Entreposage, analyse en ligne et fouille de données

Houssem Jerbi

IRIT - SIG/ED

jerbi@irit.fr

PLAN

- Introduction
- Bases de données
- Entrepôt de données
- Technologie OLAP
- Fouille de données
- Conclusion

Introduction: Système d'information

- L'information: capital immatériel pour toute organisation
- Les « Data Trucs »
  - Data (Données)
  - Database (Base de données)
  - Data warehouse (Entrepôt de données)
  - Datamart (Magasin de données)
  - Data mining (fouille des données)

PLAN

- Introduction
- Bases de données
- Entrepôt de données
- Technologie OLAP
- Fouille de données
- Conclusion

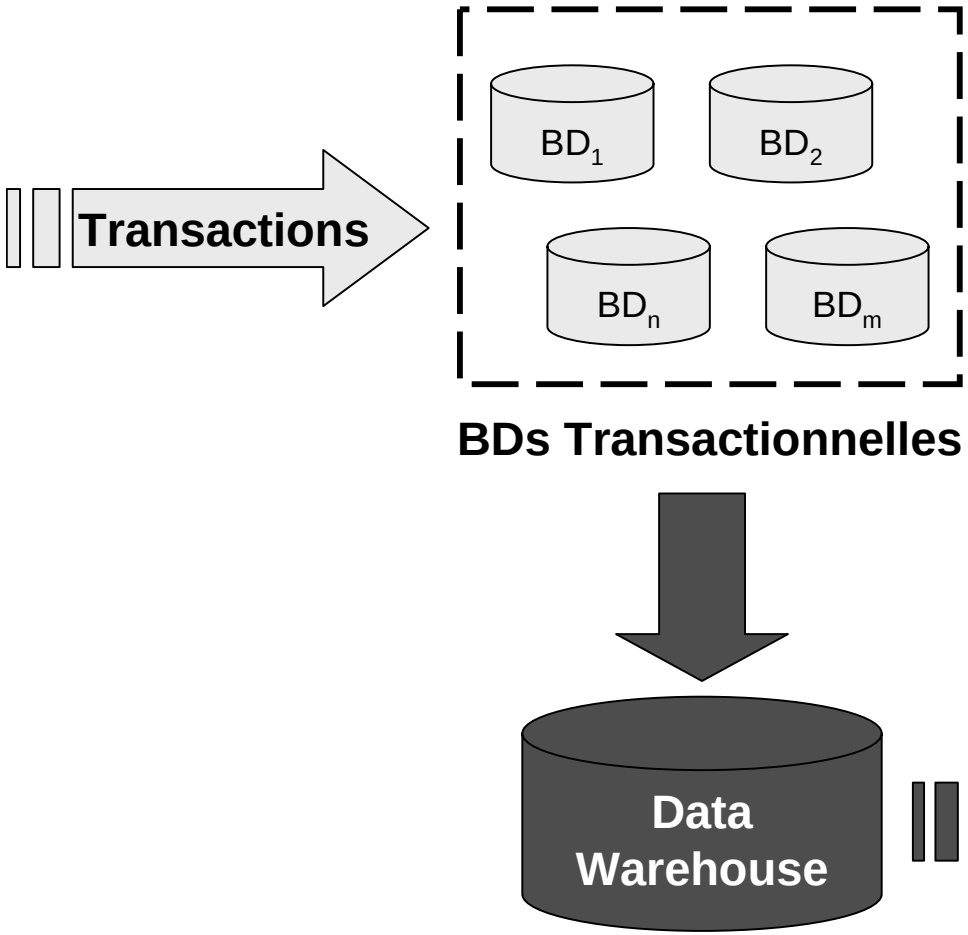
Bases de données

- Données transactionnelles
  - Exemple: données de stock, facturation,…
- Modèle de données
  - Le plus utilisé: relationnel
  - Normalisation: cohérence et non-redondance des données
- Requêtes ponctuelles, fréquentes

