

Gestion de données Réparties (DW)

Data warehouse

- Motivations et architecture
- Conception de la BD support
- Alimentation du DW
- Exploitation OLAP
- Conclusion

Data warehouse

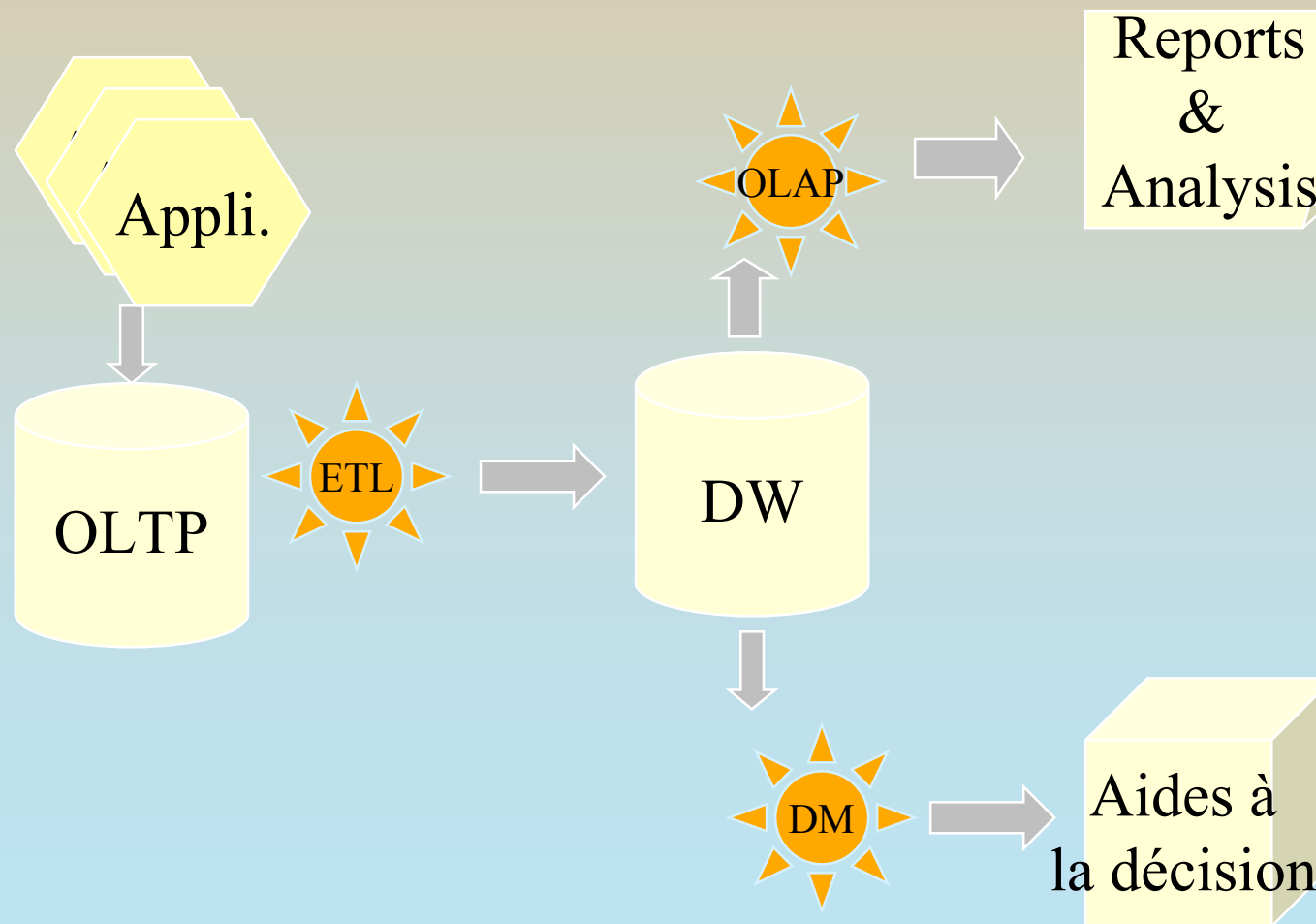
Définition, vocabulaire, composantes

Architectures

Structures multidimensionnelles

Opérations OLAP

OLTP et OLAP



Datawarehouse

- Entrepôt de données
 - Ensemble de données historisées variant dans le temps, organisé par sujets, consolidé dans une base de données unique, géré dans un environnement de stockage particulier, aidant à la prise de décision dans l'entreprise.
- Trois fonctions essentielles :
 - collecte de données de bases existantes et chargement
 - gestion des données dans l'entrepôt
 - analyse de données pour la prise de décision

Datamart

« Le marché de données est une implantation localisée d'un entrepôt de données à usage unique » (*traduction libre Devlin 1997*)

« L'entrepôt de données est prévu pour l'entreprise dans son ensemble alors que le marché de données est sectoriel (il peut être un sous-ensemble exact ou modifié de l'entrepôt de données) » (*Bédard et al, 1997*)

