Entrepôts de données

NEGRE Elsa
Université Paris-Dauphine
2016-2017

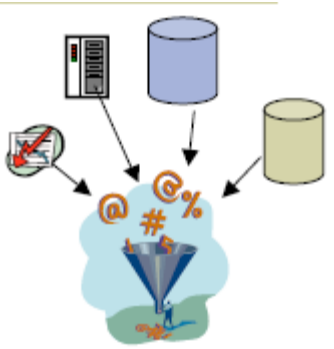
- Contexte et problématique
  - Le processus de prise de décision
- L'entrepôt de données
  - Définition
  - Différence avec un SGBD
  - Caractéristiques
- Architecture d'un système décisionnel
- Modélisation multidimensionnelle
  - Niveau conceptuel
  - Niveau logique
  - Niveau physique
- Réalisation d'un entrepôt
- Représentation et manipulation
  - Le cube OLAP
- Solutions existantes

Contexte (1)

- Besoin :
  - Prise de décisions stratégiques et tactiques
  - Réactivité
- Qui :
  - les décideurs (non informaticiens, non statisticiens)
- Comment :
  - Répondre aux demandes d'analyse de données
  - Dégager des informations qualitatives nouvelles

Contexte (2)

- Type de données : données opérationnelles (de production)
  - Bases de données, Fichiers, Paye, Gestion RH, …



- Caractéristiques des données :
  - Distribuées : systèmes éparpillés
  - Hétérogènes : systèmes et structures de données différents
  - Détaillées : organisation de données selon les processus fonctionnels et données trop abondantes pour l'analyse
  - Peu/pas adaptées à l'analyse : des requêtes lourdes peuvent bloquer le système transactionnel
  - Volatiles : pas d'historisation systématique

Problématique (1)

Nous avons donc :

- Une grande masse de données
  - Distribuées
  - Hétérogènes
  - Très détaillées
- à traiter
  - Synthétiser / résumer
  - Visualiser
  - Analyser
- pour une utilisation par des
  - Experts / analystes d'un métier
  - Non informaticiens
  - Non statisticiens

Problématique (2)

- Comment répondre aux besoins de décideurs afin d'améliorer les performances décisionnelles de l'entreprise?
  - En donnant un accès rapide et simple à l'information stratégique
  - En donnant du sens aux données
  - En donnant une vision transversale des données de l'entreprise (intégration de différentes bases de données)
  - En extrayant, groupant, organisant, corrélant et transformant (résumé, agrégation) les données

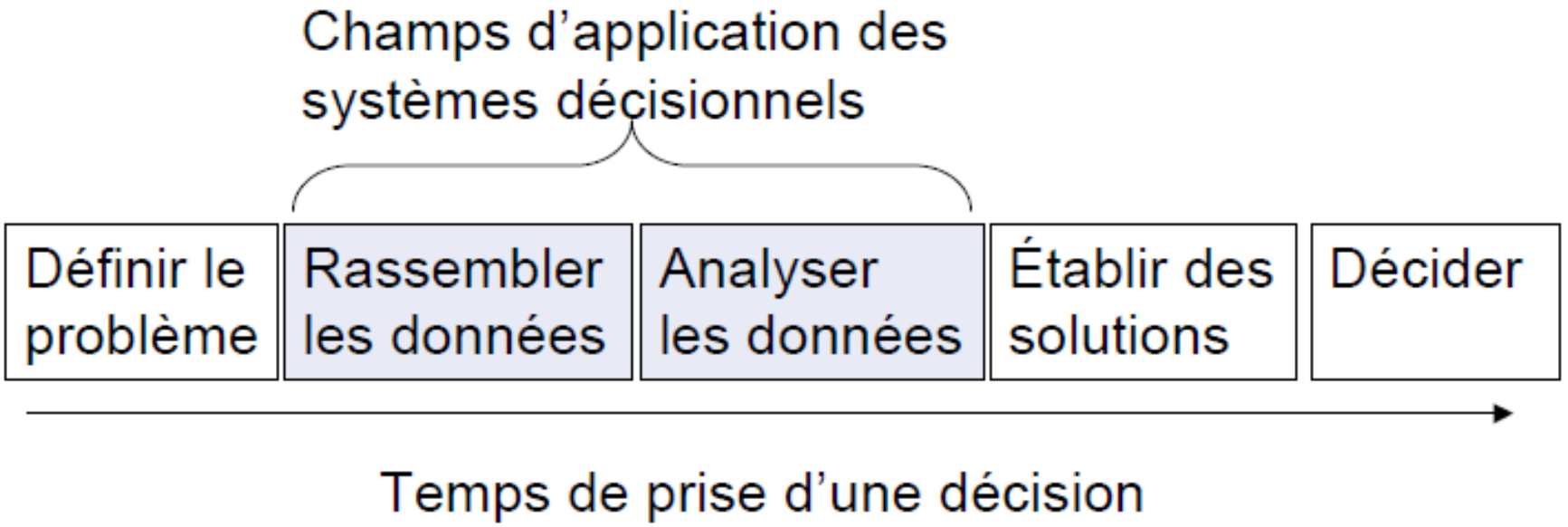
Problématique (3)

Mettre en place un SI dédié aux applications décisionnelles : un entrepôt de données (*datawarehouse*)

- Transformer des données de production en informations stratégiques

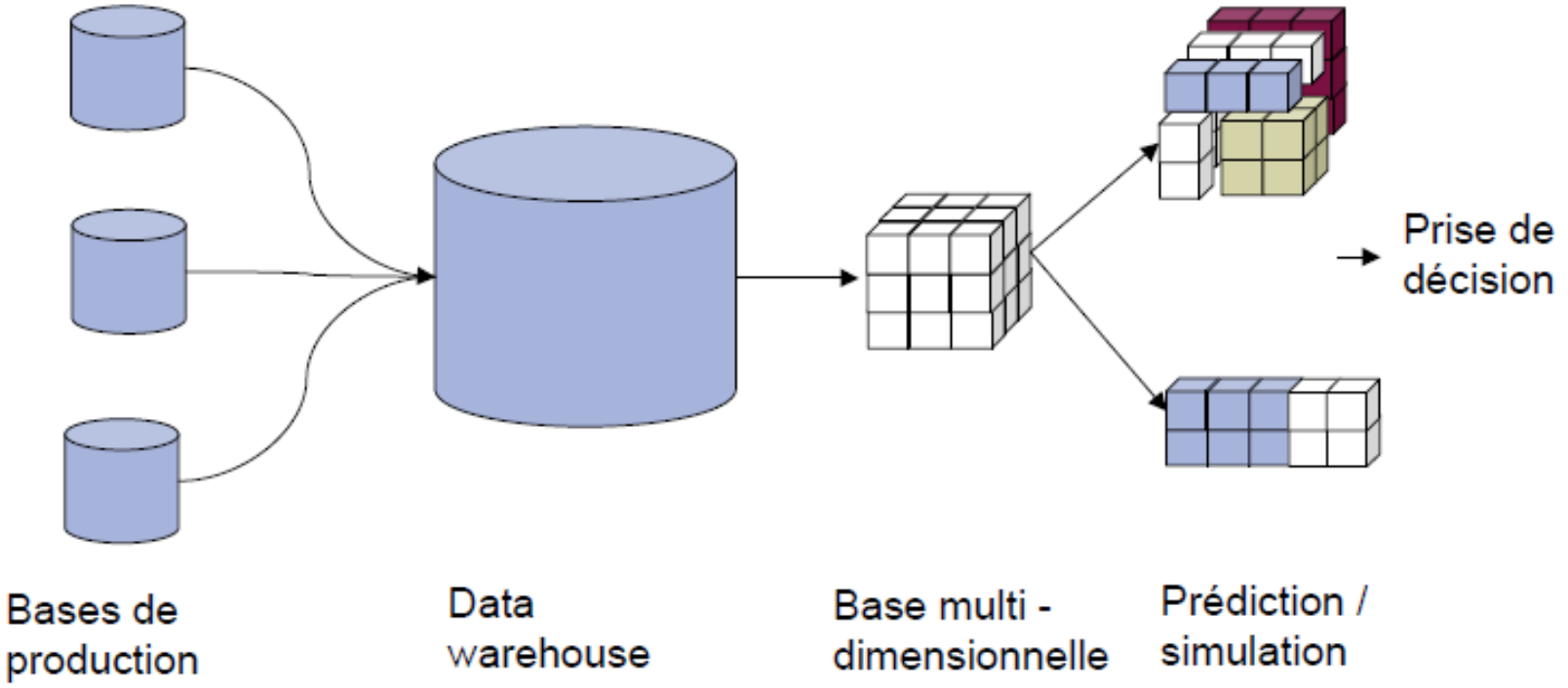
Sources : Th. Ester, HEC Lausanne

Le processus de prise de décision (1)



Sources : Lydie Soler, AgroTechParis

Le processus de prise de décision (2)



Sources : Lydie Soler, AgroTechParis

L’entrepôt : Définition

- *Le DW est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d’un processus d’aide à la décision.*

  W.H. Inmon (1996)

- C’est une BD à des fins d’analyse !!

Pourquoi pas un SGBD ? (1)

- Fonctions d'un SGBD :
  - Systèmes transactionnels (OLTP)
  - Permettre d'insérer, modifier, interroger rapidement, efficacement et en sécurité les données de la base
  - Sélectionner, ajouter, mettre à jour, supprimer des tuples
  - Répondre à de nombreux utilisateurs simultanément

Pourquoi pas un SGBD ? (2)

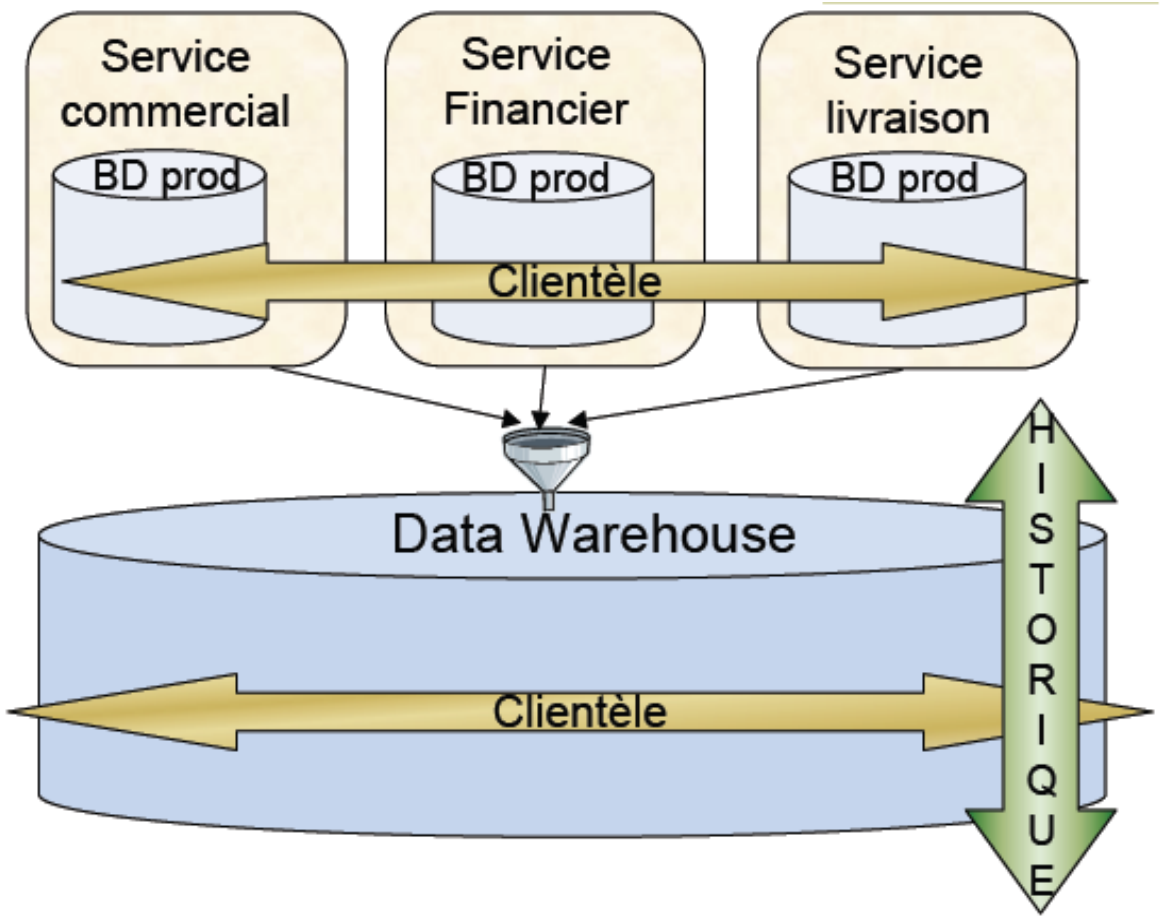
- Fonctions d'un DW :
  - Systèmes pour l'aide à la prise de décision (OLAP)
  - Regrouper, organiser des informations provenant de sources diverses
  - Intégrer et stocker les données pour une vue orientée métier
  - Retrouver et analyser l'information rapidement et facilement