PLAN

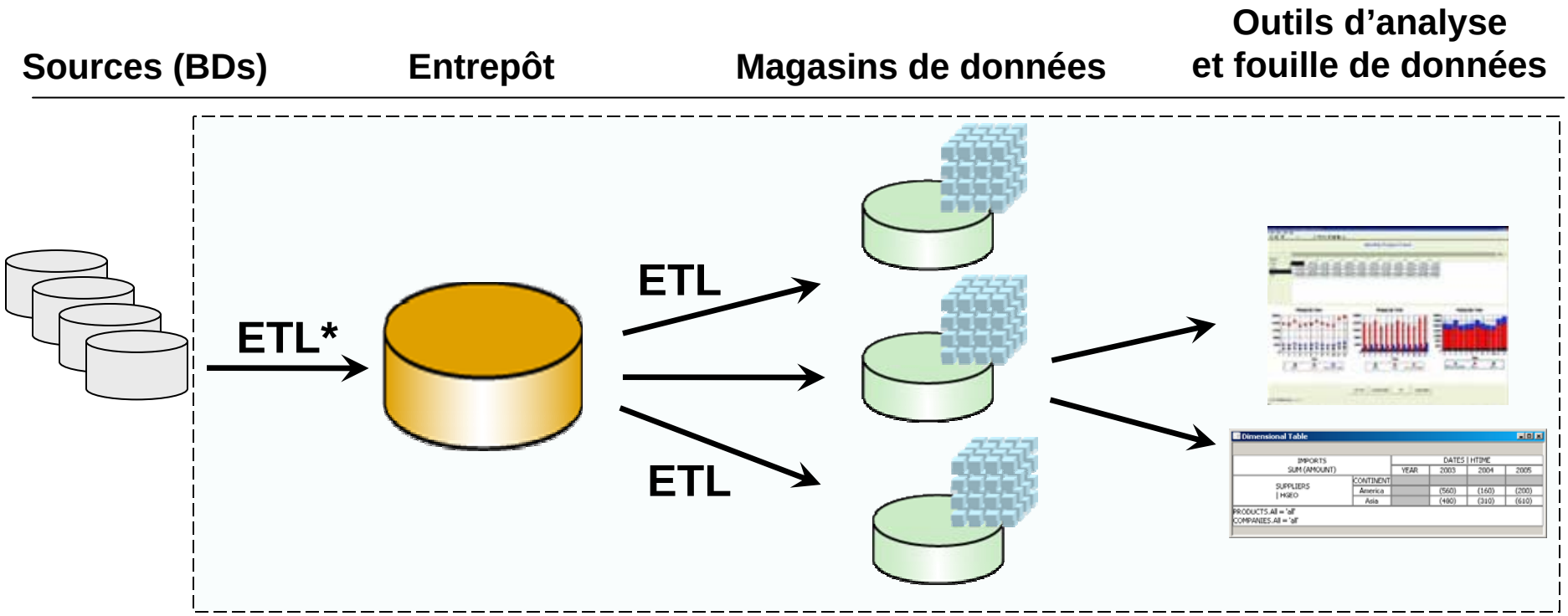
- Introduction
- Bases de données
- Entrepôt de données
- Technologie OLAP
- Fouille de données
- Conclusion