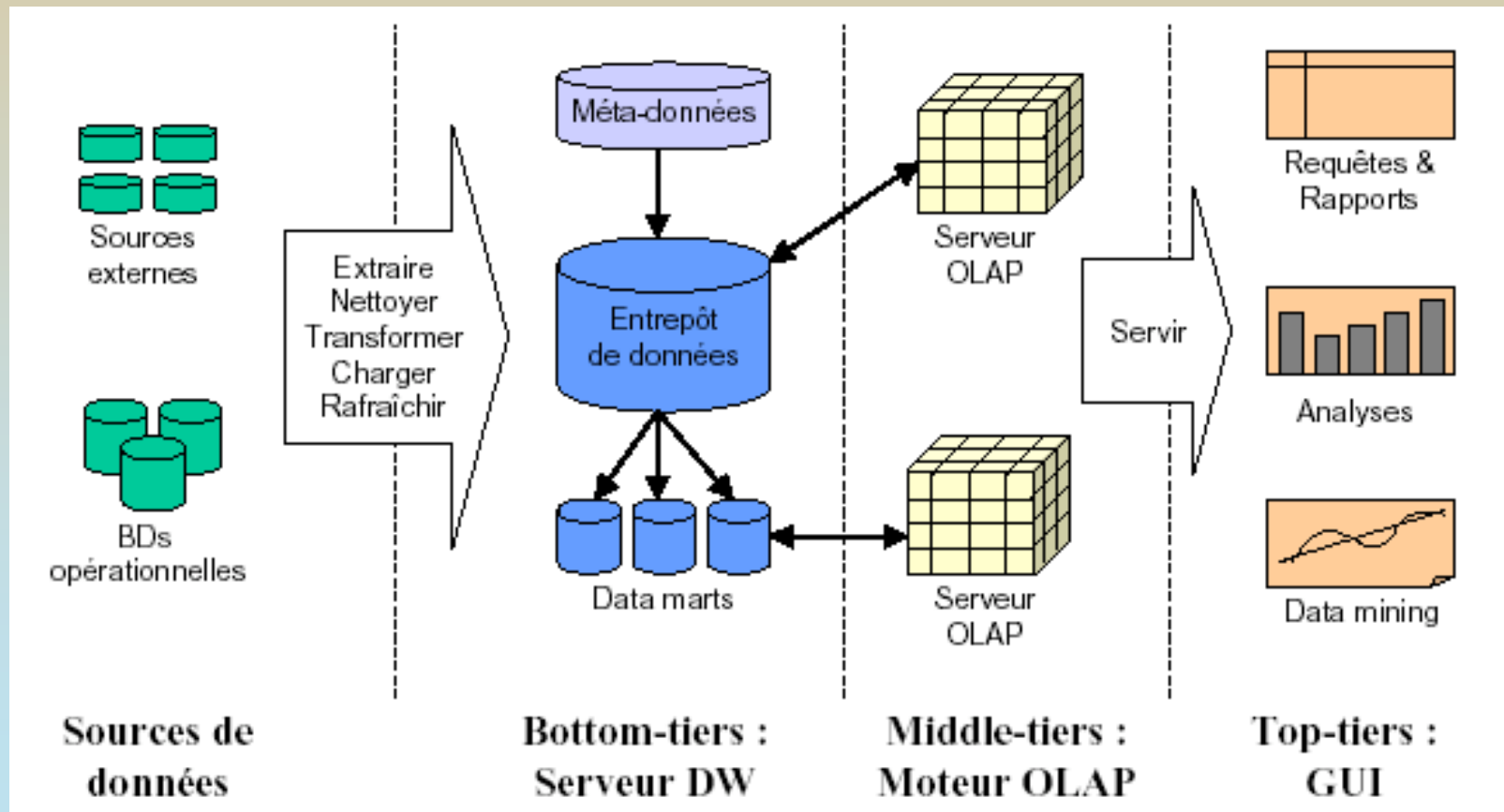
Définitions

Systèmes transactionnels (ST)	Entrepôts de données	Marchés de données
Construit pour les transactions (OLTP)	Construit pour l'analyse	Construit pour l'analyse
Données détaillées	Données détaillées et résumées	Données détaillées et résumées
Intégré selon les applications	Intégré pour l'entreprise	Intégré par sujet ou département
Mis à jour continuellement	Jamais mis à jour, seulement ajout de nouvelles données	Jamais mis à jour, seulement ajout de nouvelles données
Données actuelles	Données actuelles et d'archive	Données actuelles et d'archive
Source originale des données	Données importées des ST	Données importées des ST et/ou d'entrepôts
Structure normalisée	Structure dénormalisée*	Structure dénormalisée*

Architecture type

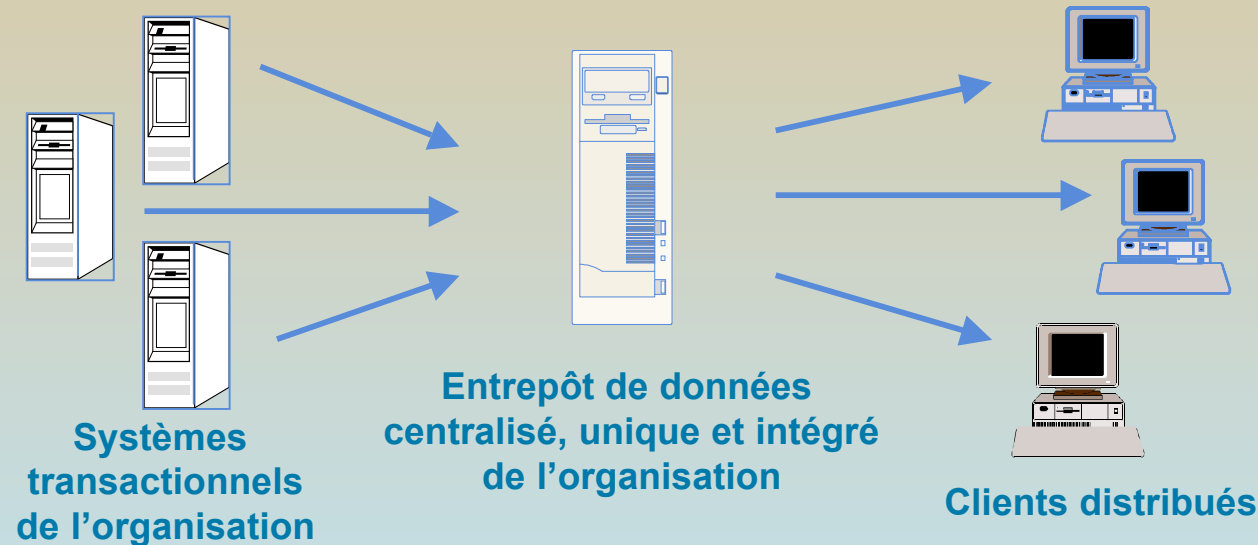
- Une architecture d'entrepôt de données possède les caractéristiques suivantes :
 - les données sources sont extraites de systèmes, de bases de données et de fichiers
 - les données sources sont nettoyées, transformées et intégrées avant d'être stockées dans l'entrepôt
 - l'entrepôt est en lecture seulement et est défini spécifiquement pour la prise de décision organisationnelle
 - les usagers accèdent à l'entrepôt à partir d'interfaces et d'applications (clients)