Pourquoi pas un SGBD ? (3)

	OLTP	DW
Utilisateurs	Nombreux Employés	Peu Analystes
Données	Alphanumériques Détaillées / atomiques Orientées application Dynamiques	Numériques Résumées / agrégées Orientées sujet Statiques
Requêtes	Prédéfinies	« one-use »
Accès	Peu de données (courantes)	Beaucoup d'informations (historisées)
But	Dépend de l'application	Prise de décision
Temps d'exécution	Court	Long
Mises à jour	Très souvent	Périodiquement

Pourquoi pas un SGBD ? (4)

OLTP: On-Line Transactional Processing

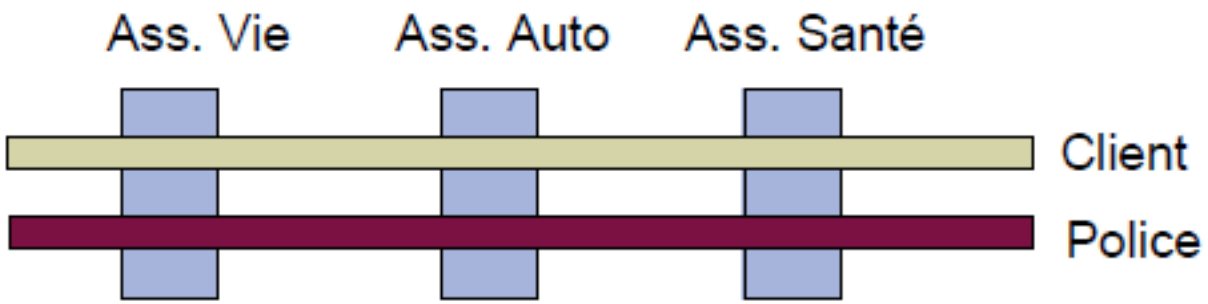


OLAP: On-Line Analitical Processing

Sources : Lydie Soler, AgroTechParis

Caractéristiques d'un DW (1)

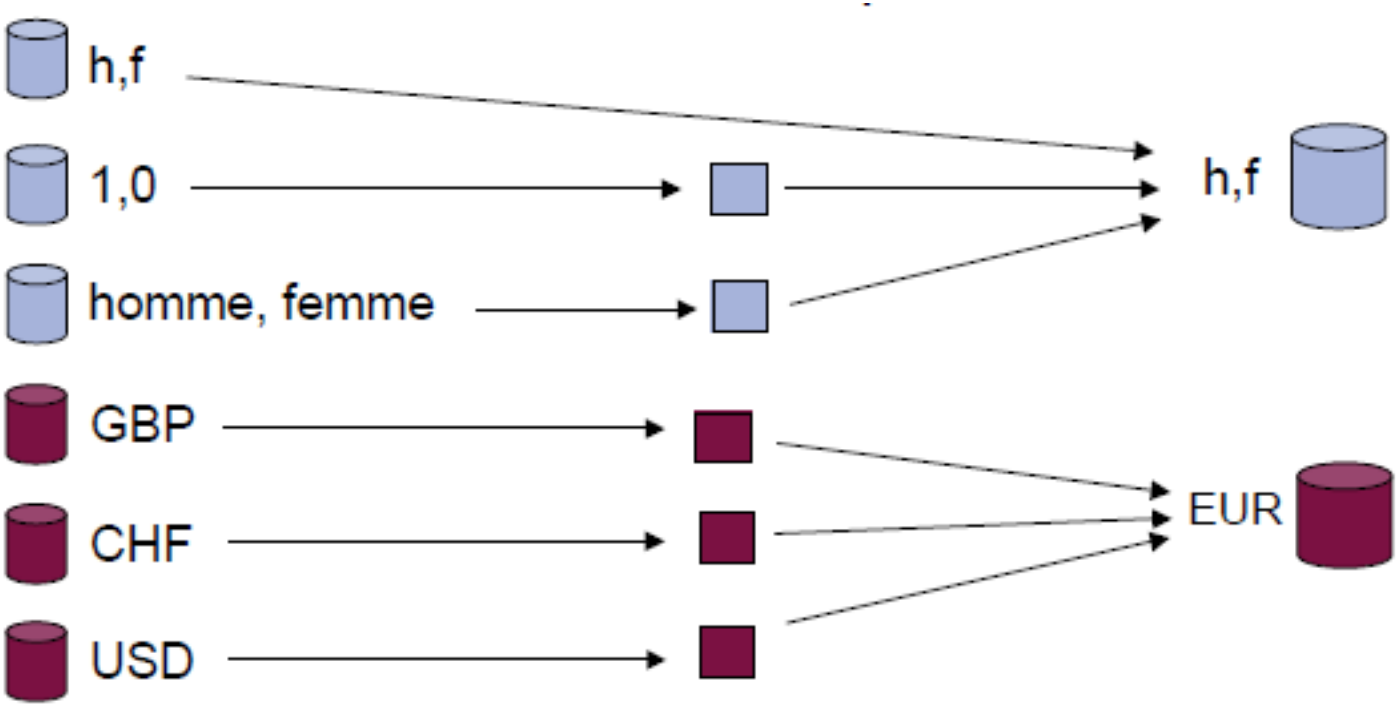
- Données orientées sujet
  - Regroupe les informations des différents métiers
  - Ne tiens pas compte de l'organisation fonctionnelle des données



Sources : Lydie Soler, AgroTechParis

Caractéristiques d'un DW (2)

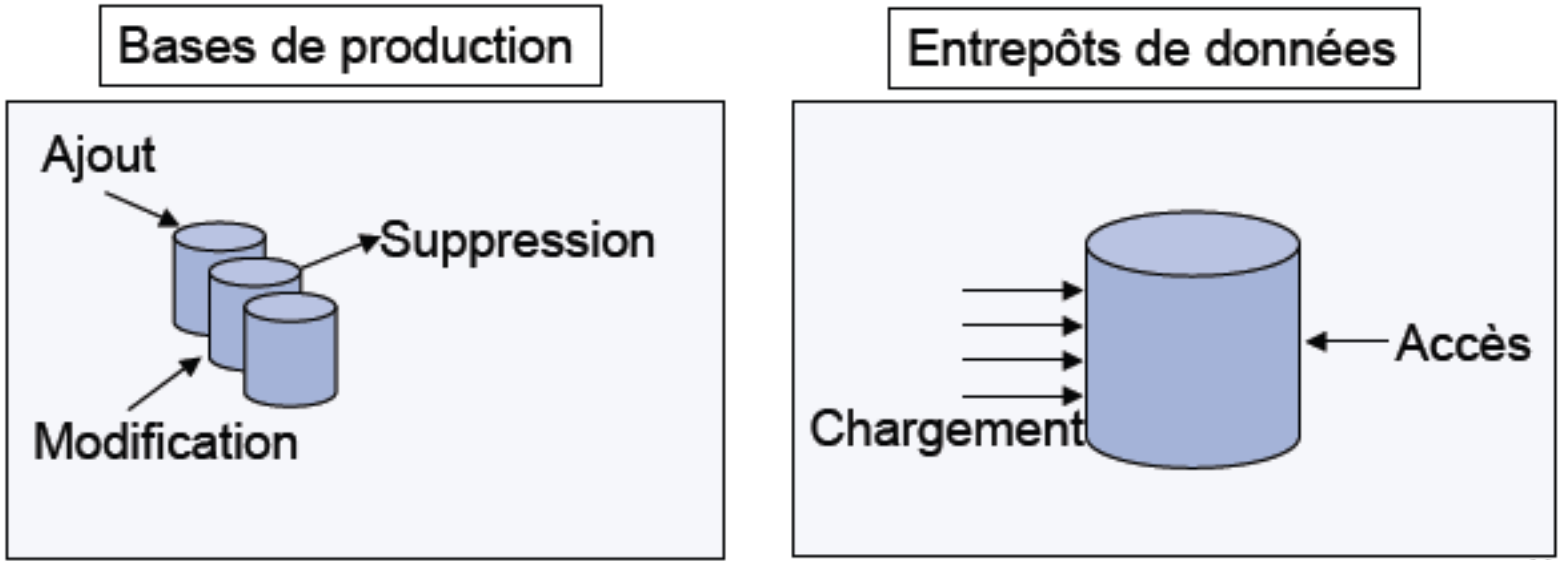
- Données intégrées
  - Normalisation des données
  - Définition d'un référentiel unique



Sources : Lydie Soler, AgroTechParis

Caractéristiques d'un DW (3)

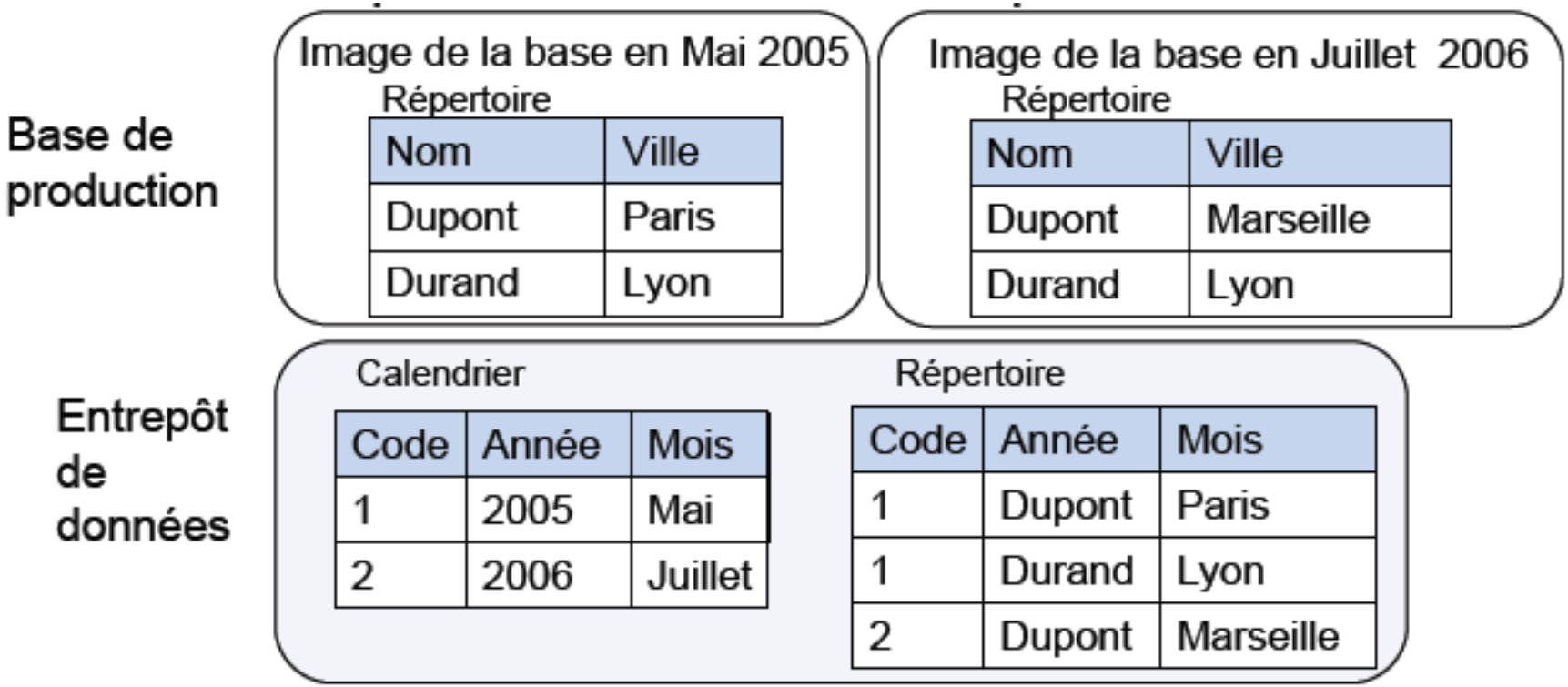
- Données non volatiles
  - Traçabilité des informations et des décisions prises
  - Copie des données de production



Sources : Lydie Soler, AgroTechParis

Caractéristiques d'un DW (4)

- Données historisées / datées
  - Les données persistent dans le temps
  - Mise en place d'un référentiel temps



