- **Drill-down** ou **forage vers le bas** : consiste à représenter les données du cube à un niveau de granularité de niveau inférieur sous forme plus détaillée.
- **RQ : Roll-up et Drill-down Opérateurs sur les dimensions**

Exemple "Drill-Down"



Exemple "Roll-Up"



Concept Faits

- Le **fait** modélise le **sujet de l'analyse**. Un fait est **formé de mesures** correspondant aux informations de l'activité analysée (ex : **Quantités** vendues, **montant** des ventes,. . .).
- **Les mesures** d'un fait sont numériques et généralement valorisées de manière continue.
- Les **mesures sont numériques** pour permettre de résumer un grand nombre d'enregistrements en quelques enregistrements (on peut les *additionner*, les *dénombrer* ou bien calculer le *minimum*, le *maximum* ou la *moyenne*).

Table de Faits

- Table principale du modèle multidimensionnel.
- Contient les données **observables** (les **faits**) sur le sujet étudié selon divers **axes d'analyse (les dimensions)**.

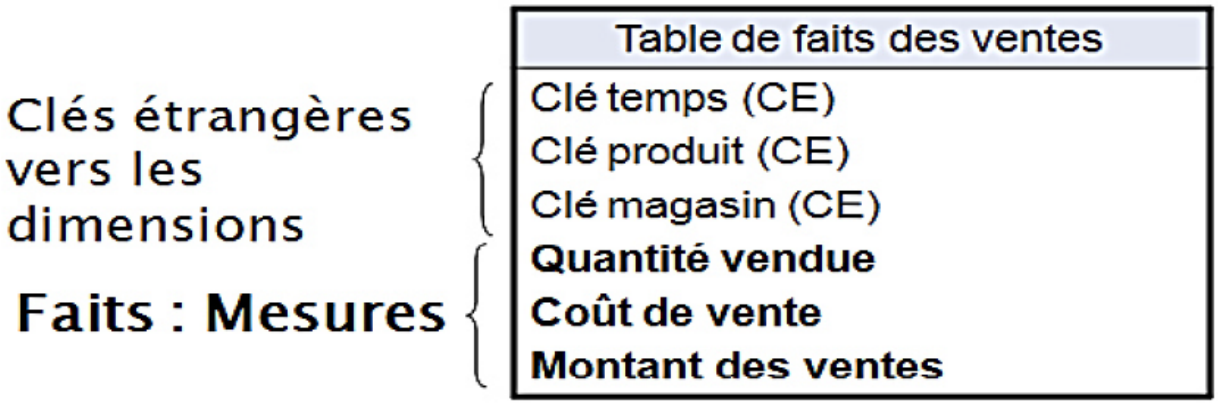


Table de Faits (suite)

- Contient les clés étrangères des axes d'analyse (dimension) : *CléDate*, *Cléproduit*, *Clémagasin*
- Trois types de faits :
  - **Additif**.
  - **Semi Additif**.
  - **Non Additif**.
- **Les faits les plus utiles sont numériques et additifs.**