Architecture type



Architecture centralisée

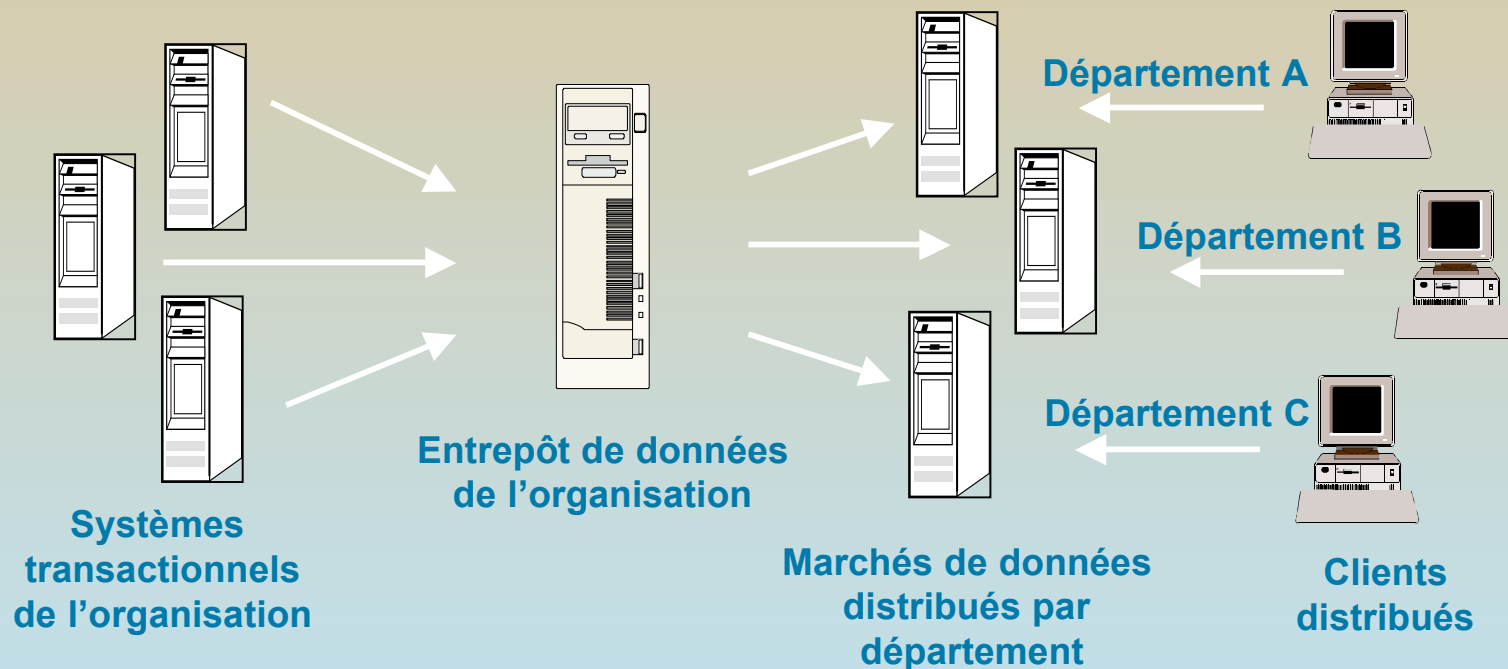
(*Corporated architecture*)



Il s'agit de la version *centralisée et intégrée* d'un entrepôt regroupant l'ensemble des données de l'entreprise. Les différentes bases de données sources sont intégrées et sont distribuées à partir de la même plate-forme physique

Architecture fédérée

(Federated architecture)



Il s'agit de la version intégrée d'un entrepôt où les données sont introduites dans les marchés de données orientés selon les différentes fonctions de l'entreprise