Sources : Lydie Soler, AgroTechParis

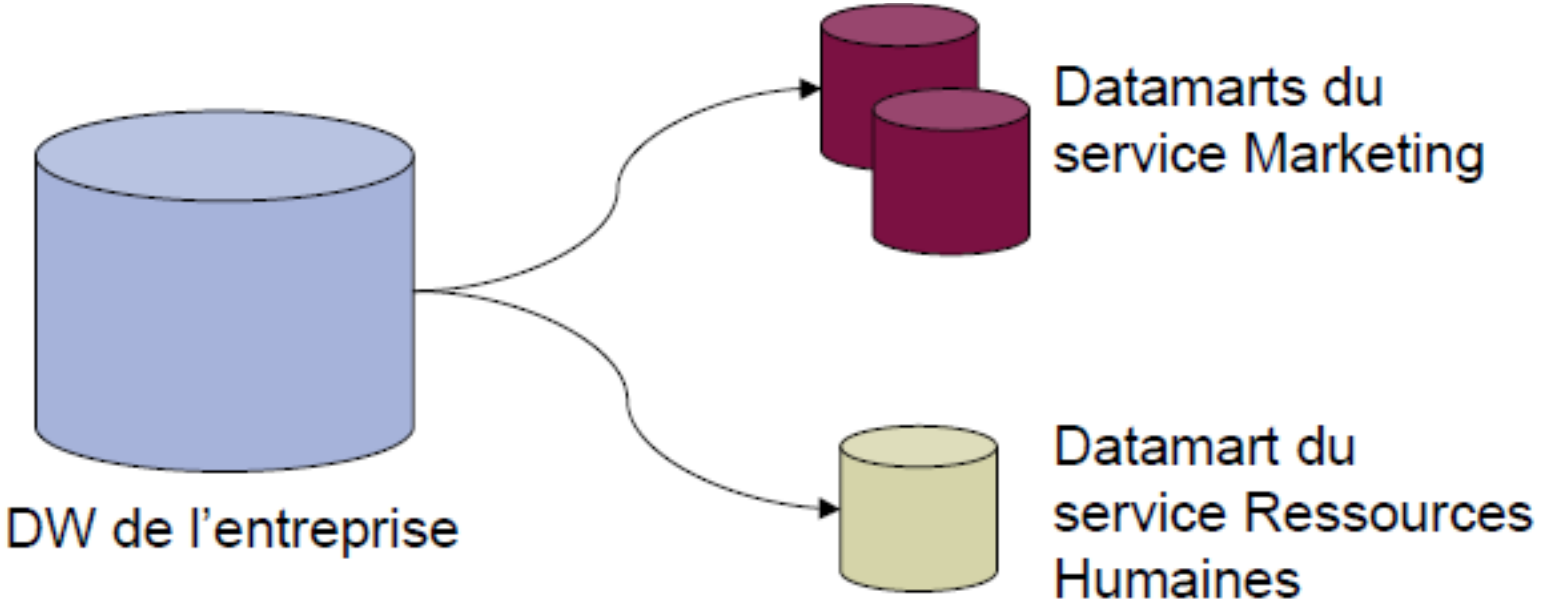
Caractéristiques d’un DW (5)

- Inconvénient :

De par sa taille, le DW est rarement utilisé directement par les décideurs car il contient plus que nécessaire pour une classe de décideurs

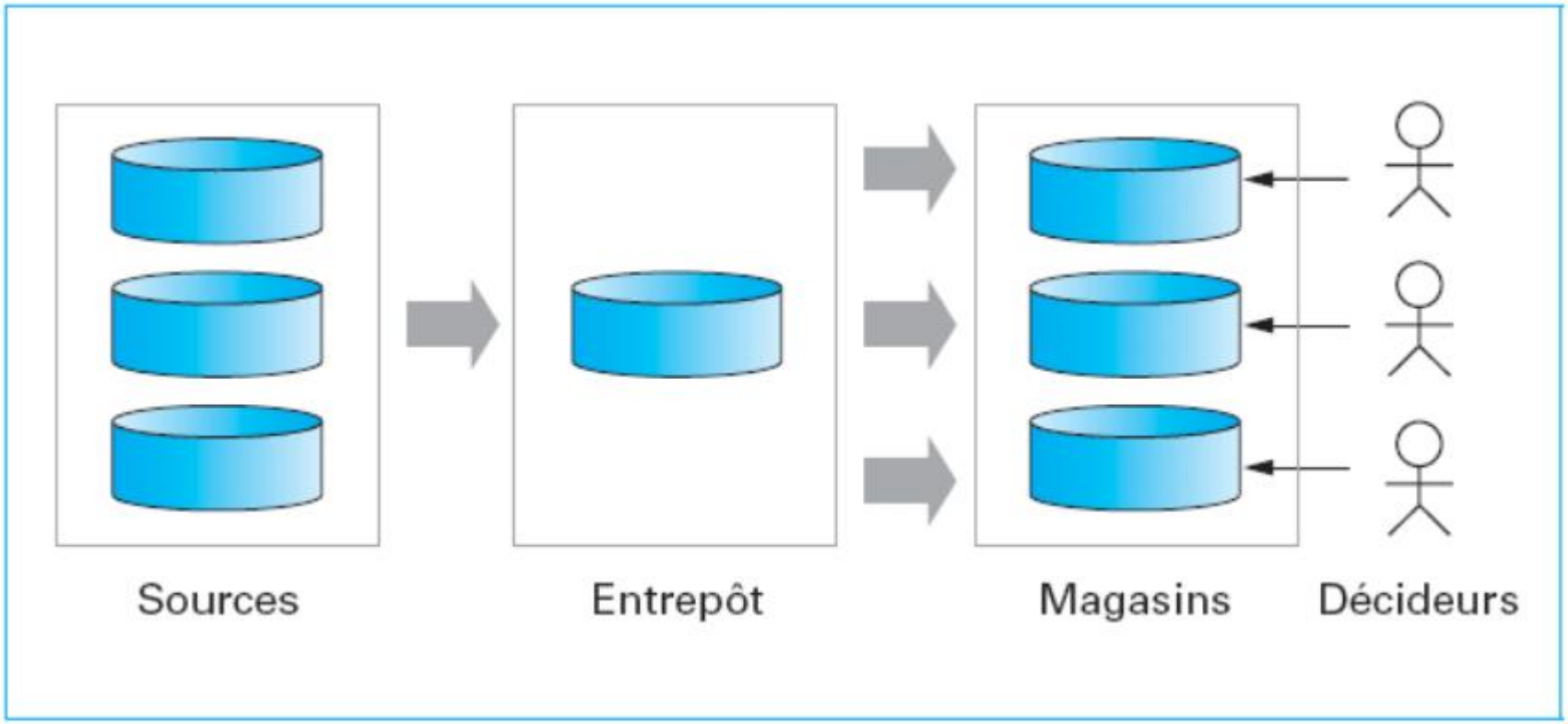
Le datamart

- Sous-ensemble d'un entrepôt de données
- Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
- Point de vue spécifique selon des critères métiers