Entrepôt de données



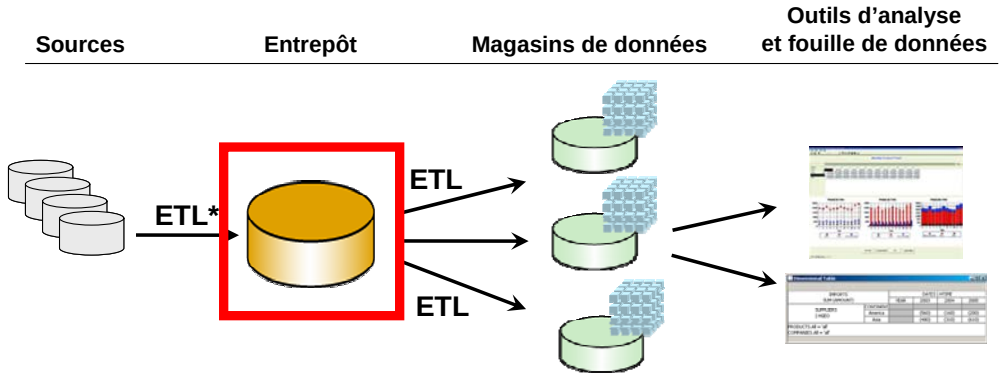
- **OLAP**
- **Fouille de données**

Entrepôt de données: Architecture de système d'information décisionnel



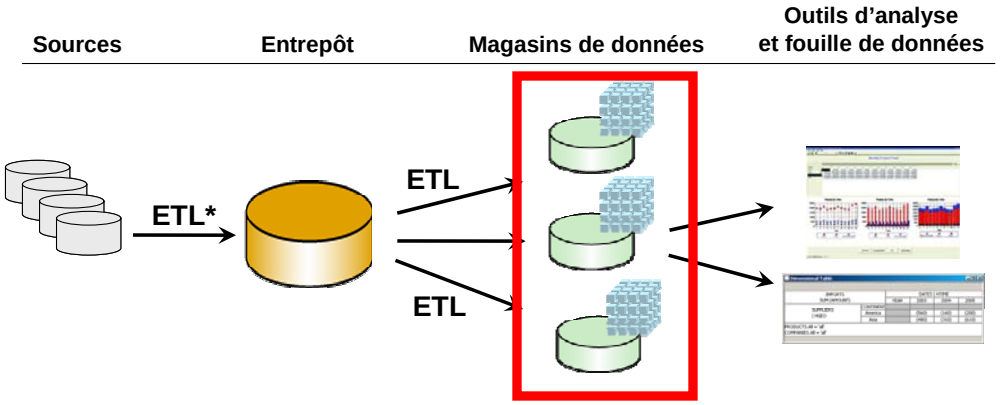
*** ETL: Extract, Transform, Load**

Entrepôt de données



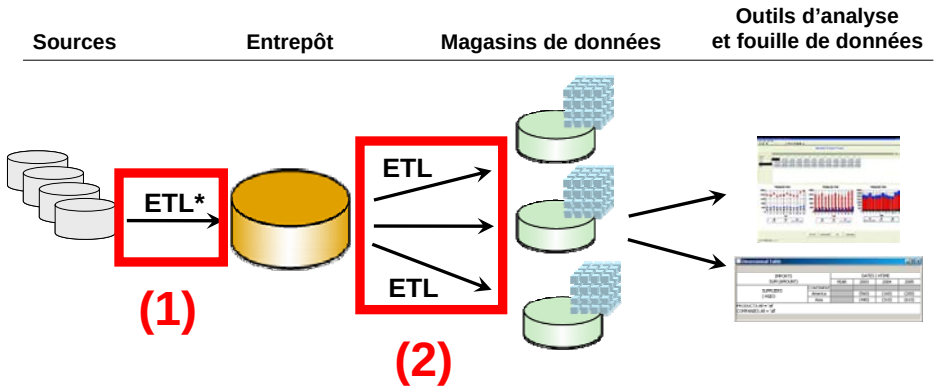
- Collecte toutes les informations sur tous les sujets pour l'organisation
- Espace de stockage centralisé qui permet de stocker et d'historiser des données résumées nécessaires à la prise de décision

Magasin de données



- Extrait de l'entrepôt destiné à une classe de décideurs
- Modèle multidimensionnel qui facilite les traitements décisionnels