Types de Faits

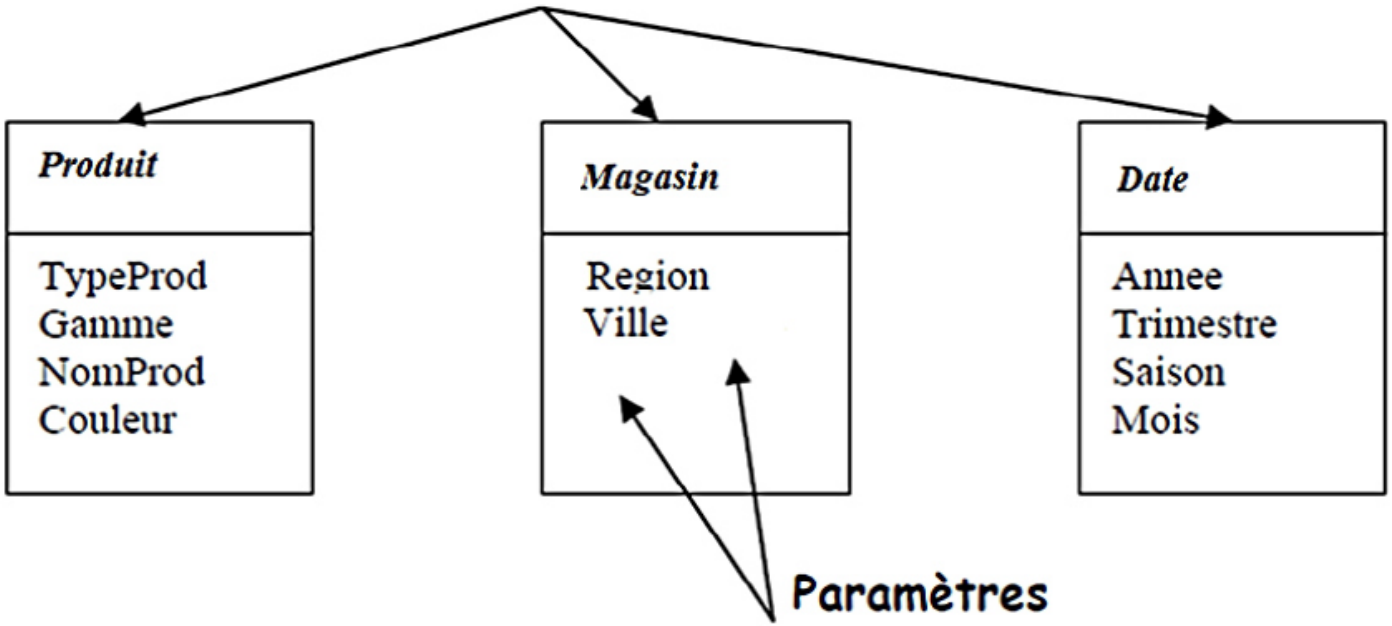
- **Additif** : additionnable suivant toutes les dimensions
  - Quantités vendues, chiffre d'affaire.
  - Peut être le résultat d'un calcul : Bénéfice = montant vente - coût
- **Semi additif** : additionnable suivant certaines dimensions
  - Solde d'un compte bancaire : Pas de sens d'additionner sur les dates car cela représente des instantanés d'un niveau.
- **Non additif** : fait non additionnable quelque soit la dimension.
  - Prix unitaire : l'addition sur n'importe quelle dimension donne un nombre dépourvu de sens.

Dimension

- Le sujet (les faits, mesures) est analysé suivant différentes **perspectives**.
- Ces perspectives correspondent à une **catégorie** utilisée pour **caractériser** les **mesures** d'activité analysées, on parle de **dimensions**.
- Une **dimension** modélise une **perspective** d'analyse.
- Une dimension se compose d'**attributs** et *niveaux* correspondant aux informations **faisant varier les mesures** de l'activité.

Exemples dimension

Vente (**Produit**, **client**, **magasin**, **vendeur**, **date**, coût de vente, montant des ventes, Qté vendue)