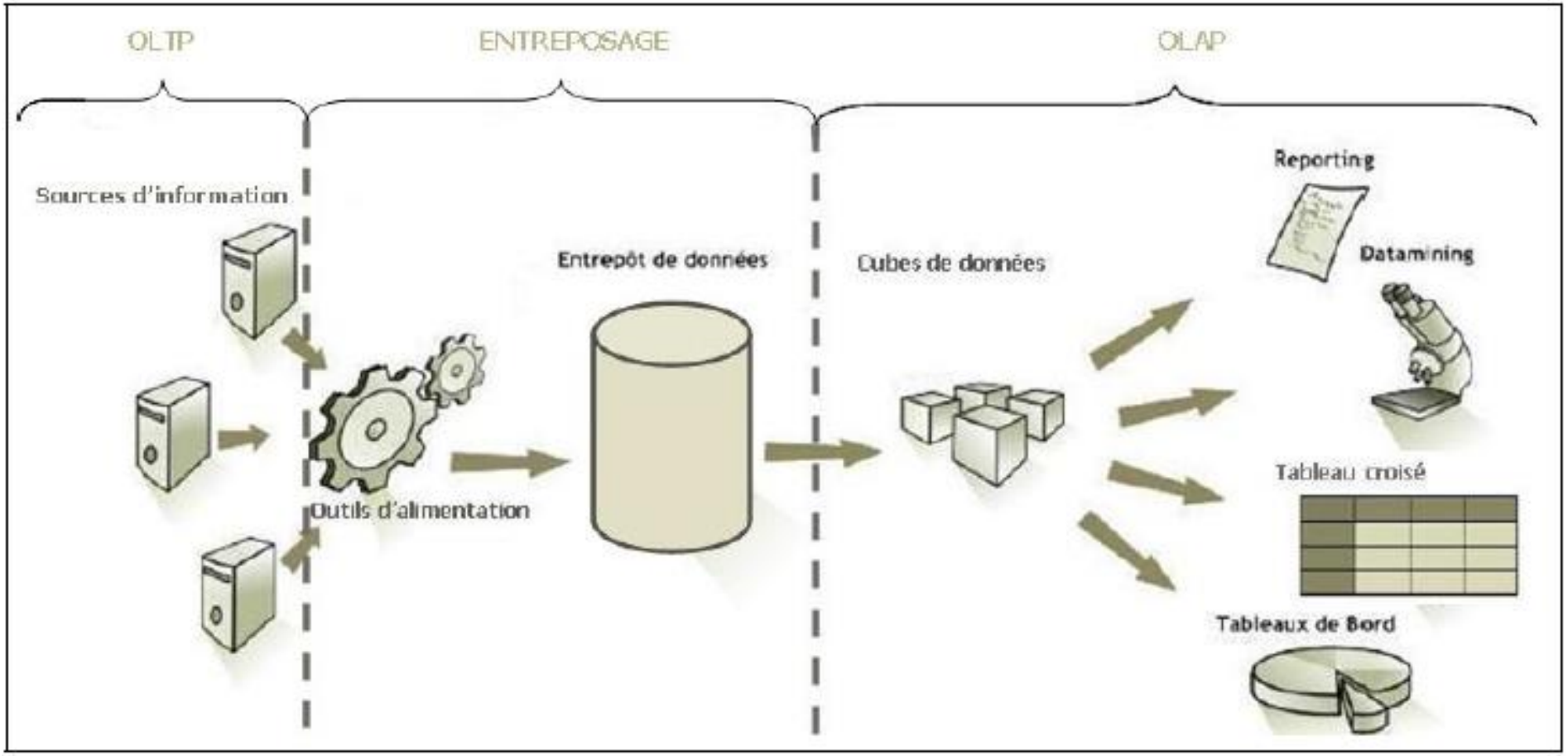
Sources : Lydie Soler, AgroTechParis

Architecture d'un système décisionnel



Sources : C. Chrisment, IRIT

Plus en détails…



Modélisation multidimensionnelle

- Niveau conceptuel
- Niveau logique
- Niveau physique

Niveau conceptuel

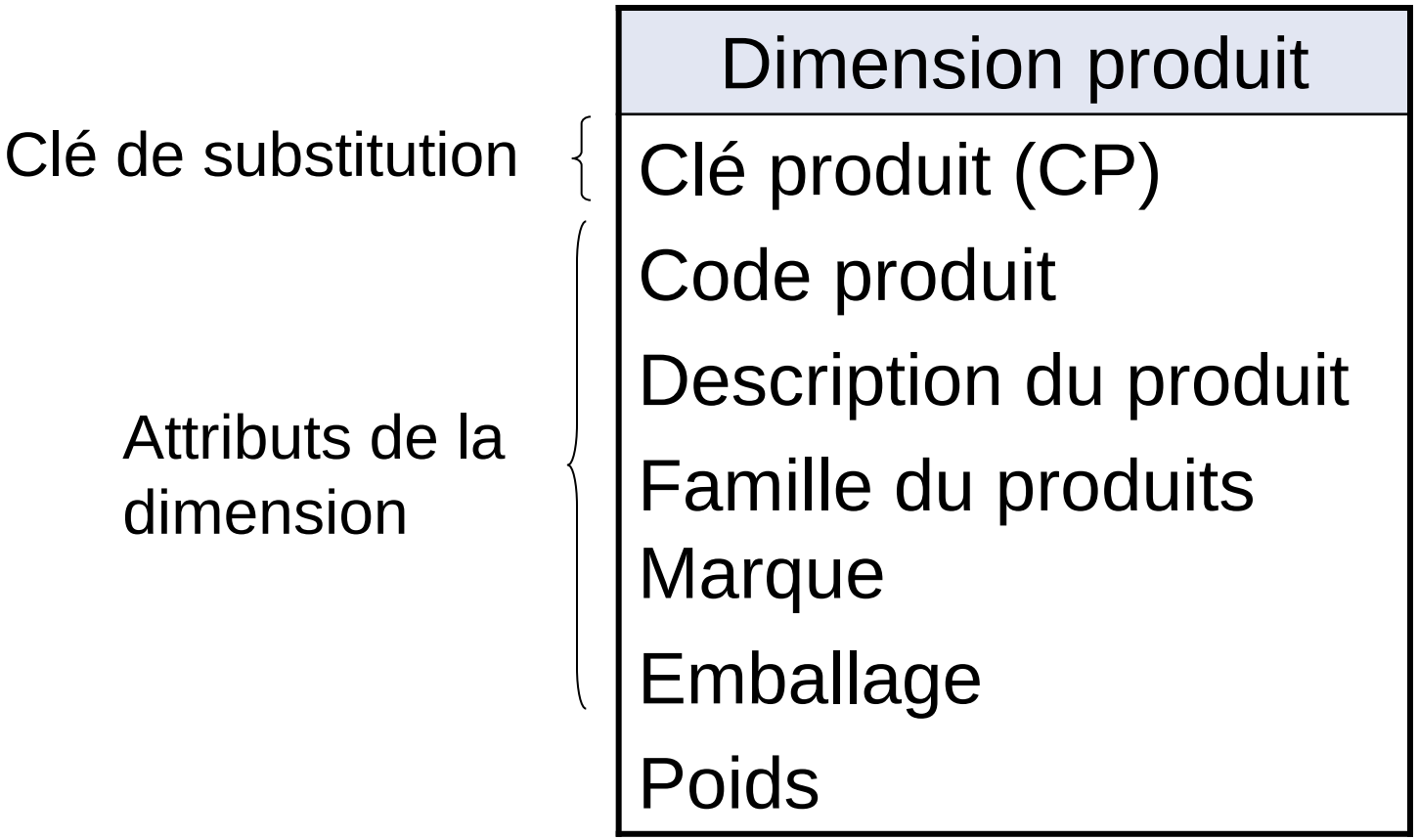
- Description de la base multidimensionnelle indépendamment des choix d'implantation

- Les concepts:
  - Dimensions et hiérarchies
  - Faits et mesures

Dimension (1)

- Axes d'analyse avec lesquels on veut faire l'analyse
  - Géographique, temporel, produits, etc.
- Chaque dimension comporte un ou plusieurs attributs/membres
- **Une dimension est tout ce qu'on utilisera pour faire nos analyses**.
- Chaque membre de la dimension a des caractéristiques propres et est en général textuel
- ***Remarque importante*** :
  - tables de dimension << Table de fait

Dimension (2)

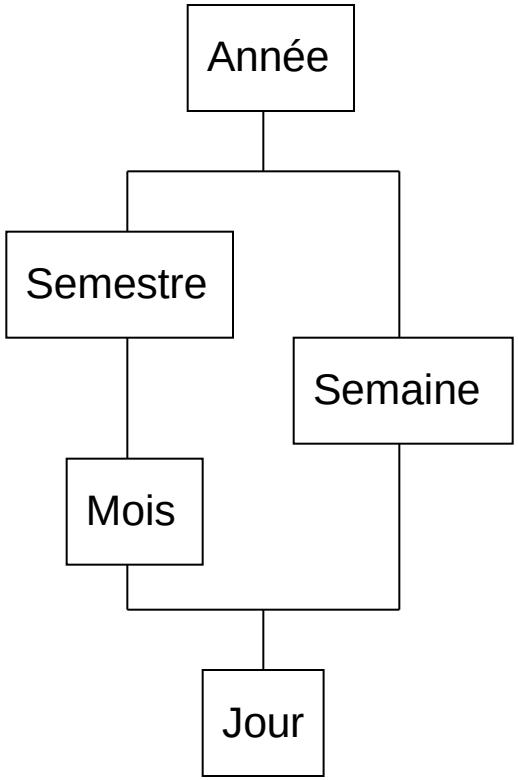

Dimension produit
Clé de substitution
Clé produit (CP)
Attributs de la dimension
Code produit
Description du produit
Famille du produits
Marque
Emballage
Poids

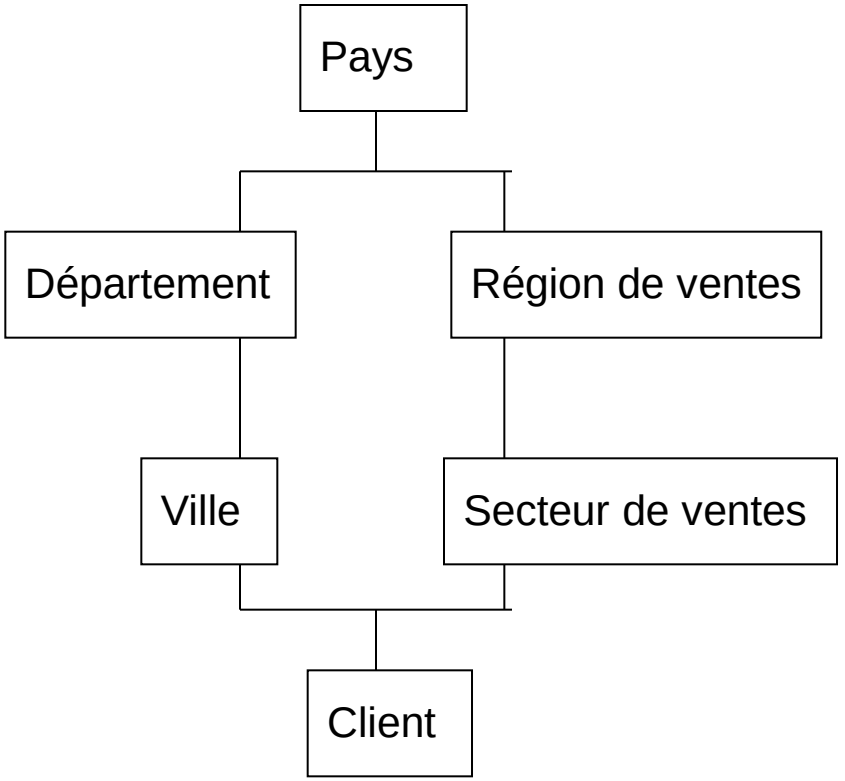
Hiérarchie (1)

- Les attributs/membres d'une dimension sont organisés suivant des hiérarchies
  - Chaque membre appartient à un niveau hiérarchique (ou niveau de granularité) particulier
  - Exemples :
    - Dimension temporelle : jour, mois, année
    - Dimension géographique : magasin, ville, région, pays
    - Dimension produit : produit, catégorie, marque, etc.
- Attributs définissant les niveaux de granularité sont appelés paramètres
- Attributs informationnels liés à un paramètre sont dits attributs faibles

Hiérarchie (2)

- Hiérarchies multiples dans une dimension

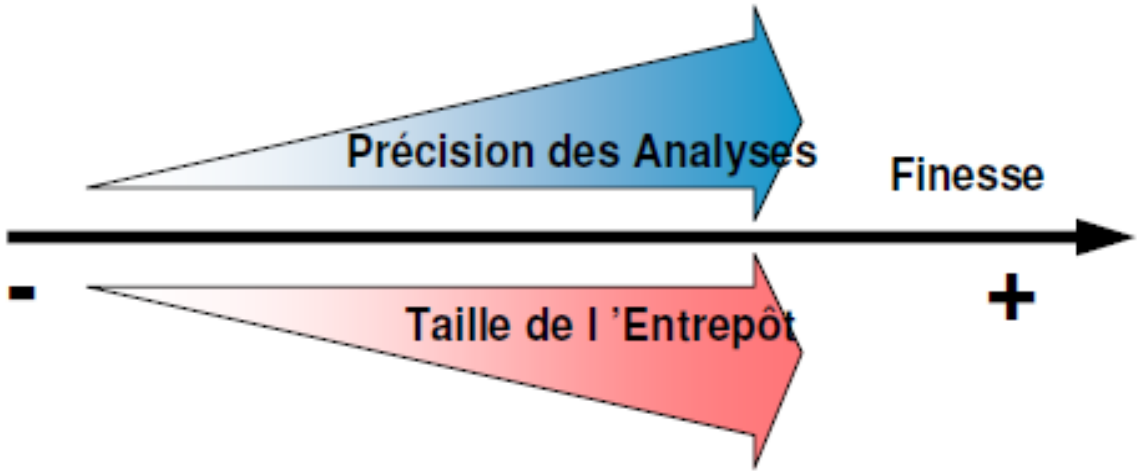




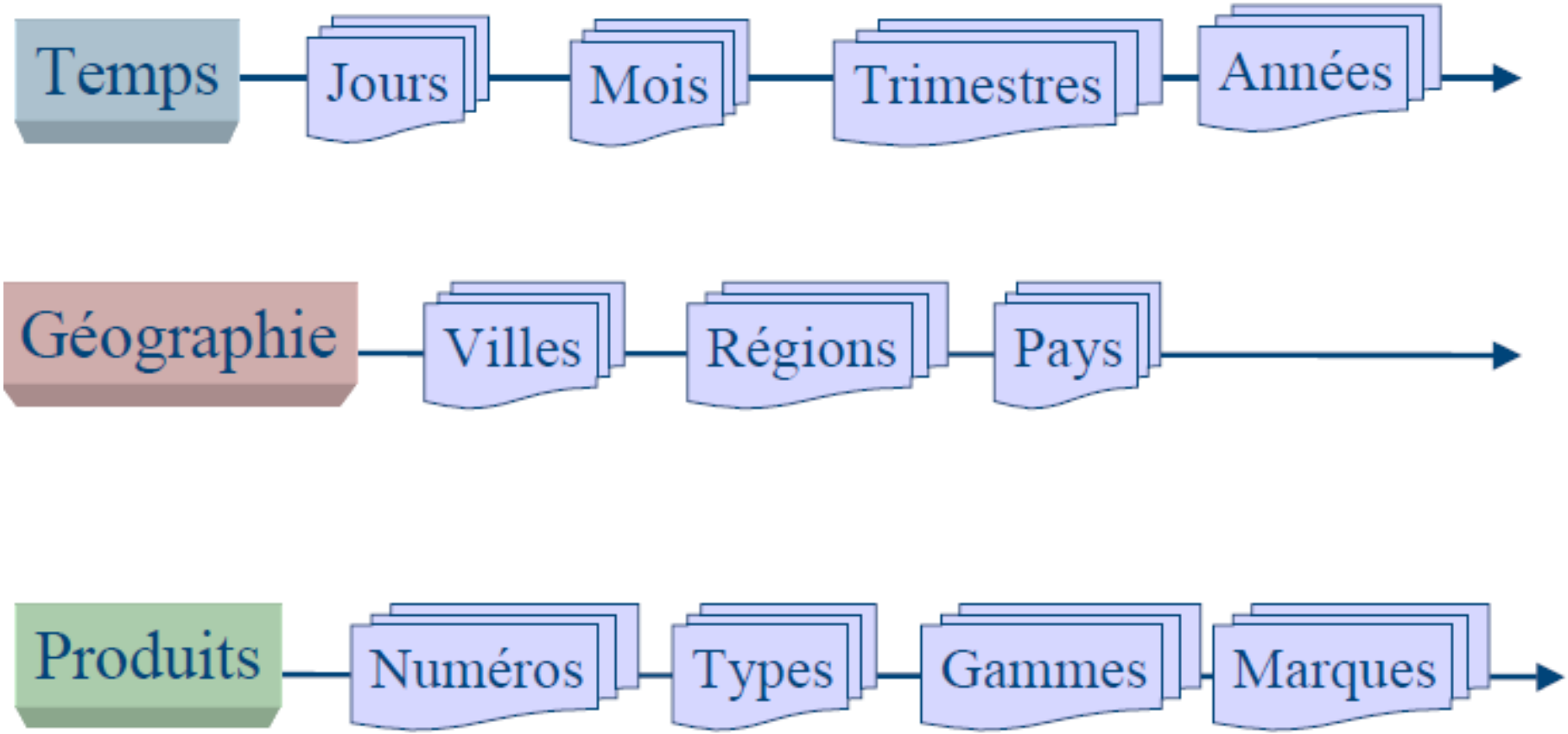
Granularité (1)

- Niveau de détail de représentation
  - Journée > heure du jour
  - Magasin > rayonnage

- Choix de la granularité



Granularité (2)



Fait

- Sujet analysé
- un ensemble d'attributs appelés mesures (informations opérationnelles)
  - les ventes (chiffre d'affaire, quantités et montants commandés, volumes des ventes, ...)
  - les stocks (nombre d'exemplaires d'un produit en stock, ...),
  - les ressources humaines (nombre de demandes de congés, nombre de démissions, …).