Outils ETL



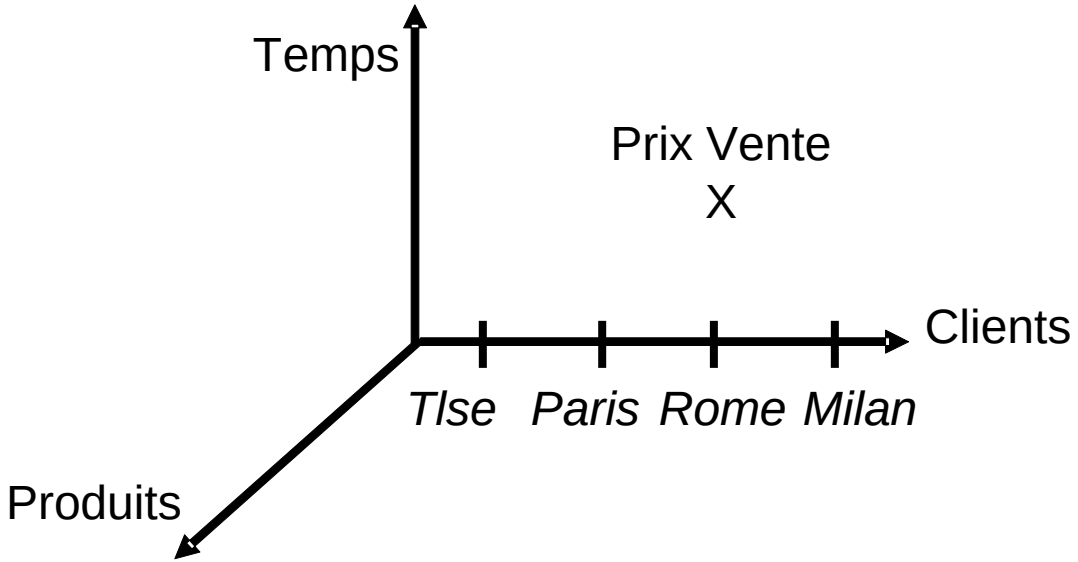
- Alimentation de l'entrepôt (1) et extraction des magasins (2)
- Extract
  - Accès aux différentes sources
  - Selon des règles (déclencheurs) ou requêtes
  - Périodique

Outils ETL

- Transform
  - Unification des modèles (sources hétérogènes)
  - Gestion des inconsistances des données sources, élimination des doubles, etc.
- Load
  - Chargement dans l'entrepôt ou dans les magasins
  - Périodicité parfois longue

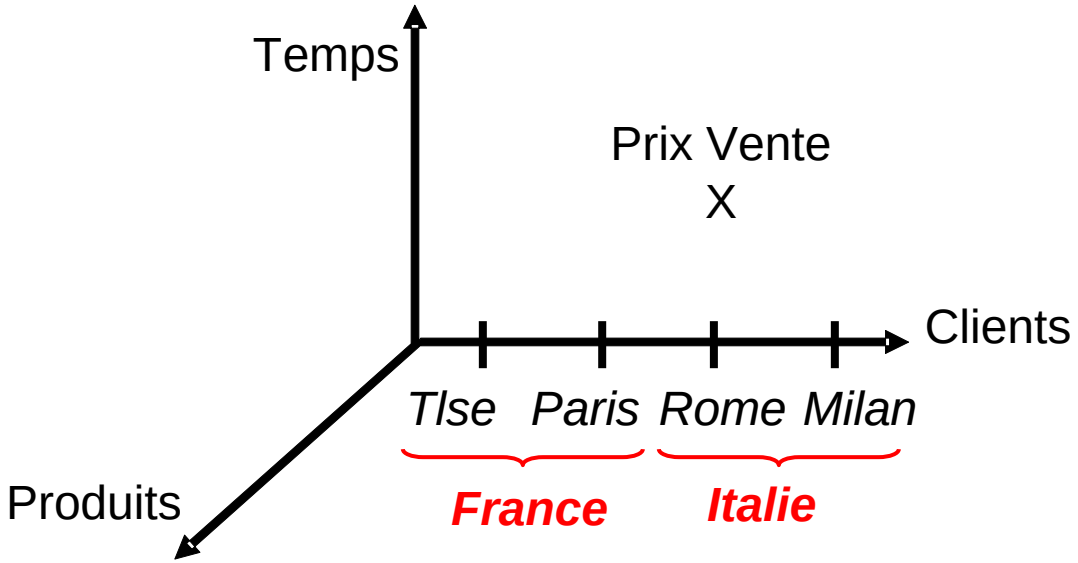
Magasins: BD multidimensionnelles

- Modèle facilitant l'analyse décisionnelle
  - Sujets (faits) et
  - Axes d'analyse (dimensions)
    - Niveaux de granularité