Concevoir le DW

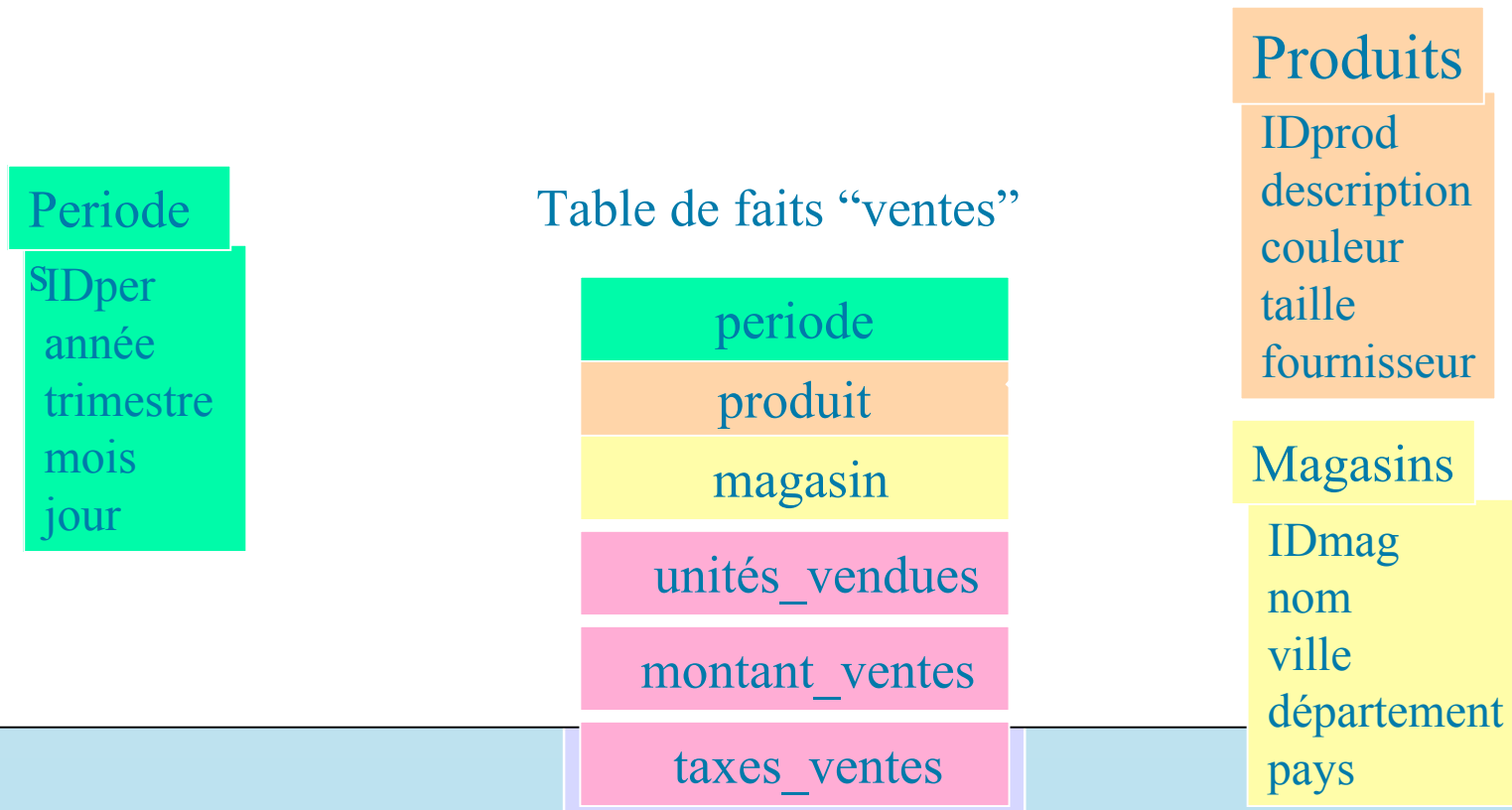
- Export de données des sources
 - Hétérogènes et variées
 - Fichiers, BD patrimoniales, Web, ...
 - Définition des vues exportées
- Définition d'un schéma global
 - Intègre les données utiles
 - S'appuie sur le modèle relationnel
- Nécessité d'une gestion de méta-données
 - Description des sources
 - Description des vues exportées
 - Description du schéma global

Organisation par sujet

- Les données sont organisées par sujets majeurs:
 - Clients, produits, ventes, ...
- **Sujet = faits + dimensions**
 - Collecte les données utiles sur un sujet
 - Exemple: ventes
 - Synthétise une vue simple des événements à analyser
 - Exemple: Ventes (N°, produit, période, magasin,)
 - Détaille la vue selon les dimensions
 - Exemple: Produits(IDprod, description, couleur, taille, ...)
 - Magasins(IDmag, nom, ville, dept, pays)
 - Periodes(IDper, année, trimestre, mois, jour)

Schémas en étoile

- Une table de faits encadrées par N tables de dimensions



Vocabulaire

Les outils traditionnels de gestion et d'exploitation des données sont du type transactionnel ou OLTP (*On-Line Transaction Processing*)

Les nouveaux outils d'exploitation des données sont de type analytique :

- Entrepôts de données (Data Warehouses)

- Marchés de données (Data Marts)

- Requêteurs et rapporteurs (Querying and Reporting Tools)

- OLAP (On-Line Analytical Processing)

- Fouille de données automatique (Data Mining)

Schémas en flocons

- Raffinement du schéma étoile avec des tables normalisées par dimensions

Ventes

Produits

IDprod
description
couleur
taille
IDfour

Fournisseurs

IDfour
description
type
Adresse

- Avantages
 - Évite les redondances
 - Conduit aux constellations (plusieurs tables de faits à dimensions partagées)

Conception du schéma intégré

- Isoler les faits à étudier
 - Schéma des tables de faits
- Définir les dimensions
 - Axes d'analyse
- Normaliser les dimensions
 - Éclater en plusieurs tables liés par contraintes référentielles
- Intégrer l'ensemble
 - Plusieurs tables de faits partagent quelques tables de dimension (constellation d'étoiles)

Bilan conception

- Le datawarehouse regroupe, historise, résume les données de l'entreprise
- Le concepteur définit schéma exportés et intégrés
 - des choix fondamentaux !
 - Ciblage essentiel !
- Le datamart c'est plus ciblé et plus petit.
- Questions ?
 - Peut-on ajouter des données au niveau de l'entrepôt ?