- Un fait représente la valeur d'une mesure, calculée ou mesurée, selon un membre de chacune des dimensions

- **Un fait est tout ce qu'on voudra analyser**.
  - Exemple : ***250 000 euros*** est un fait qui exprime la valeur de la mesure ***Coût des travaux*** pour le membre ***2002*** du niveau ***Année*** de la dimension ***Temps*** et le membre ***Versailles*** du niveau ***Ville*** de la dimension ***Découpage administratif***.

- La table de fait contient les valeurs des mesures et les clés vers les tables de dimensions

Mesure

- Élément de donnée sur lequel portent les analyses, en fonction des différentes dimensions.

- Ces valeurs sont le résultat d'opérations d'agrégation sur les données
  - Exemple :
    - Coût des travaux
    - Nombre d'accidents
    - Ventes
    - …

Clés

- Tables de dimension
  - Clé primaire
- Tables de fait
  - Clé composée
    - Clés étrangères des tables de dimension

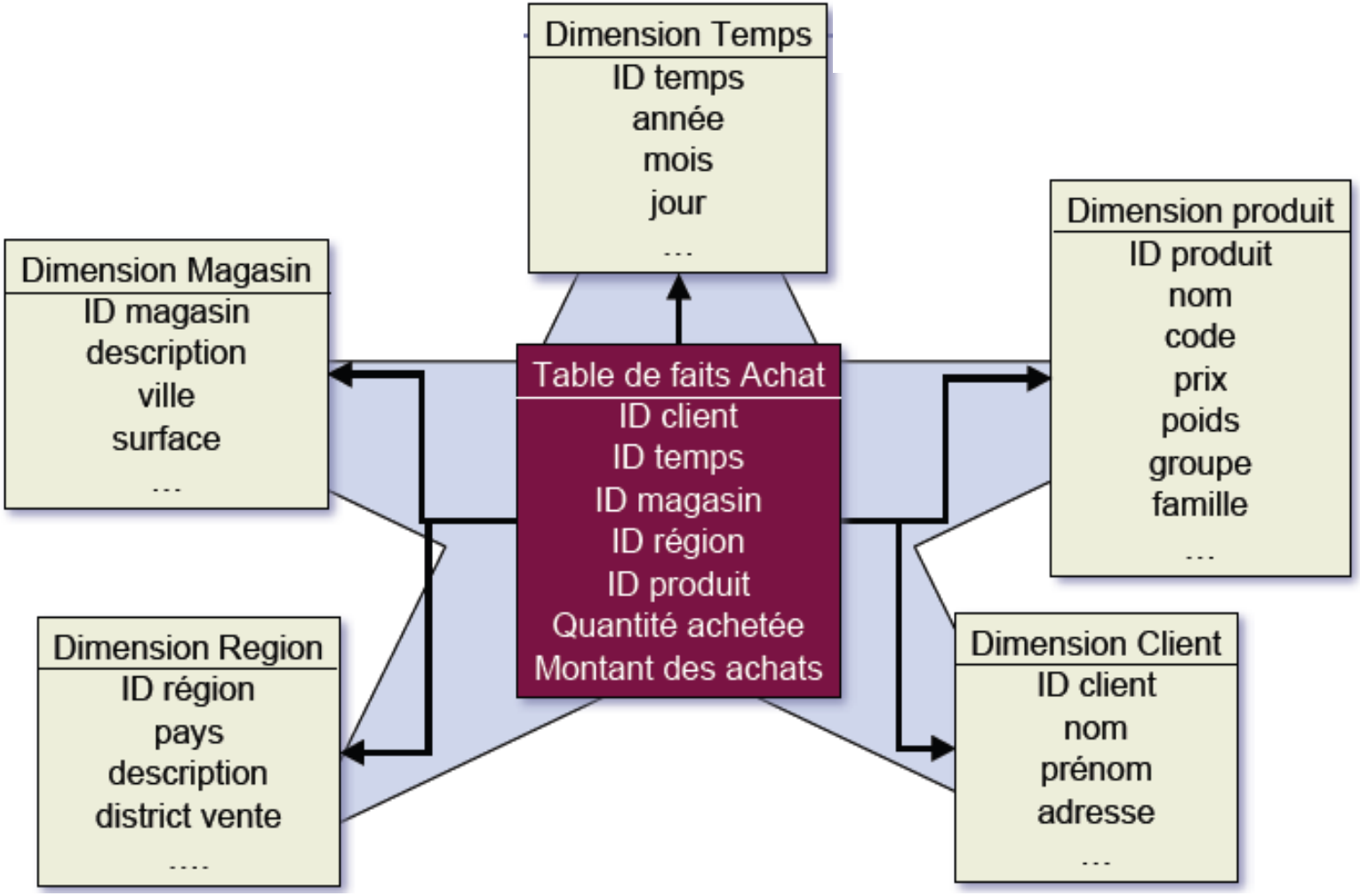
Modélisation

- Au niveau conceptuel, il existe 2 modèles :
  - en étoile (*star schema*)
  - ou en constellation (*fact constellation schema*)

Modèle en étoile (1)

- Une table de fait centrale et des dimensions
- Les dimensions n'ont pas de liaison entre elles
- Avantages :
  - Facilité de navigation
  - Nombre de jointures limité
- Inconvénients :
  - Redondance dans les dimensions
  - Toutes les dimensions ne concernent pas les mesures

Modèle en étoile (2)

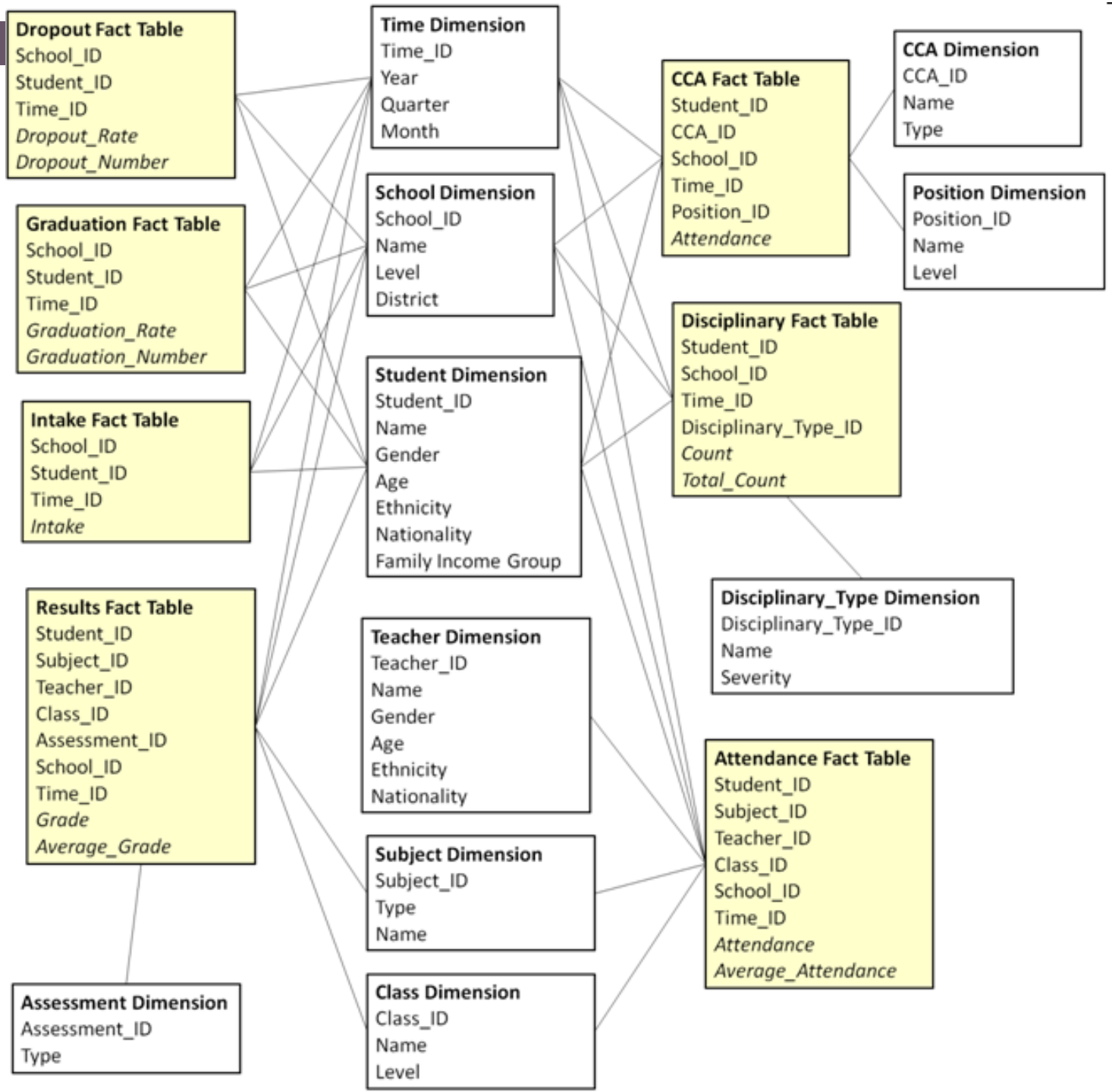


Sources : Lydie Soler, AgroTechParis

Constellation (1)

- Série d’étoiles
  - Fusion de plusieurs modèles en étoile qui utilisent des dimensions communes
  - Plusieurs tables de fait et tables de dimensions, éventuellement communes

Constellation (2)



Sources : http://gankahhwee.com

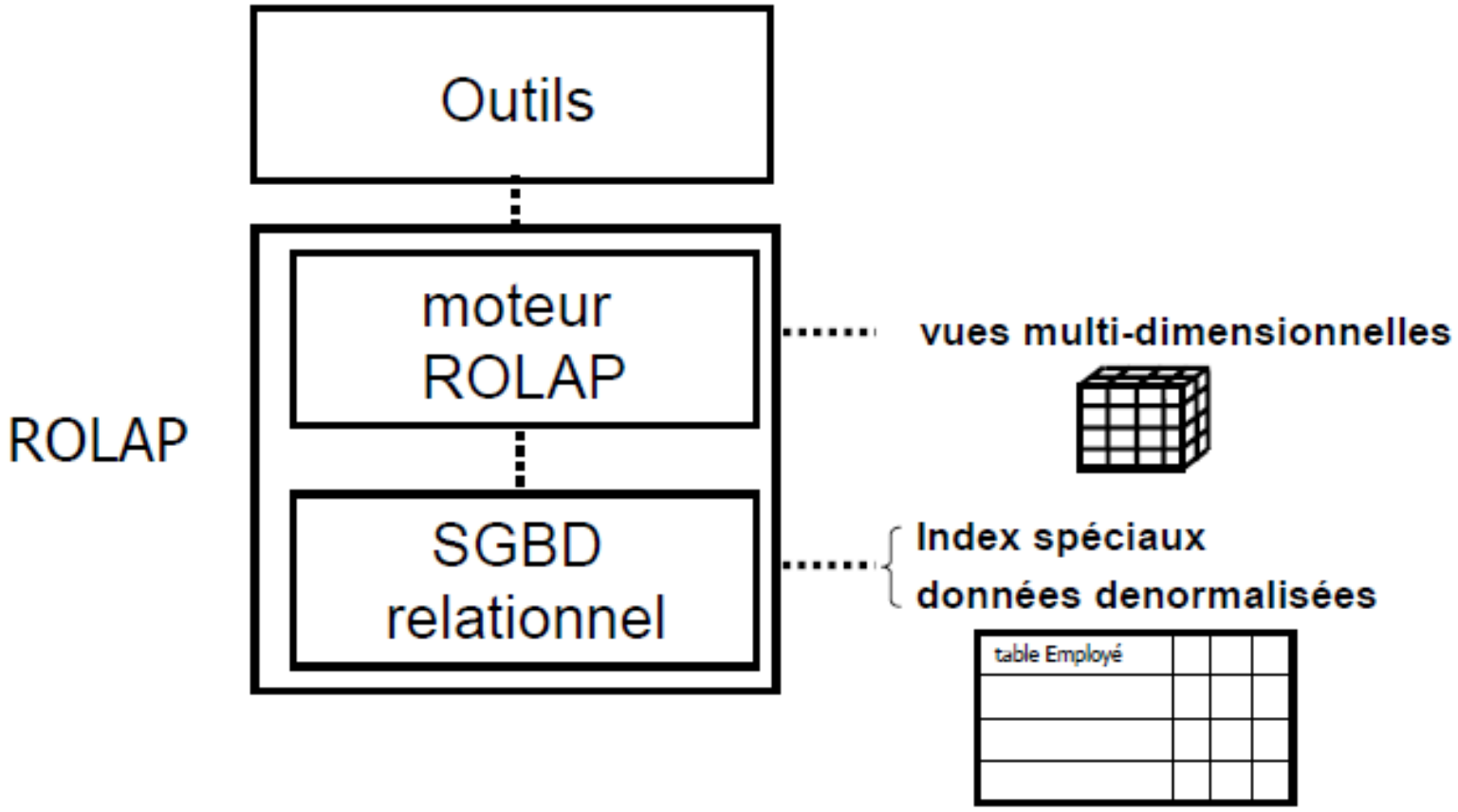
Niveau logique

- Description de la base multidimensionnelle suivant la technologie utilisée :
  - ROLAP (*Relational-OLAP*)
  - MOLAP (*Multidimensional-OLAP*)
  - HOLAP (*Hybrid-OLAP*)