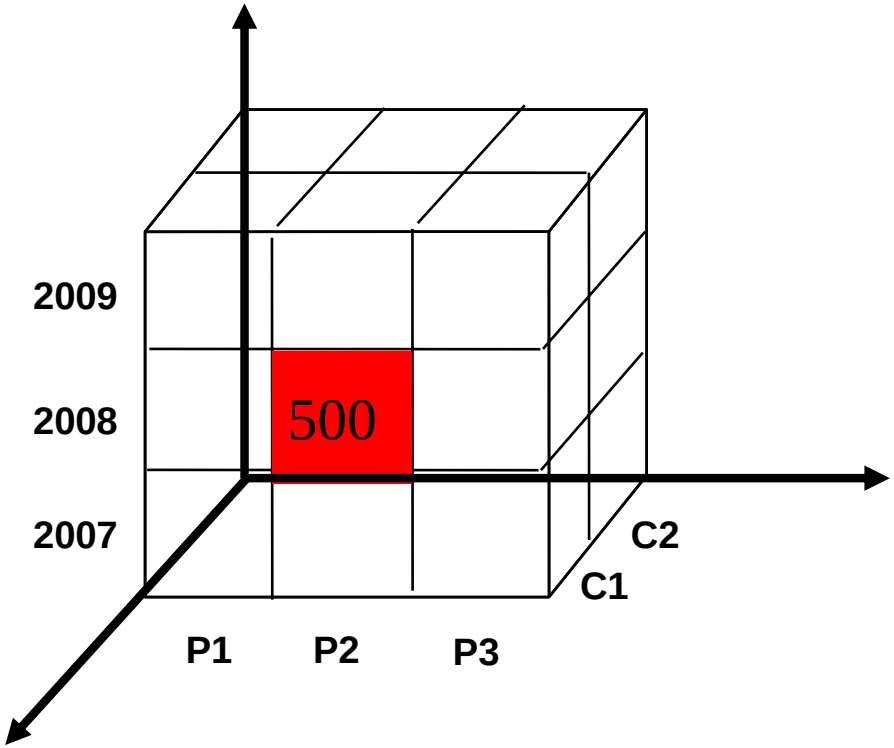
Magasins: BD multidimensionnelles

- Modèle facilitant l'analyse décisionnelle
  - Sujets (faits) et
  - Axes d'analyse (dimensions)
    - Niveaux de granularité



Magasins: BD multidimensionnelles

- Métaphore du cube
  - 500: Somme des ventes en *2008* du produit *P2* pour le client *C1*



PLAN

- Introduction
- Bases de données
- Entrepôt de données
- Technologie OLAP
- Fouille de données
- Conclusion

Analyse en ligne OLAP

- On-Line Analytical Processing: OLAP
- Opérations de manipulation de données
  - Forage
    - Roll up : Agréger selon une dimension
      - Jour ➔ Mois
    - Drill down : Détailler selon une dimension
      - Mois ➔ Jour
  - Sélection et projection selon un axe
    - Année = 2010 (année en cours)