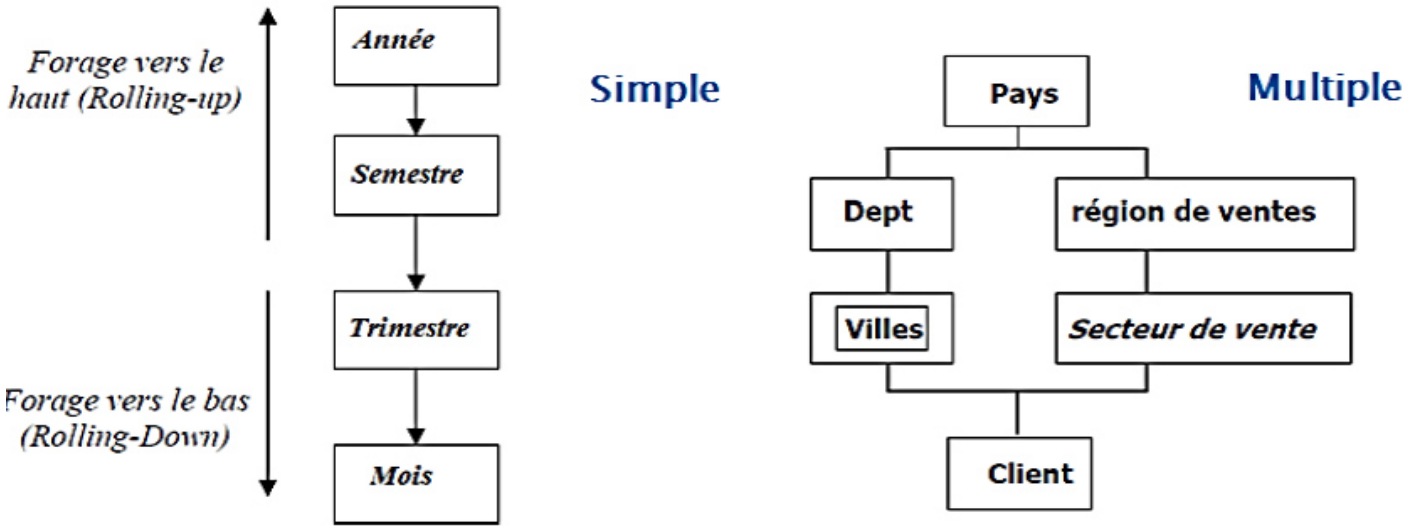


Attributs et Niveaux

- Une dimension est généralement formée de paramètres (ou attributs) textuels et discrets.
- **Les paramètres textuels** sont utilisés pour restreindre la portée des requêtes afin de limiter la taille des réponses.
- **Les paramètres sont discrets**, c'est à dire que les valeurs possibles sont bien déterminées et sont des descripteurs constants.

Concept Hiérarchie/Granularité

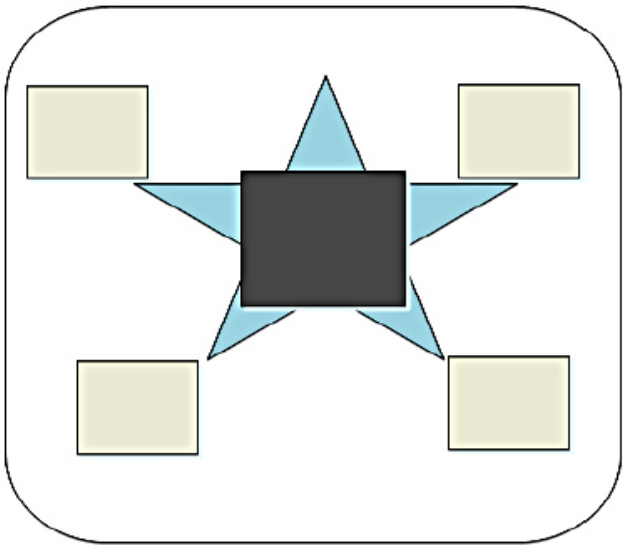
- Une **hiérarchie** organise les attributs d'une dimension selon une relation "est plus fin" conformément à leur **niveau** de détail ou **granularité**.
- On distingue deux types de hiérarchies **Simple** et **Multiple** :



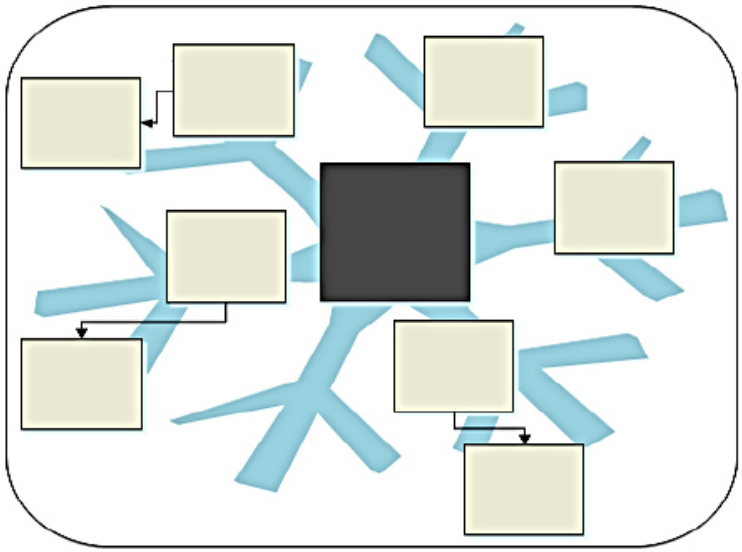
Modèle Multidimensionnel

- **Table de faits** + **Tables de dimensions** reliées à la table de faits par une jointure.
- Chaque enregistrement de la table de faits stocke les clés des tables de dimensions et les mesures faites à un instant précis.
- Chacune des tables de dimension possède une clé primaire unique correspondant à l'un des composants de la clé multiple de la table de faits.
- On obtient un schéma en étoile (dans le cas le plus simple).