ROLAP (1)

- Les données sont stockées dans une BD relationnelle
- Un moteur OLAP permet de simuler le comportement d'un SGBD multidimensionnel
- Avantages :
  - Facile à mettre en place
  - Peu couteux
  - Evolution facile
  - Stockage de gros volumes
- Inconvénients :
  - Moins performant lors des phases de calculs
- Exemple de moteur ROLAP : Mondrian

ROLAP (2)

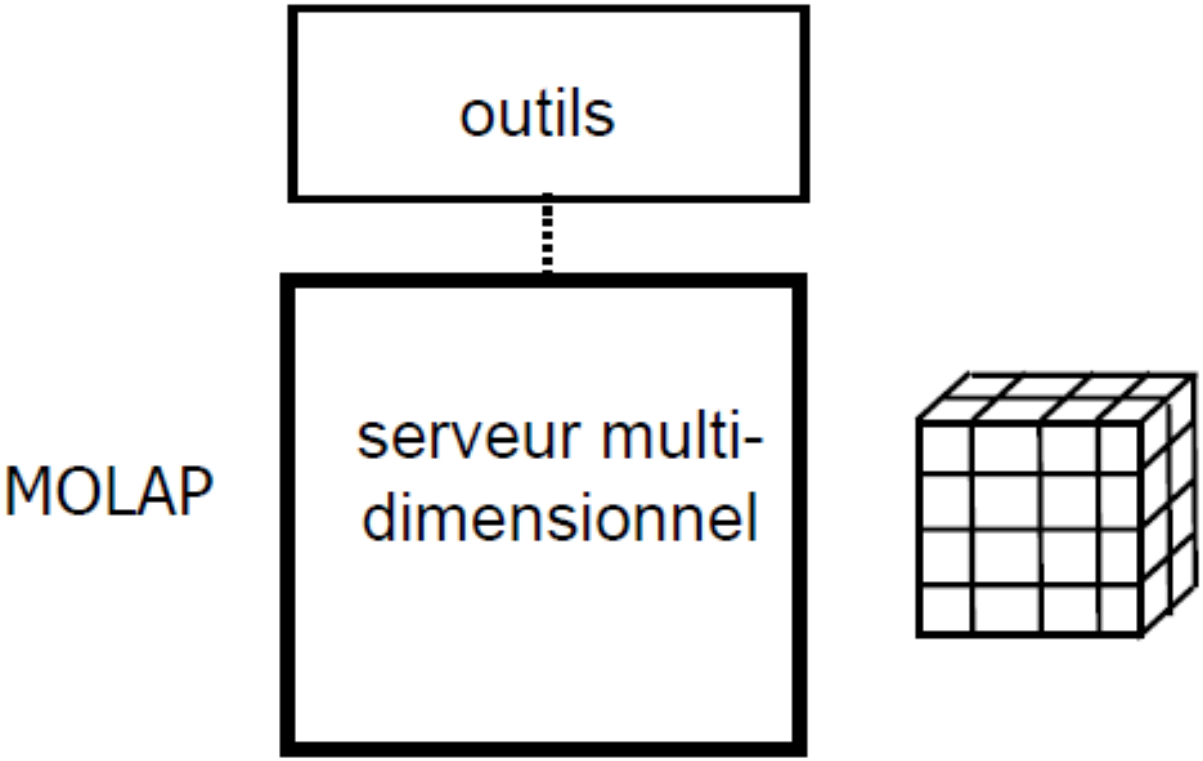


Sources : EPFL, Lausanne

MOLAP (1)

- Les données sont stockées comme des matrices à plusieurs dimensions : Cube[1:m,1:n,1:p](mesure)
- Accès direct aux données dans le cube
- Avantages :
  - Rapidité
- Inconvénients :
  - Difficile à mettre en place
  - Formats souvent propriétaires
  - Ne supporte pas de rtès gros volumes de données
- Exemple de moteurs MOLAP :
  - Microsoft Analysis Services
  - Hyperion

MOLAP (2)



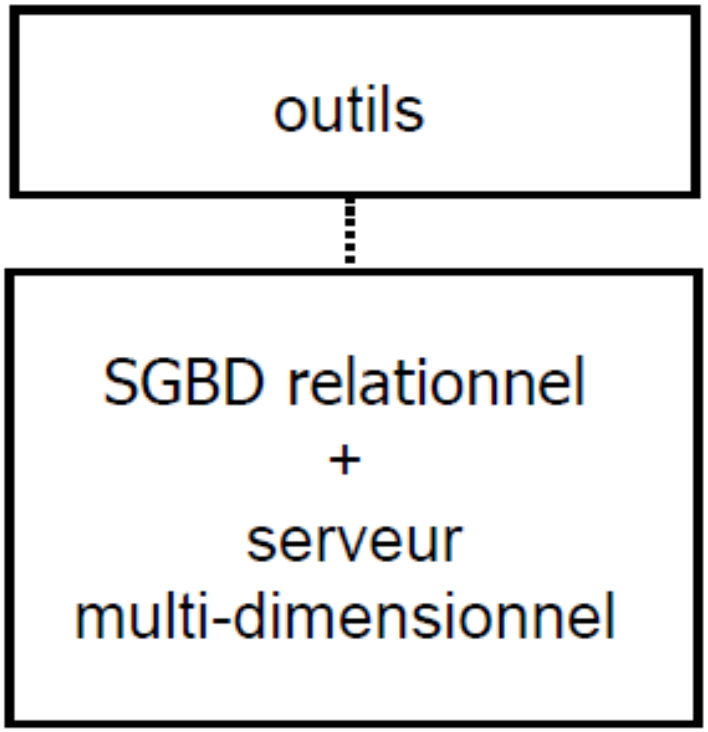
Sources : EPFL, Lausanne

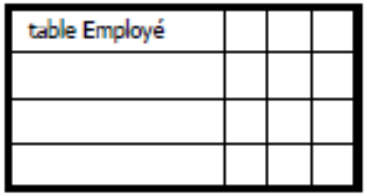
HOLAP (1)

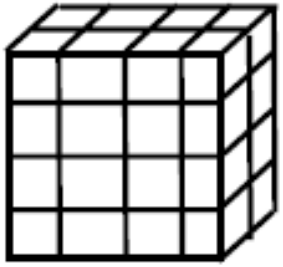
- Solution hybride entre ROLAP et MOLAP
- Données de base stockées dans un SGBD relationnel (tables de faits et de dimensions) + données agrégées stockées dans un cube
- Avantages / inconvénients :
  - Bon compromis au niveau des coûts et des performances (les requêtes vont chercher les données dans les tables et le cube)

HOLAP (2)

HOLAP







Sources : EPFL, Lausanne

Modélisation

- Au niveau logique, il existe 1 modèle :
  - en flocon (*snowflake schema*)

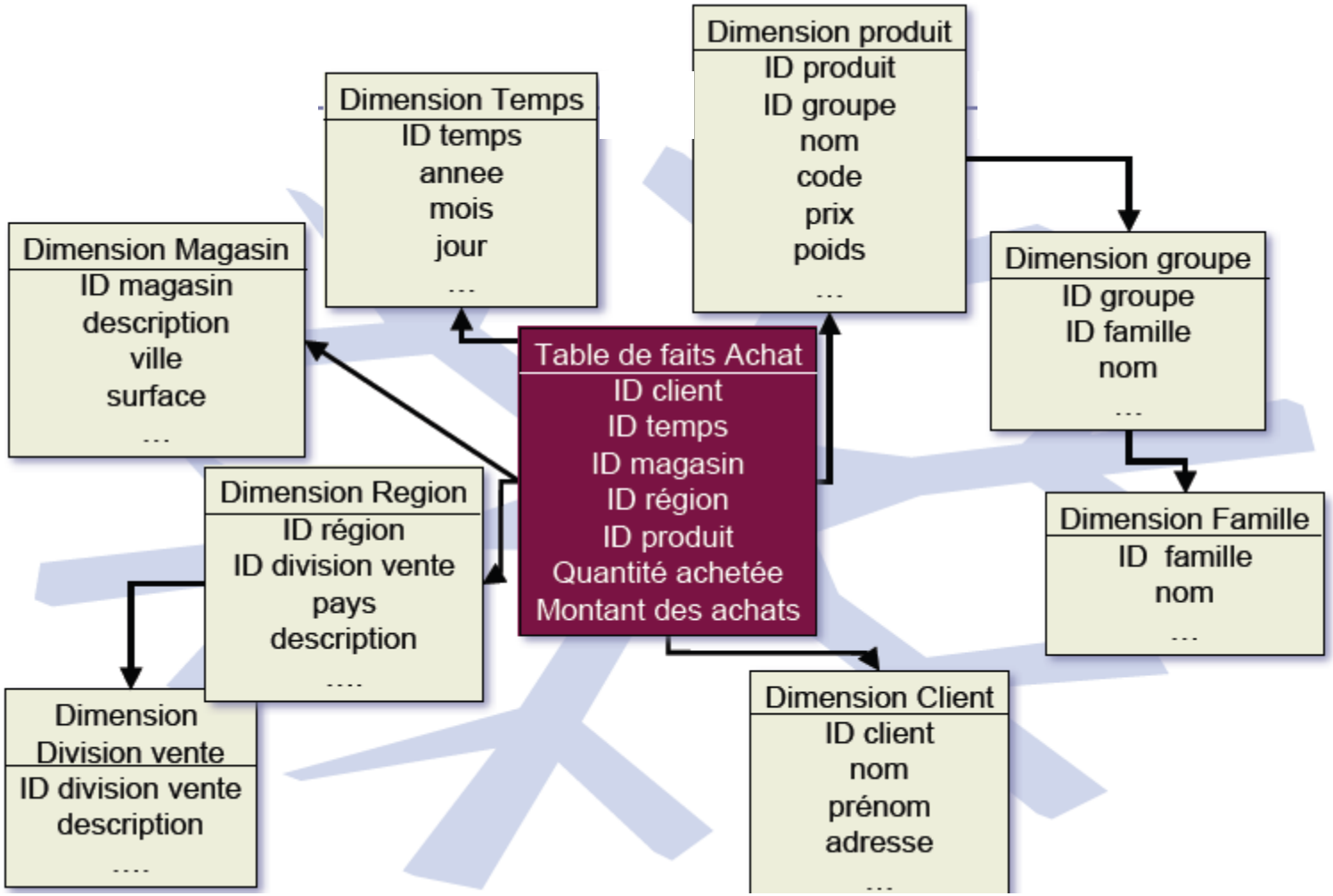
Modèle en flocon (1)

- Modèle en étoile + normalisation des dimensions
  - Une table de fait et des dimensions 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 (elle a la granularité la plus fine)

- Avantages :
  - Normalisation des dimensions
  - Economie d'espace disque (réduction du volume)

- Inconvénients :
  - Modèle plus complexe (nombreuses jointures)
  - Requêtes moins performantes
  - Navigation difficile

Modèle en flocon (2)