OLAP: Opérations de manipulation

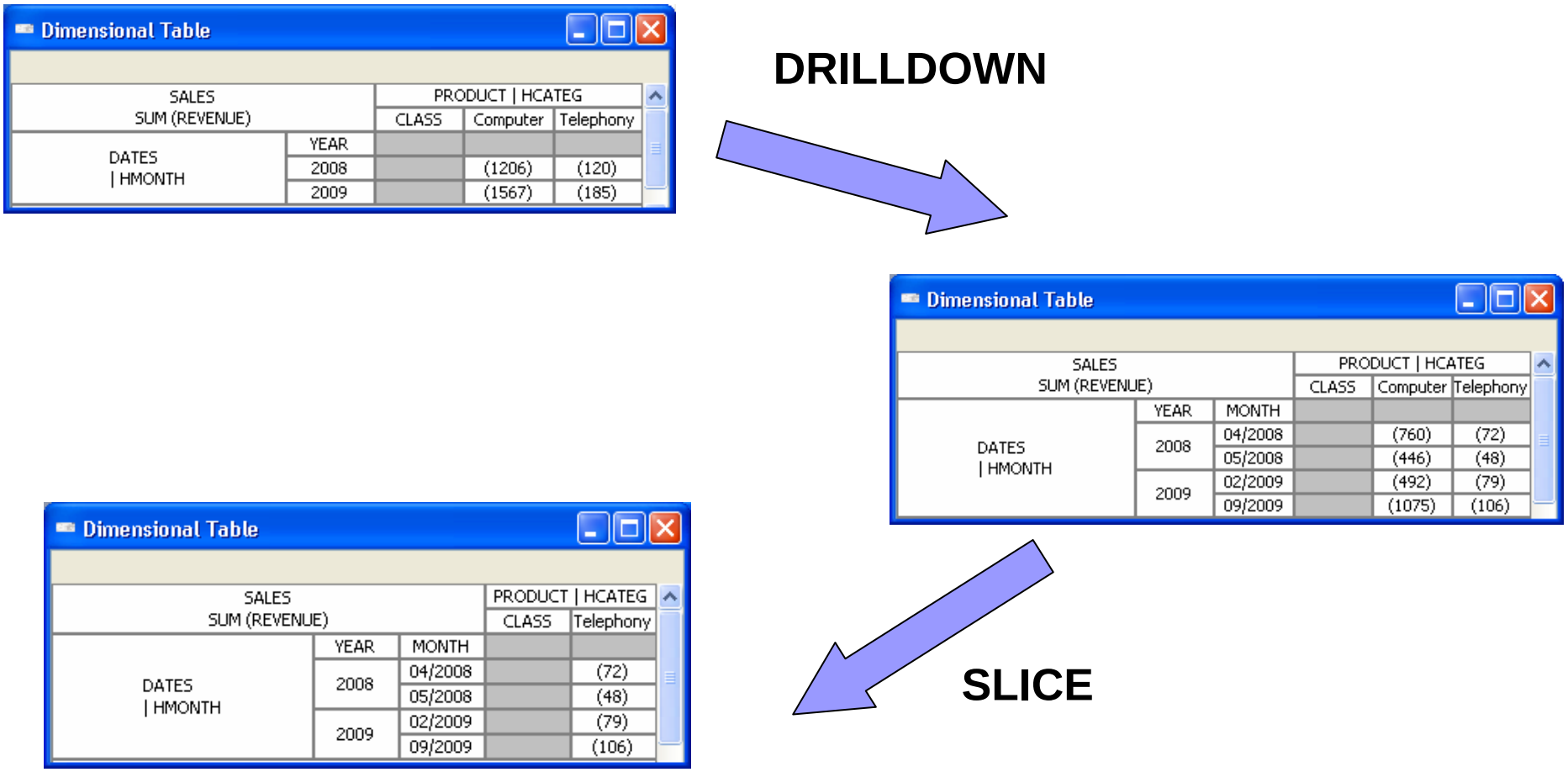
- Opérations de manipulation de données
  - Rotation : Changer un axe de l'analyse
    - (Année,Produit)➔(Ville, Produit)

- Structure de visualisation adaptée: Table multidimensionnelle

Dimensional Table

SALES SUM (REVENUE)		PRODUCT | HCATEG		
		CLASS	Computer	Telephony
DATES | HMONTH	YEAR			
	2008		(1206)	(120)
	2009		(1567)	(185)

OLAP: Exemple



PLAN

- Introduction
- Bases de données
- Entrepôt de données
- Technologie OLAP
- Fouille de données
- Conclusion

Fouille de données (Data mining)

- Au-delà de l'OLAP: la fouille des données
  - OLAP: offrir une vue de « ce qui se passe »
  - Data mining: prévoir « ce qui se passera » et « pourquoi »
    - s'appuie sur des techniques d'intelligence artificielle
    - met en évidence des liens cachés entre les données.

Fouille de données

- Prévoir ce qui se passera dans le futur
- Classifier des personnes ou des entités en groupes
  - par reconnaissance de motifs
  - En se basant sur leurs attributs
- Associer les évènements qui pourraient survenir ensemble

Exemple: Données médicales

- Entrepôt pour le suivi de maladies infectieuses
- Analyse OLAP
  - Analyse du taux d'affectation par ville, par département, par année
- Fouille des données
  - Corrélation entre le taux d'affectation et le taux de présence de certains polluants

Conclusion

- Besoin de prise de décision: entrepôt de données
- Dichotomie Entrepôt/Magasin de données
- Application de l'OLAP aux données scientifiques