Alimenter le DW

- ETL = Extracteur+Intégrateur
 - Extract + Transform + Load
- Extraction
 - Depuis les bases sources ou les 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

- Accès unifiés aux données
 - Unification des modèles
 - Traduction de fichiers, BD réseaux, annuaires en tables
 - Evolution vers XML (modèle d'échange) plus riche
 - Unification des accès
 - Rowset, SQL limité, SQL complet, ...
- Mapping plus ou moins sophistiqué
 - Unification des noms
 - Appeler pareil les mêmes choses et différemment les choses différentes
 - Application des "business rules"
 - Elimination des doubles
 - Jointure, projection, agrégation (SUM, AVG)
- Cleaning des données

Data Cleaning

- Valeurs manquantes (nulles)
 - Ignorer le tuple
 - Remplacer par une valeur fixe ou par la moyenne
- Valeurs erronées ou inconsistantes
 - Générées en présence de bruits
 - Détecter par une analyse de voisinage
 - Écart par rapport à la moyenne
 - Factorisation en groupes (outliers)
 - Remplacer par une valeur fixe ou par la moyenne
- Inspection manuelle de certaines données possible

Chargement

- Pas de mise à jour
 - Insertion de nouvelles données
 - Archivage de données anciennes
- De gros volumes
 - Périodicité parfois longue
 - Chargement en blocs (bulk load)
 - Mise à jour des index et résumés
- Problèmes
 - Cohabitation avec l'OLAP ?
 - Procédures de reprises ?

Gérer l'entrepôt

- Base relationnelle
 - Support de larges volumes (qq 100 gigas à qq téras)
 - Historisation des données (fenêtres)
 - Importance des agrégats et chargements en blocs
- Base spécialisée
 - Base multidimensionnelle
 - Combinaison des deux
- Machine support parallèle
 - Multiprocesseurs
 - Mémoire partagée, cluster, bus partagé, etc.

Le multidimensionnel

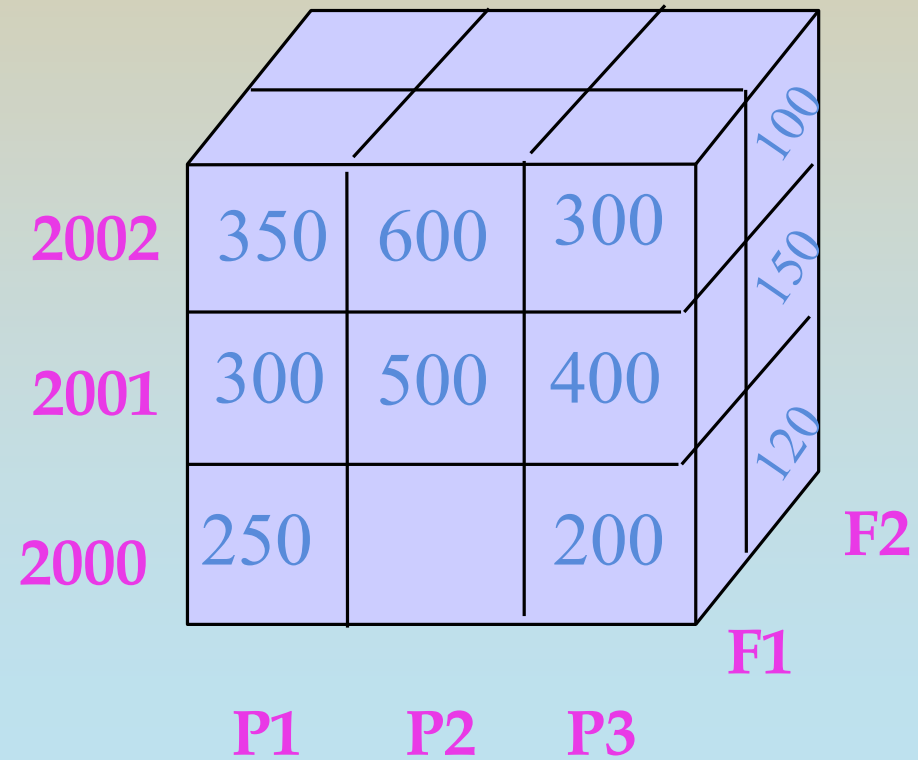
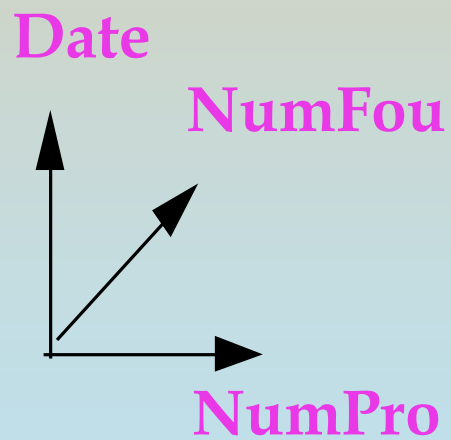
- Dimensions:

- Temps
- Géographie
- Produits
- Clients
- Canaux de ventes.....