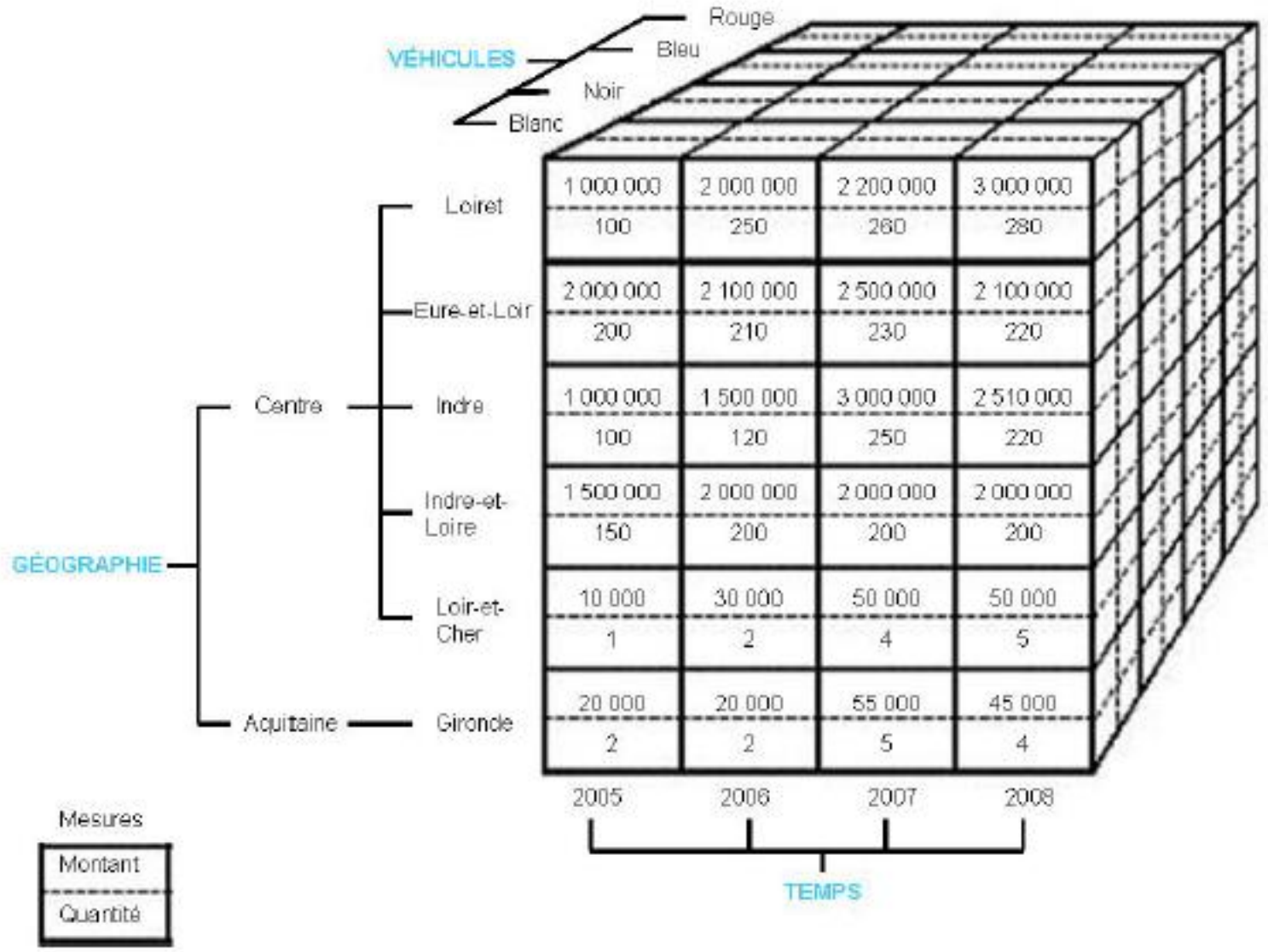


Sources : Lydie Soler, AgroTechParis

Cube (1)

- 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.

Cube (2)


VÉHICULES
Rouge
Bleu
Noir
Blanc
GÉOGRAPHIE
Centre
Acquitaine
Loiret
Eure-et-Loir
Indre
Indre-et-Loire
Loir-et-Cher
Gironde
1 000 000
100
2 000 000
250
2 200 000
260
3 000 000
280
2 000 000
200
2 100 000
210
2 500 000
230
2 100 000
220
1 000 000
100
1 500 000
120
3 000 000
250
2 510 000
220
1 500 000
150
2 000 000
200
2 000 000
200
2 000 000
200
10 000
1
30 000
2
50 000
4
50 000
5
20 000
2
20 000
2
55 000
5
45 000
4
2005
2006
2007
2008
TEMPS
Mesures
Montant
Quantité

Niveau physique

- C'est l'implantation et dépend donc du logiciel utilisé.

- Globalement : insuffisance des instructions SQL classiques
  - CREATE TABLE … AS … : recopie physique, à reprendre intégralement lors de l'évolution des sources
  - CREATE VIEW … AS … : recalculé à chaque requête, temps de réponse inacceptable sur les volumes manipulés

- Optimisation : indexes, …

Réalisation d'un DW

- Evolution des besoins et des sources
  → démarche itérative

- 3 techniques :
  - Top-down [Inmon]
  - Bottom-up [Kimball]
  - Middle-out

Top-Down

- Concevoir tout l'entrepôt intégralement
  - Il faut donc connaître à l'avance toutes les dimensions et tous les faits.
- Objectif : Livrer une solution technologiquement saine basée sur des méthodes et technologies éprouvées des bases de données.
- Avantages :
  - Offrir une architecture intégrée : méthode complète
  - Réutilisation des données
  - Pas de redondances
  - Vision claire et conceptuelle des données de l'entreprise et du travail à réaliser
- Inconvénients :
  - Méthode lourde
  - Méthode contraignante
  - Nécessite du temps

- **Bottom-Up** (approche inverse)
  - Créer les datamarts un par un puis les regrouper par des niveaux intermédiaires jusqu'à obtention d'un véritable entrepôt.
  - Objectif : Livrer une solution permettant aux usager d'obtenir facilement et rapidement des réponses à leurs requêtes d'analyse
  - Avantages :
    - Simple à réaliser,
    - Résultats rapides
    - Efficace à court terme
  - Inconvénients :
    - Pas efficace à long terme
    - Le volume de travail d'intégration pour obtenir un entrepôt de données
    - Risque de redondances (car réalisations indépendantes).

- **Middle-Out** (approche hybride)
  - Concevoir intégralement l'entrepôt de données (toutes les dimensions, tous les faits, toutes les relations), puis créer des divisions plus petites et plus gérables.
  - Avantages :
    - Prendre le meilleur des 2 approches
    - Développement d'un modèle de données d'entreprise de manière itérative
    - Développement d'une infrastructure lourde qu'en cas de nécessité
  - Inconvénients :
    - implique, parfois, des compromis de découpage (dupliquer des dimensions identiques pour des besoins pratiques).

Ne pas oublier… (1)

- Le volume de données manipulées

Grandes distribution :

CA annuel : 80 000 M$
Prix moyen d'un article d'un ticket : 5$
Nbre d'articles vendus pour un an : $80 * 10^9 / 5 = 16 * 10^9$
Volume du DW :

*$16*10^9$ *3 ans * 24 octets =* **1,54 To** *($1,54*10^{12} = 1\,540$ Go)*

Téléphonie :

Nbre d'appels quotidiens : 100 millions
Historique : 3 ans * 365 jours = 1 095 jours
Volume du DW :

*100 millions * 1 095 jours * 24 octets =* **3,94 To**

Cartes de crédit :

Nbre de clients : 50 millions
Nbre moyen mensuel de transactions : 30
Volume :

*50 millions * 26 mois * 30 transactions * 24 octets =* **1,73 To**

Ne pas oublier… (2)

- Voici 5 étapes importantes pour la réalisation d'un DW :
  - Conception
  - Acquisition des données
  - Définition des aspects techniques de la réalisation
  - Définition des modes de restitution
  - Stratégies d'administration, évolution, maintenance

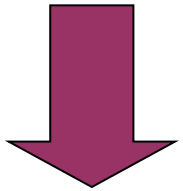
1 - Conception

- Définir la finalité du DW :
  - Quelle activité de l'entreprise faut-il piloter?
  - Quel est le processus de l'entreprise à modéliser?
  - Qui sont les décideurs?
  - Quels sont les faits numériques?
    - Qu'est ce qui va être mesurer?
  - Quelles sont les dimensions ?
    - Comment les gestionnaires décrivent-ils des données qui résultent du processus concerné?

- Définir le modèle de données :
  - Modèle en étoile / flocon ?
  - et/ou Cube?
  - et/ou Vues matérialisées?

2 – Acquisition des données

- Pour l'alimentation ou la mise à jour de l'entrepôt
  - Mise à jour régulière



Besoin d'un outil pour automatiser les chargements de l'entrepôt :

ETL (*Extract, Transform, Load*)

ETL :

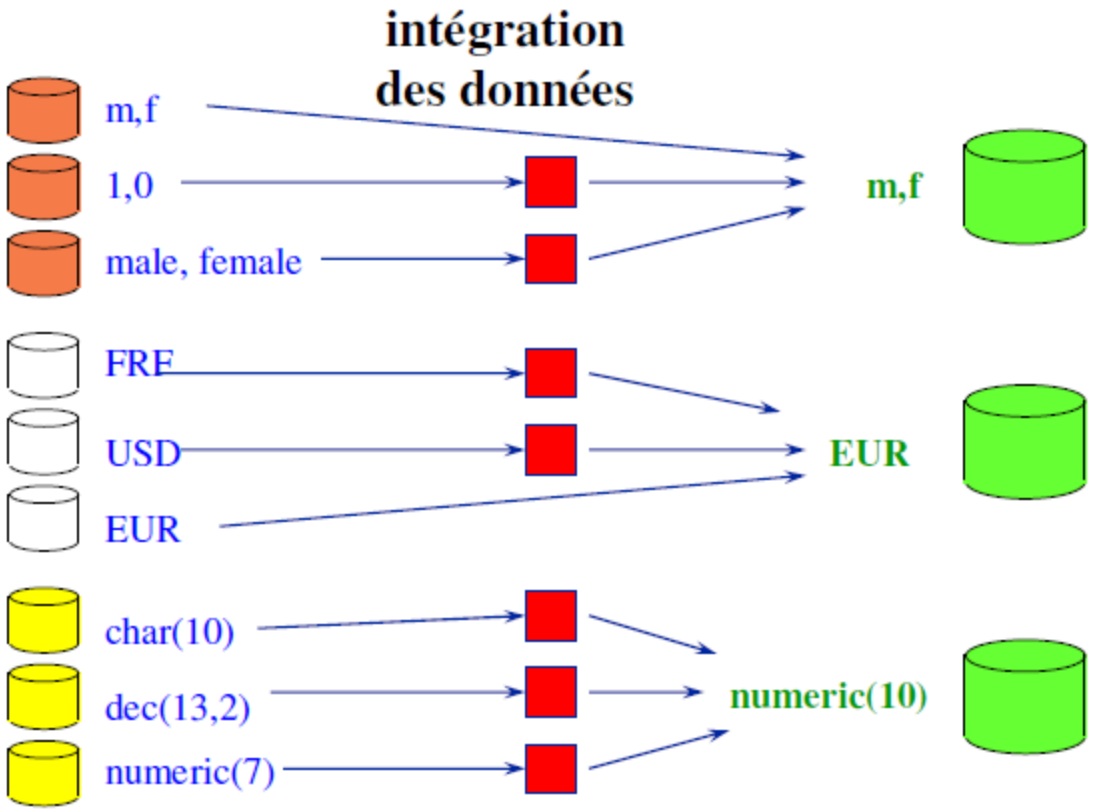
- Modèle entité-relation (BD de production)
  → Modèle à base de dimensions et de faits

- Outil :
  - Offrant un environnement de développement
  - Offrant des outils de gestion des opérations et de maintenance
  - Permettant de découvrir, analyser, et extraire les données à partir de sources hétérogènes
  - Permettant de nettoyer et standardiser les données
  - Permettant de charger les données dans un entrepôt

- Extraction :
  - Depuis différentes sources (bd, fichiers, journaux, …)
  - Différentes techniques :
    - Push : règles (triggers)
    - Pull : requêtes (queries)
  - Périodique et répétée
    - Dater ou marquer les données envoyées
  - Difficulté :
    - Ne pas perturber les applications OLTP

- Transformation : Etape très importante qui garantit la cohérence et la fiabilité des données

  - Rendre cohérentes les données issues de différentes sources
    - Unifier les données
      - Ex. dates : MM/JJ/AA -> JJ/MM/AA
      - Ex. noms : D-Naiss, Naissance, Date-N -> « Date-Naissance »
    - Trier, Nettoyer
      - Eliminer les doubles
      - Jointures, projection, agrégation (SUM, AVG, …)
      - Gestion des valeurs manquantes (NULL) (ignorer ou corriger ?)
      - Gestion des valeurs erronées ou inconsistantes (détection et correction)
      - Vérification des contraintes d'intégrité (pas de violation)
    - Inspection manuelle de certaines données possible…


intégration
des données
m,f
1,0
male, female
m,f
FRF
USD
EUR
EUR
char(10)
dec(13,2)
numeric(7)
numeric(10)

- Chargement : Insérer ou modifier les données dans l'entrepôt
  - Alimentation incrémentale ou totale?, offline ou online?, fréquence des chargements?, taille de l'historique?, …
  - Si pas de MAJ :
    - insertion de nouvelles données
    - Archivage des données anciennes
  - Sinon (attention en cas de gros volumes)
    - Périodicité parfois longue
    - MAJ des indexes et des résumés

Attention…

- ETL ≠ ELT
  - L'approche ELT (Extraction, Loading, Transformation) génère du code SQL natif pour chaque moteur de BD impliqué dans le processus – sources et cibles
  - Cette approche profite des fonctionnalités de chaque BD mais les requêtes de transformation doivent respecter la syntaxe spécifique au SGBD

3 – Aspects techniques

- Contraintes
  - logicielles,
  - matérielles,
  - humaines,
  - …

4 - Restitution

- = But du processus d'entreposage,
- = Conditionne souvent le choix de l'architecture et de la construction du DW
- Toutes les analyses nécessaires doivent être réalisables !