Types de modèles



Modèle en étoile

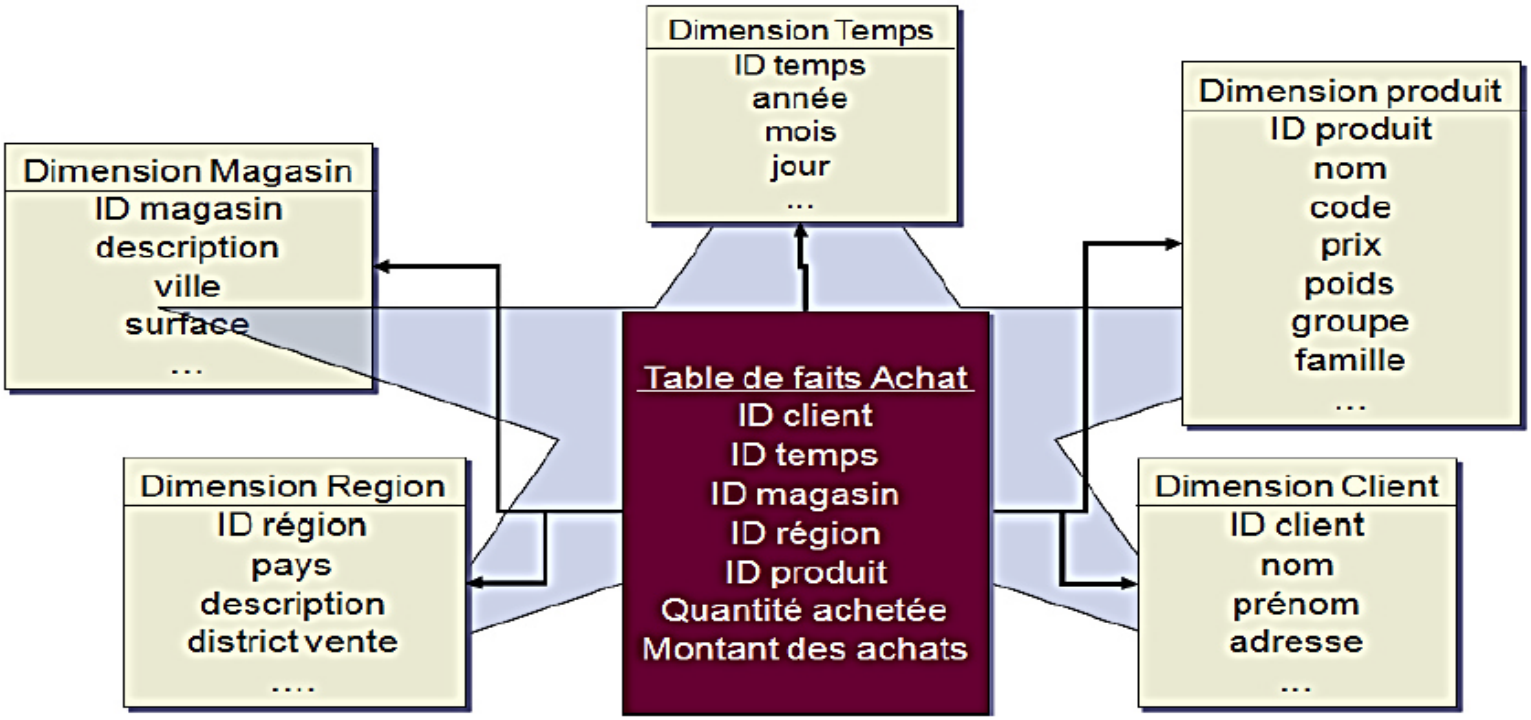


Modèle en flocon

Modèle en étoile

- Une table de fait **centrale** et des **dimensions**.
- Les dimensions n'ont pas de liaison entre elles
- Chacune des tables de dimension possède une clé primaire unique correspondant à l'un des composants de la clé multiple de la table de faits.
- **Avantages** :
  - Facilité de navigation.
  - Nombre de jointures limité.
- **Inconvénients** :
  - Redondance dans les dimensions.
  - Toutes les dimensions ne concernent pas les mesures.