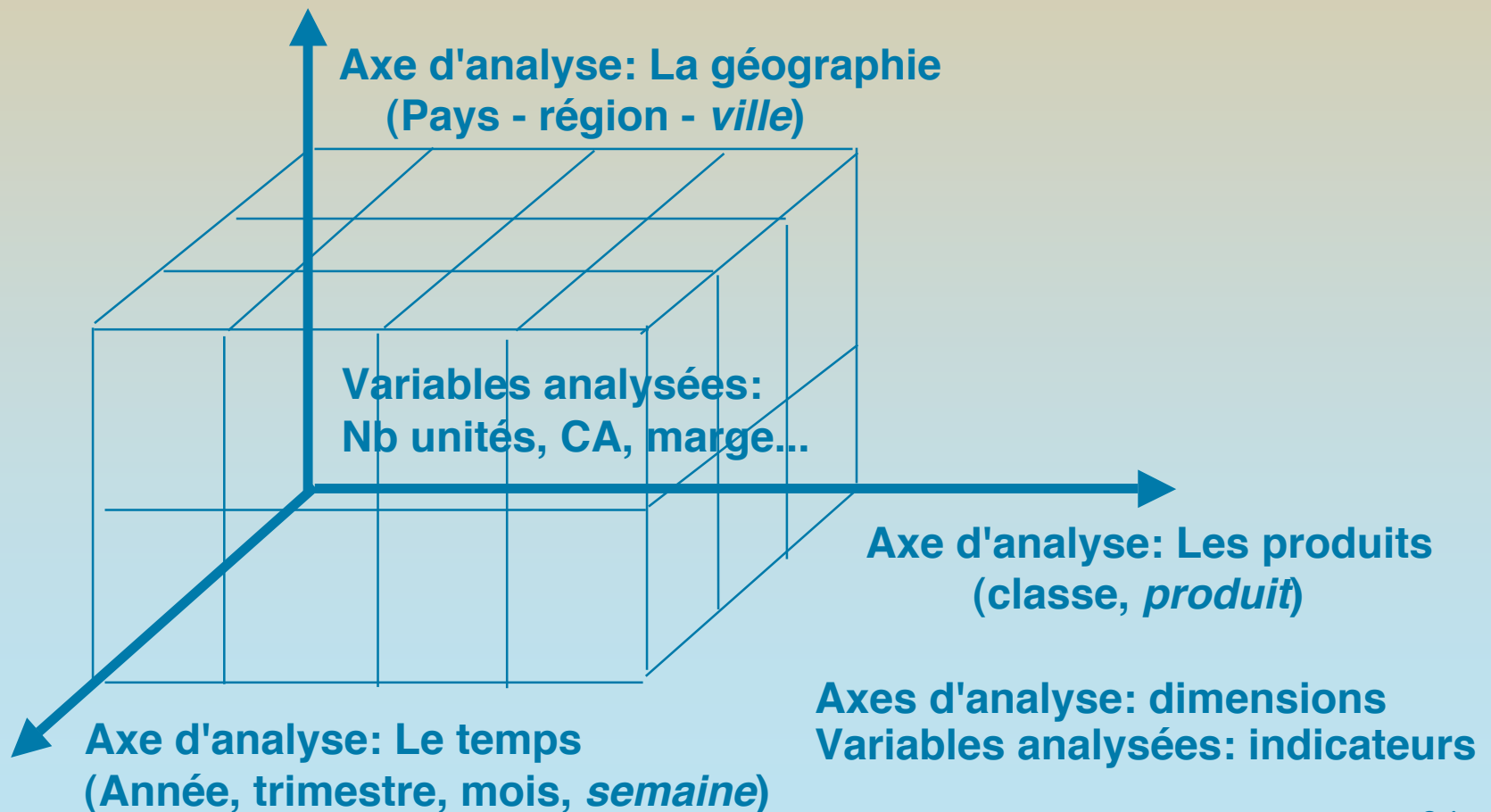
- Indicateurs:

- Nombre d'unités vendues
- CA
- Coût
- Marge.....