- Types d'outils de restitution :
  - Requêteurs et outils d'analyse
  - Outils de data mining

5 – Administration, maintenance

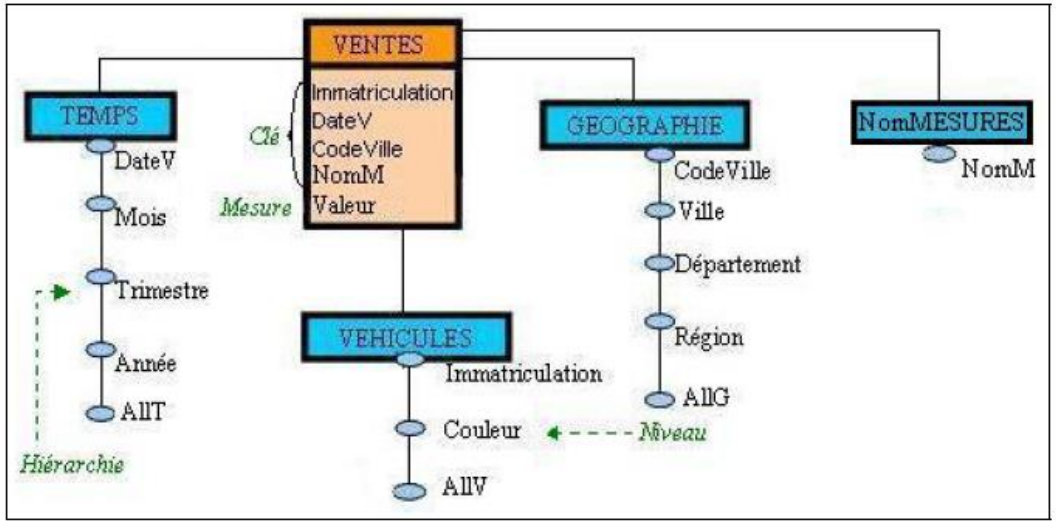
- Toutes les stratégies à mettre en place pour l'administration, l'évolution et la maintenance
  - Ex : fréquences des rafraichissements (global ou plus fin?)

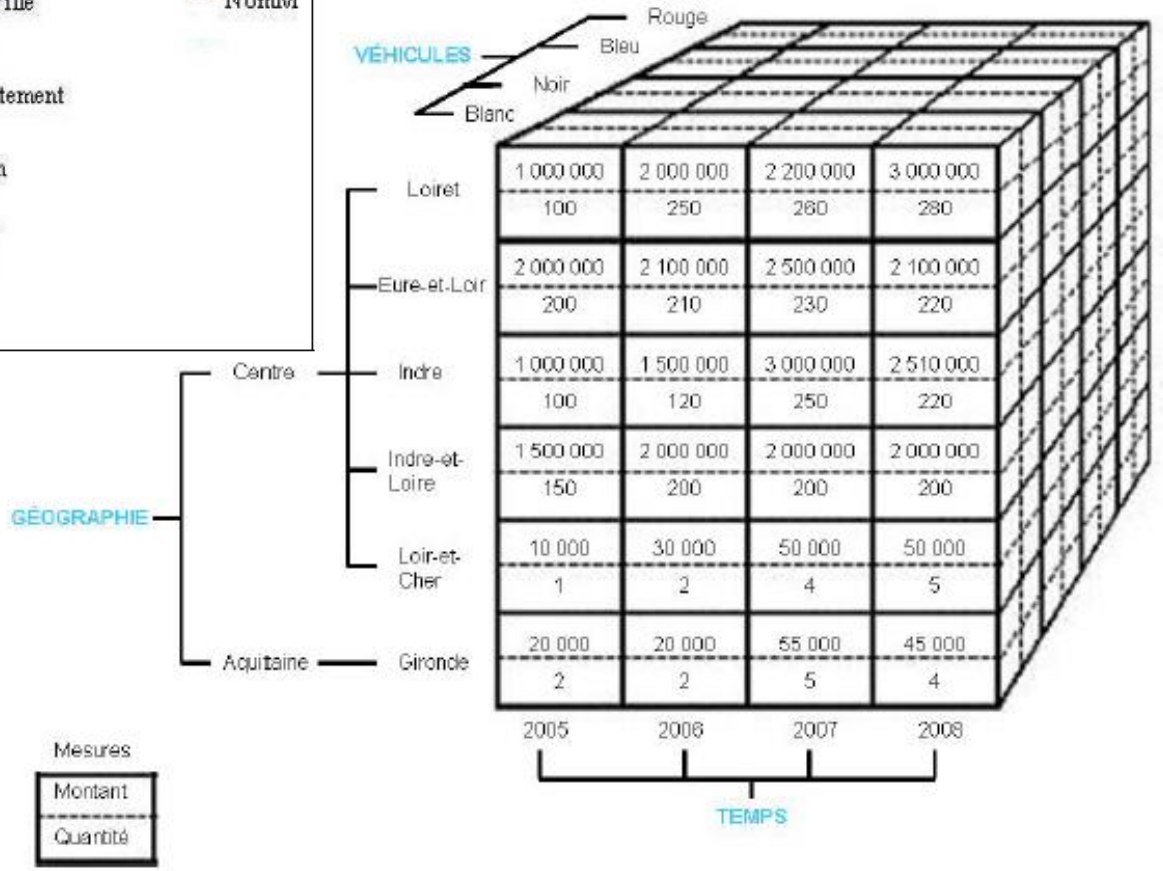
Représentation et manipulation (1)

- Le cube de données
- est traditionnellement représenté sous forme de table multidimensionnelle
- et manipulé via différents opérateurs

Représentation et manipulation (2)

- La table multidimensionnelle
  - Présente les valeurs des mesures d'un fait en fonction des valeurs des paramètres des dimensions représentées en lignes et en colonnes étant données des valeurs des autres dimensions
    - les lignes et les colonnes sont les axes selon lesquels le cube est exploré et chaque cellule contient la (ou les) mesure(s) calculée(s).
  - correspond à une tranche du cube multidimensionnel





Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Représentation et manipulation (3)

- Opérateurs de visualisation du cube (Cube -> Cube)
  - Transformation de la granularité des données (Forage)
  - Sélection / projection sur les données du cube
  - Restructuration / réorientation du cube

- Opérations de forage (liées à la granularité)
  - Roll-up (forage vers le haut) :
    - Représente les données à un niveau de granularité supérieur selon la hiérarchie de la dimension désirée
      - Agréger selon une dimension
        - Semaine -> Mois
  - Drill-down (forage vers le bas) :
    - Inverse du roll-up
    - Représente les données à un niveau de granularité inférieur
      - Détailler selon une dimension
        - Mois -> Semaine

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Roll-Up sur la dimension Géographie **Drill-down**

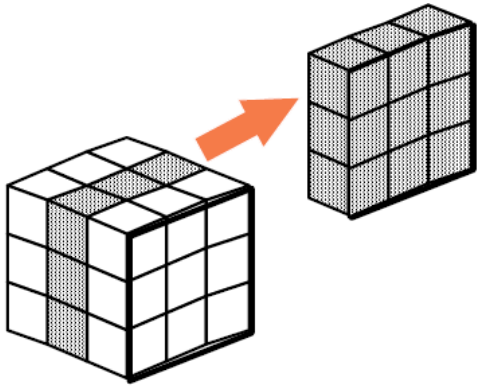
Quantité des ventes		Géographie.Région	
		Aquitaine	Centre
Temps.Année	2005	2	551
	2006	2	782
	2007	5	944
	2008	4	925
Véhicules.AllV			

Opérations de sélection / projection

- Slice :
  - Sélection
  - Tranche du cube obtenue par prédicats selon une dimension
    - Mois = « Avril 2004 »
- Dice :
  - Projection selon un axe
  - Sorte de cumuls de sélection
    - Projeter(Région, Produit)

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Slice (Année = « 2005 »)



Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
Véhicules.AllV							

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Dice (Département = « Loir et Cher » ou « Gironde », Année = « 2007 » ou « 2008 »)

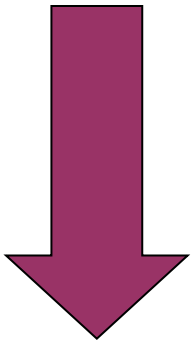
Quantité des ventes		Géographie.Département	
		Loir et Cher	Gironde
Temps.Année	2007	4	5
	2008	5	4
Véhicules.AllV			

- Opérations de restructuration / réorientation
  - Pivot (ou Rotate)
    - Tourne le cube pour visualiser une face différente
      - (Région, Produit) -> (Région, Mois)
  - Switch (ou Permutation)
    - Inter-change la position des membres d'une dimension
  - Nest
    - Imbrique des membres issus de dimensions différentes
  - Push (ou Enfoncement)
    - Combine les membres d'une dimension aux mesures (les membres deviennent le contenu des cellules)
  - AddM, DelM
    - Pour l'ajout et la suppression de mesures à afficher
  - …

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Pivot

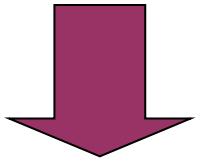
(Temps.Année, Géographie.Département
-> Temps.Année, Véhicules.Couleur)



Quantité des ventes		Véhicules.Couleur			
		Blanc	Noir	Bleu	Rouge
Temps.Année	2005	120	200	150	83
	2006	130	220	150	284
	2007	140	250	259	300
	2008	150	280	249	250
Géographie.AllG					

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

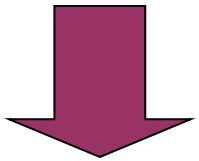
Nest (Véhicules.Couleur, Temps.Année)



Quantité des ventes		Géographie.Département					
Véhicules.Couleur	Temps.Année	Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Blanc	2005	...	...	...	...	...	...
	2006	...	...	...	...	...	...
	2007	...	...	...	...	...	...
	2008	...	...	...	...	...	...
Noir	2005	...	...	...	...	...	...
	2006	...	...	...	...	...	...
	2007	...	...	...	...	...	...
	2008	...	...	...	...	...	...
Bleu	2005	...	...	...	...	...	...
	2006	...	...	...	...	...	...
	2007	...	...	...	...	...	...
	2008	...	...	...	...	...	...
Rouge	2005	...	...	...	...	...	...
	2006	...	...	...	...	...	...
	2007	...	...	...	...	...	...
	2008	...	...	...	...	...	...

Quantité des ventes		Géographie.Département					
		Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
Temps.Année	2005	100	200	100	150	1	2
	2006	250	210	120	200	2	2
	2007	260	230	250	200	4	5
	2008	280	220	220	200	5	4
Véhicules.AllV							

Push (Véhicules.Couleur)



Quantité des ventes	Géographie.Département					
Temps.Année	Loiret	Eure et Loir	Indre	Indre et Loire	Loir et Cher	Gironde
2005	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...
	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...
	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...
	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...
2006	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...
	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...
	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...
	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...
2007	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...
	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...
	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...
	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...
2008	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...	Blanc ...
	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...	Noir ...
	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...	Bleu ...
	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...	Rouge ...

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

Références

- « Data Warehouse Design: Modern Principles and Methodologies » de Matteo Golfarelli et Stefano Rizzi, 2009, Ed: Osborne/McGraw-Hill.

- « Olap Solutions: Building Multidimensional Information Systems » de E. Thomsen, 2002, Ed: John Wiley & Sons Inc.

Exercice

On considère un entrepôt de données permettant d'observer les ventes de produits d'une entreprise. Le schéma des tables est le suivant :

- *CLIENT (id-client, région, ville, pays, département)*
- *PRODUIT (id-prod, catégorie, coût-unitaire, fournisseur, prix-unitaire, nom-prod)*
- *TEMPS (id-tps, mois, nom-mois, trimestre, année)*
- *VENTE (id-prod, id-tps, id-client, date-expédition, prix-de-vente, frais-de-livraison)*

Questions

1. Indiquer quelles sont la (les) table(s) de fait et les tables de dimension de cet entrepôt.
2. Donner pour chaque dimension, sa (multi-) hiérarchie.
3. Donner la représentation du schéma en étoile de l'entrepôt selon la notation de Golfarelli.
4. On veut transformer ce schéma en schéma en flocon. Donner la nouvelle représentation de la table TEMPS (ajouter des paramètres / attributs, si nécessaire)