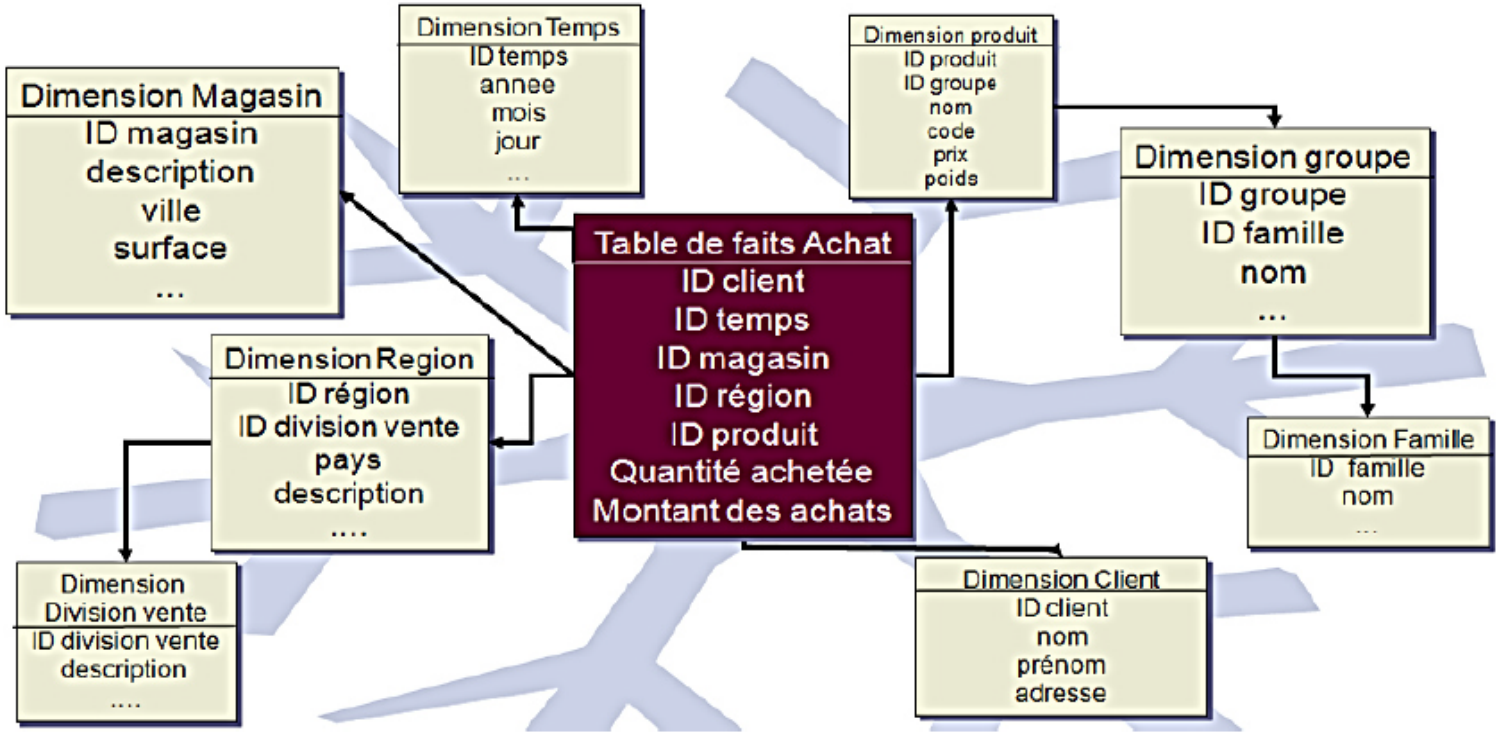
Exemple de Modèle en étoile



Modèle en flocon

- Une **table** de **fait** et des **dimensions décomposées** en sous **hiérarchies**.
- **Un seul niveau hiérarchique par table de dimension.**
- La table de dimension de **niveau** hiérarchique le **plus bas est reliée à la table de fait**.
- **Avantages** :
  - Normalisation des dimensions.
  - Économie d'espace disque.
- **Inconvénients** :
  - Modèle plus complexe (jointure).
  - Requêtes moins performantes.

Exemple de Modèle en flocon



Méthodologie de Kimball

1. Choisir le sujet.
2. Choisir la granularité des faits.
3. Identifier et adapter les dimensions.
4. Choisir les faits.
5. Stocker les pré-calculs.
6. Établir les tables de dimensions.
7. Choisir la durée de la base.
8. Suivre les dimensions lentement évolutives.
9. Décider des requêtes prioritaires, des modes de requêtes.