Cube de données



Le data cube et les dimensions

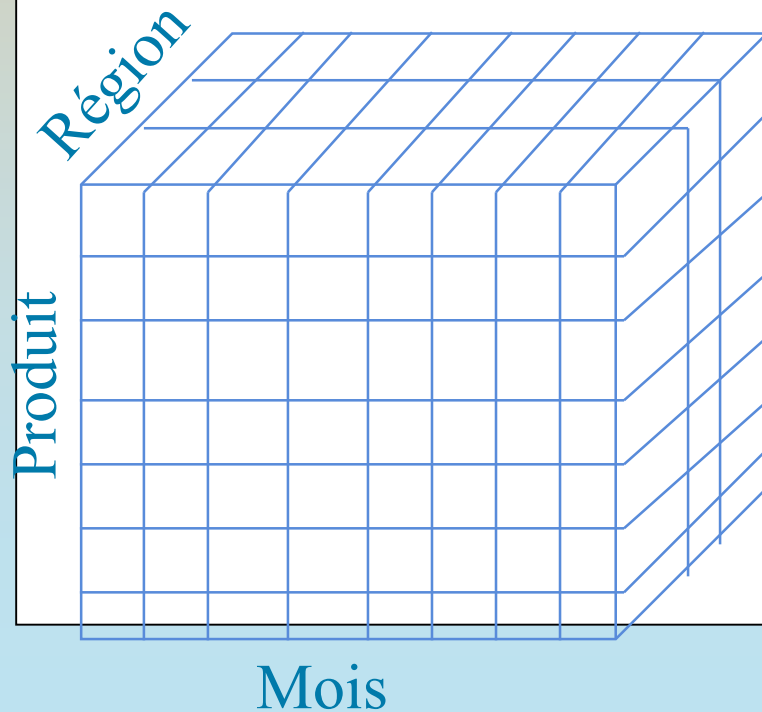


La granularité des dimensions

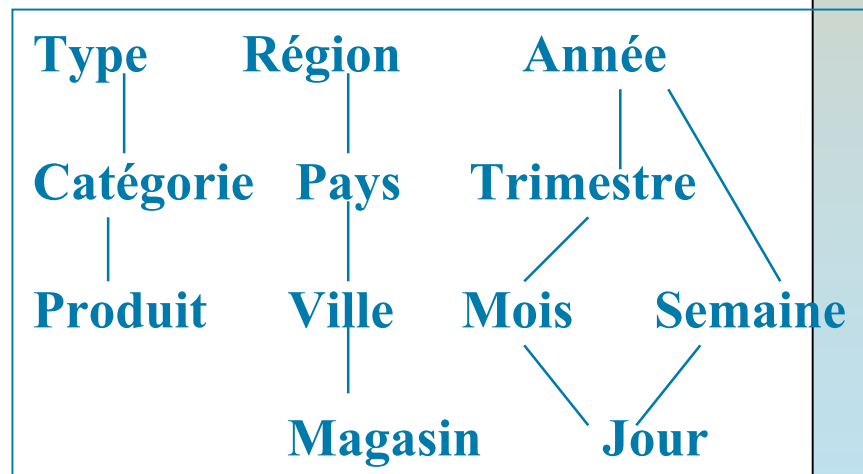


Exemple

- Montant des ventes fonction de (Mois, région, Produit)

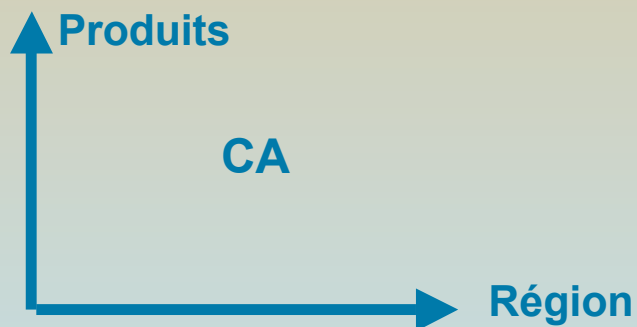


Granularité des dimensions :