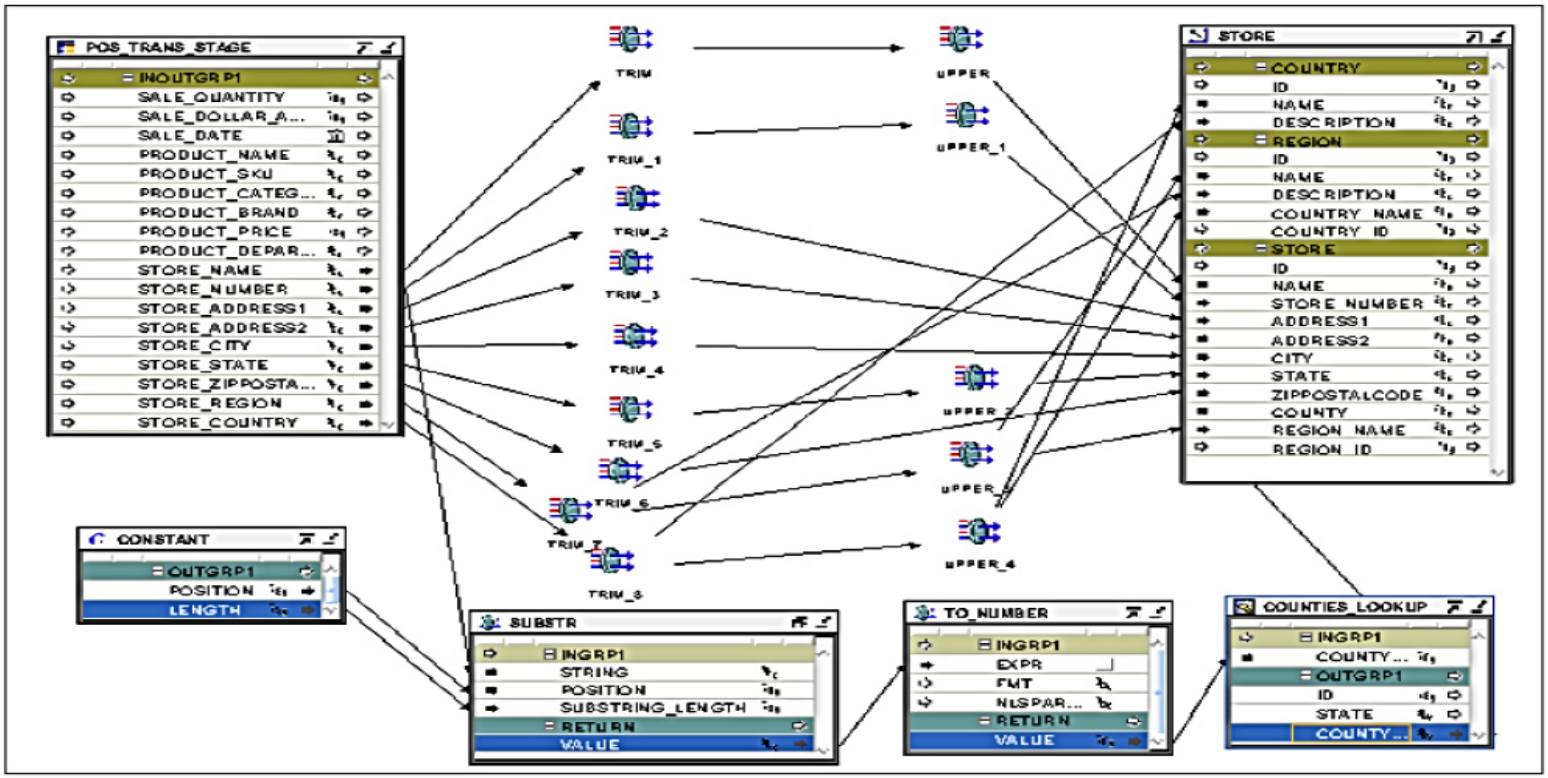
Définition d'un ETL

- Offre un **environnement de développement** ;
- Offre des outils de **gestion** des opérations et de **maintenance** ;
- Permet de **découvrir**, **analyser** et **extraire** les données à partir de sources hétérogènes ;
- Permet de **nettoyer** et standardiser les données ;
- Permet de **charger** les données dans un entrepôt.

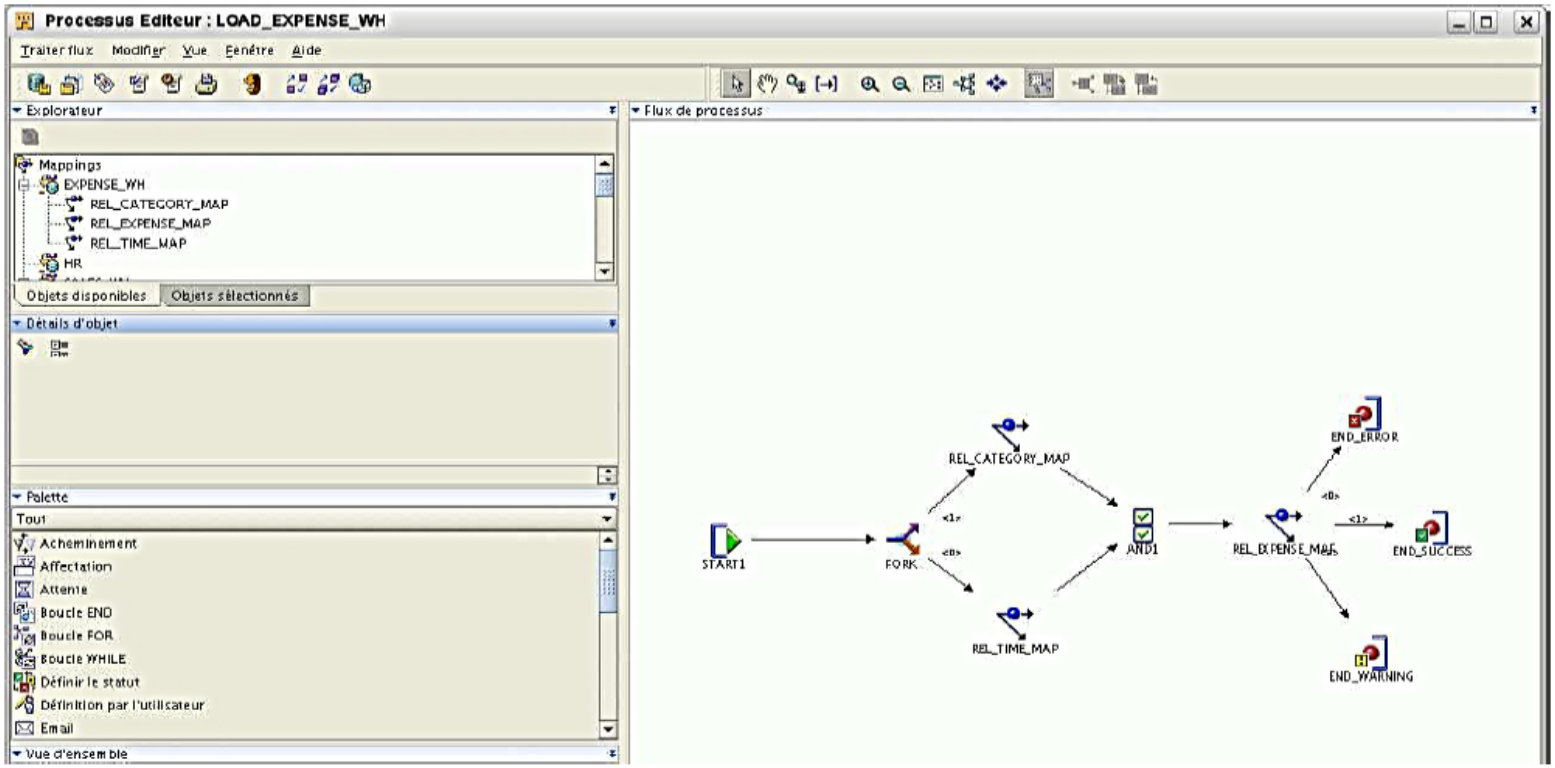
Alimentation/ mise à jour de l'entrepôt

- Entrepôt mis à jour régulièrement ;
- Chargement automatisé de l'entrepôt ;
- Utilisation d'outils **ETL (Extract, Transform, Load)**

Aperçu d'un ETL



Aperçu d'un ETL (suite)



ROLAP

- **Relational OLAP** :
  - **Données stockées dans une base de données relationnelles** ;
  - Un moteur **OLAP** permet de simuler le comportement d'un SGBD multidimensionnel ;
- Plus facile et moins cher à mettre en place ;
- Moins performant lors des phases de calcul ;
- Exemples de moteurs ROLAP : Mondrian