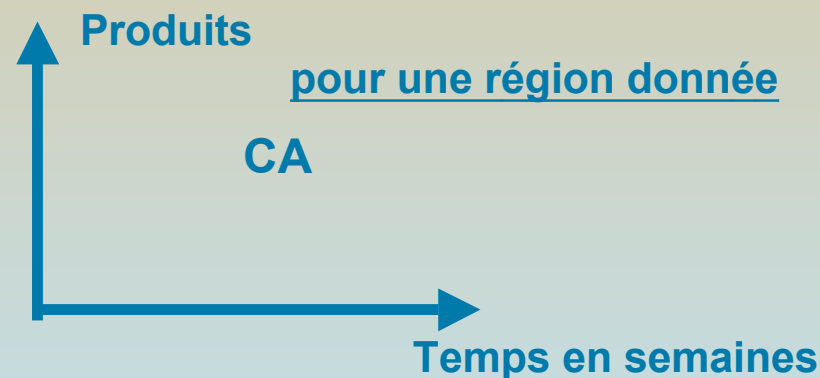


La navigation multidimensionnelle

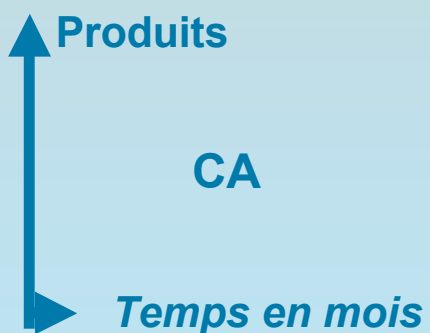
Projection en 2 dimensions



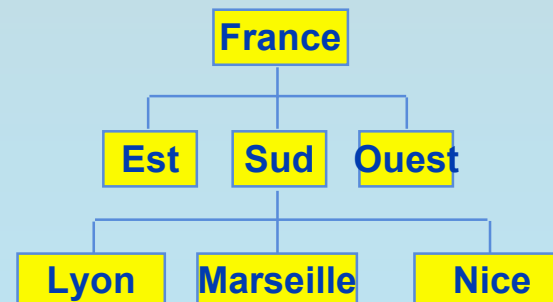
Coupe d'un cube



Réduction selon 1 dimension



Zoom selon une dimension

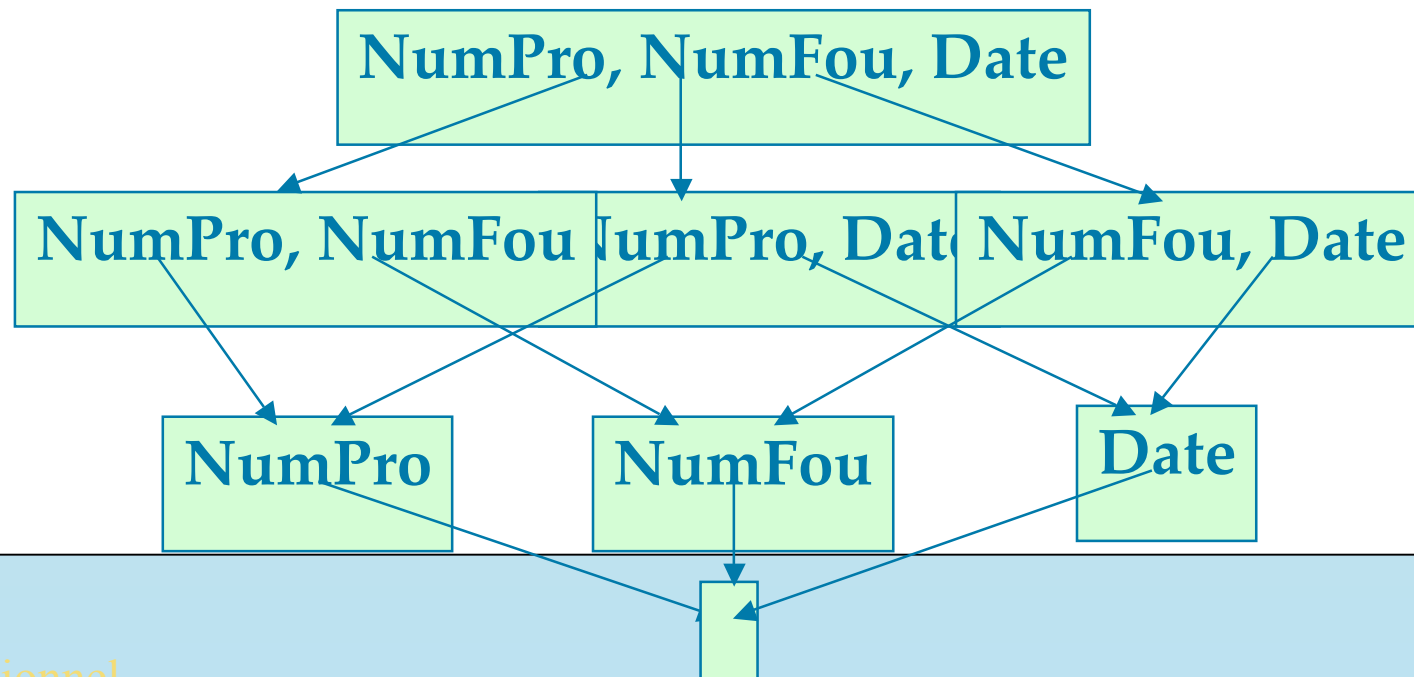


L'algèbre des cubes

- Roll up :
 - Agréger selon une dimension
 - Semaine → Mois
- Drill down :
 - Détailler selon une dimension
 - Mois → Semaine
- Slice et Dice:
 - Sélection et projection selon 1 axe
 - Mois = 04-2003 ; Projeter(Région, Produit)
- Pivot :
 - Tourne le cube pour visualiser une face
 - (Région,Produit)→(Région, Mois)

Les vues d'un cube

- Partant d'un cube 3D, il est possible d'agréger selon une dimension tournante
- On obtient un treillis de vues (calculable en SQL)



Extension de SQL

- **ROLLUP:**
 - SELECT <column list>
 - FROM <table...>
 - GROUP BY
ROLLUP(column_list);
- Crée des agrégats à $n+1$ niveaux, n étant le nombre de colonne de groupage
 - $n, n-1, n-2, \dots, 0$ colonnes

- **CUBE:**
 - SELECT <column list>
 - FROM <table...>
 - GROUP BY
CUBE(column_list);
- Crée 2^n combinaisons d'agrégats, n étant le nombre de colonne de groupage

Exemple CUBE

Animal	Lieu	Quantite
Chien	Paris	12
Chat	Paris	18
Tortue	Rome	4
Chien	Rome	14
Chat	Naples	9
Chien	Naples	5
Tortue	Naples	1

- `SELECT Animal, Lieu, SUM(Quantite) as Quantite FROM Animaux GROUP BY Animal, Lieu WITH CUBE`

Animal	Lieu	Quantite
Chat	Paris	18
Chat	Naples	9
Chat	-	27
Chien	Paris	12
Chien	Naples	5
Chien	Rome	14
Chien	-	31
Tortue	Naples	1
Tortue	Rome	4
Tortue	-	5
-	-	63
-	Paris	30
-	Naples	15
-	Rome	18

Exemple ROLLUP

Animal	Lieu	Quantite
Chien	Paris	12
Chat	Paris	18
Tortue	Rome	4
Chien	Rome	14
Chat	Naples	9
Chien	Naples	5
Tortue	Naples	1

- `SELECT Animal, Lieu, SUM(Quantite) as Quantite FROM Animaux GROUP BY Animal,Lieu WITH ROLLUP`

Animal	Lieu	Quantite
Chat	Paris	18
Chat	Naples	9
Chat	-	27
Chien	Paris	12
Chien	Naples	5
Chien	Rome	14
Chien	-	31
Tortue	Naples	1
Tortue	Rome	4
Tortue	-	5
-	-	63

Quelques outils OLAP

- Oracle
 - OLAP API = Datacube
 - Express = Analyse
 - Report = Reporting
- Business Object
 - BusinessQuery = Requêtage
 - BusinessObject = Requêtage + Analyse + Reporting
 - WebIntelligence = Datacube
- Cognos
 - Impromptu = Reporting
 - Powerplay = Datacube
 - Query = Requêtage
- Hyperion
 - ESS Base = Base MOLAP
 - ESS Analysis = Analyse + Datacube

Les Data ...

- Datawarehouse
 - entrepôt des données historisées de l'entreprise
- Datamart
 - magasin de données ciblé sur un sujet précis
- Datamining
 - exploration des données afin de découvrir des connaissances
- Datacube
 - cube de présentation d'unités selon 3 dimensions
- Datawebhouse
 - entrepôt des données collectées sur le web