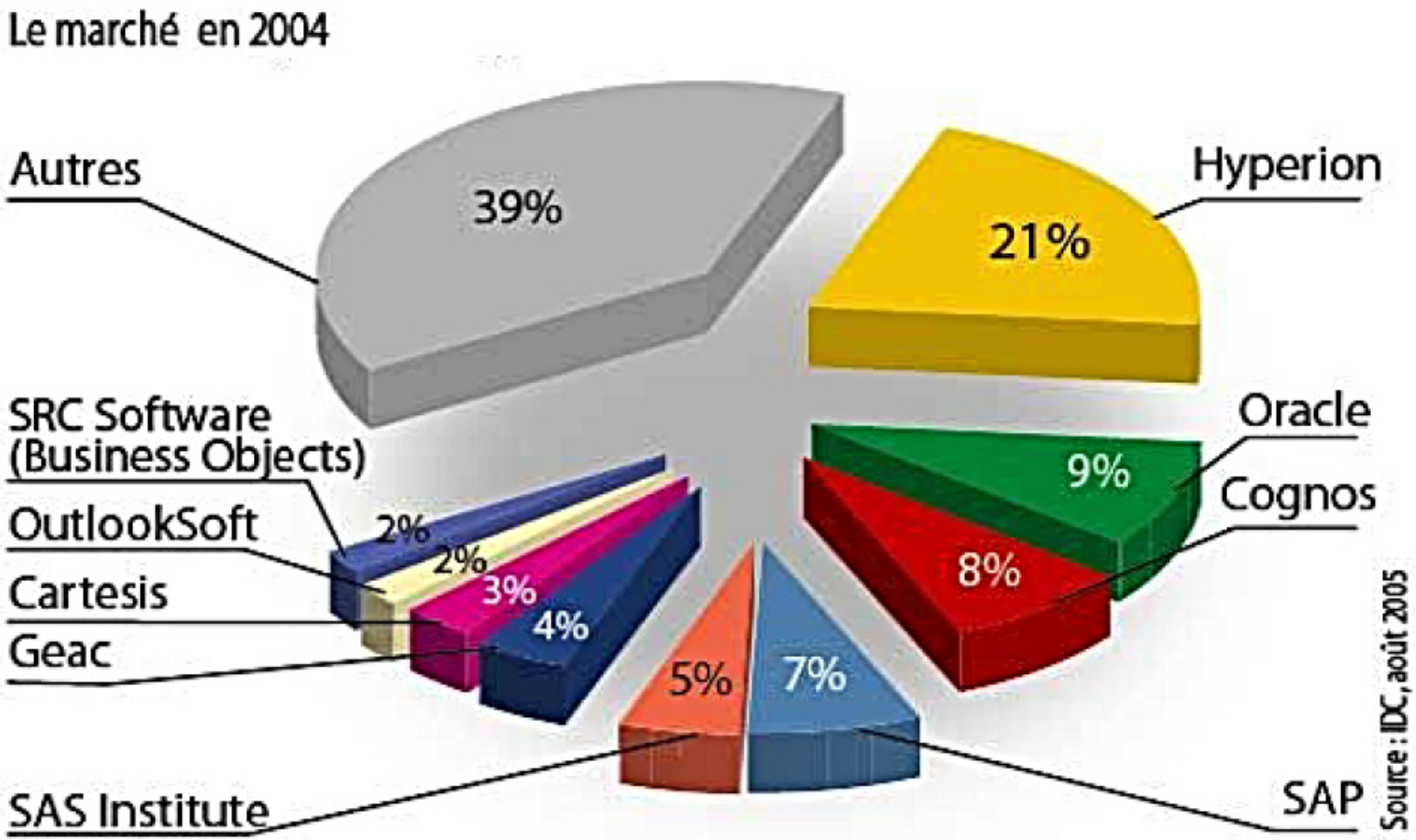
MOLAP

- **Multi dimensional OLAP : :**
  - Utilise un système multidimensionnel " **pur** " qui gère les structures multidimensionnelles natives (les cubes) ;
  - Accès direct aux données dans le cube ;
- Plus dificile à mettre en place ;
- Formats souvent propriétaires ;
- Conçu exclusivement pour l'analyse multidimensionnelle.
- Exemples de moteurs MOLAP : (Microsoft Analysis Services, Hyperion)

HOLAP

- **Hybride OLAP** :
  - tables de faits et tables de dimensions stockées dans SGBD relationnel (données de base) ;
  - données agrégées stockées dans des cubes ;
- Solution **hybride entre MOLAP et ROLAP** ;
- Bon compromis au niveau coût et performance.

Marché Décisionnel



Quelques solutions commerciales

Quelques solutions open source

ETL	Entrepôt de données	OLAP	Reporting	Data Mining
■Octopus ■Kettle ■CloverETL ■Talend	■MySql ■Postgresql ■Greenplum/Bizgres	■Mondrian ■Palo	■Birt ■Open Report ■Jasper Report ■JFreeReport	■Weka ■R-Project ■Orange ■Xelopes

Intégré
■Pentaho (Kettle, Mondrian, JFreeReport, Weka)
■SpagoBI