Systèmes d'information décisionnels (Data Warehouse)

Dr Dendani-Hadiby Nadjette
Université Badji Mokhtar Annaba
Département d'Informatique
n_dendani@yahoo.fr

2021/2022

L’ANALYSE MULTIDIMENSIONNELLE

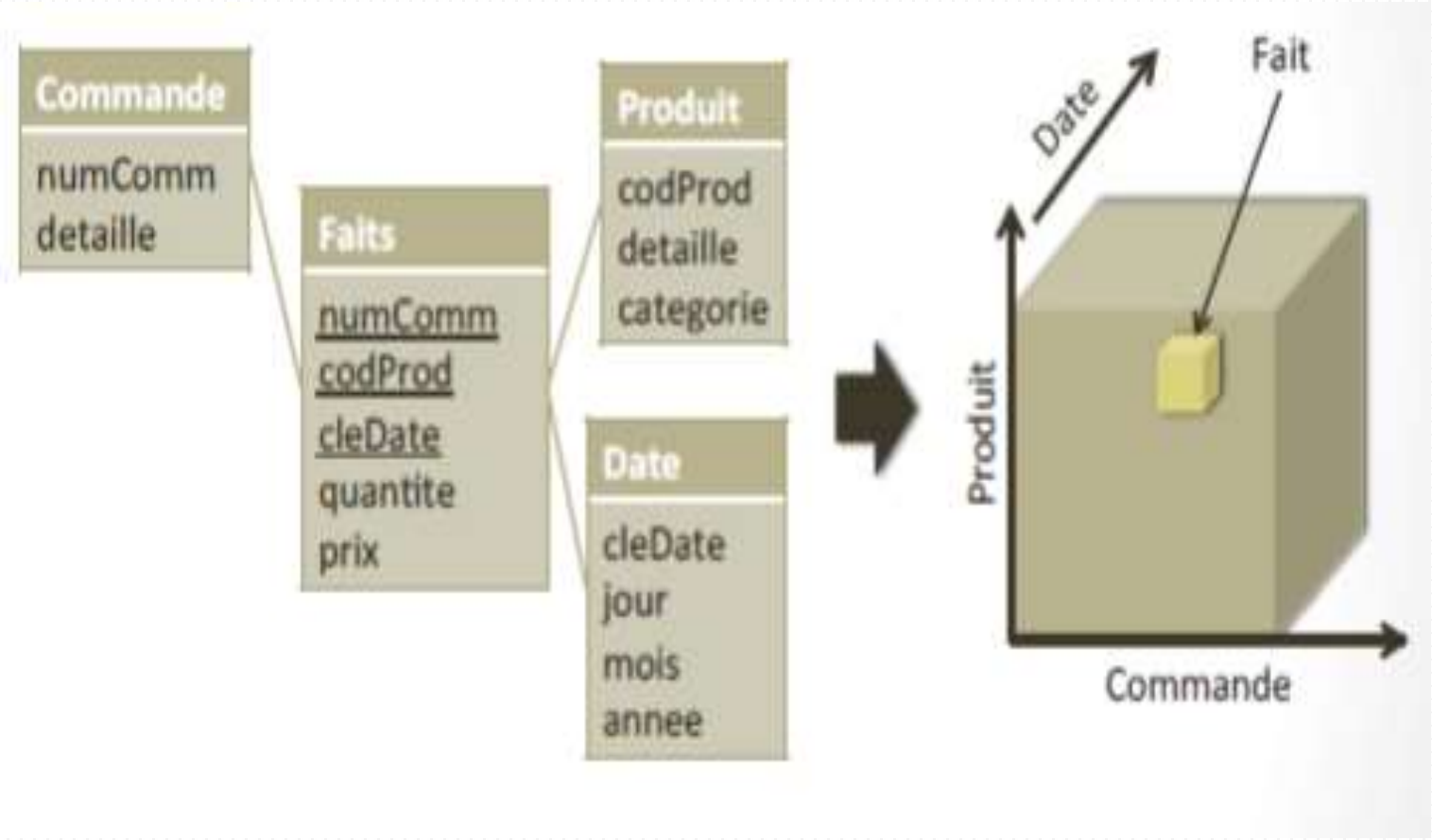
L'analyse multidimensionnelle 1

- Objectif : obtenir des informations déjà agrégées selon les besoins de l'utilisateur: simplicité et rapidité d'accès
- HyperCube OLAP : représentation de l'information dans un hypercube à N dimensions
- OLAP (OnLine Analytical Processing) : fonctionnalités qui servent à faciliter l'analyse multidimensionnelle : opérations réalisables sur l'hypercube

L’analyse multidimensionnelle 2

- 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.

(Hyper)Cube de données 1

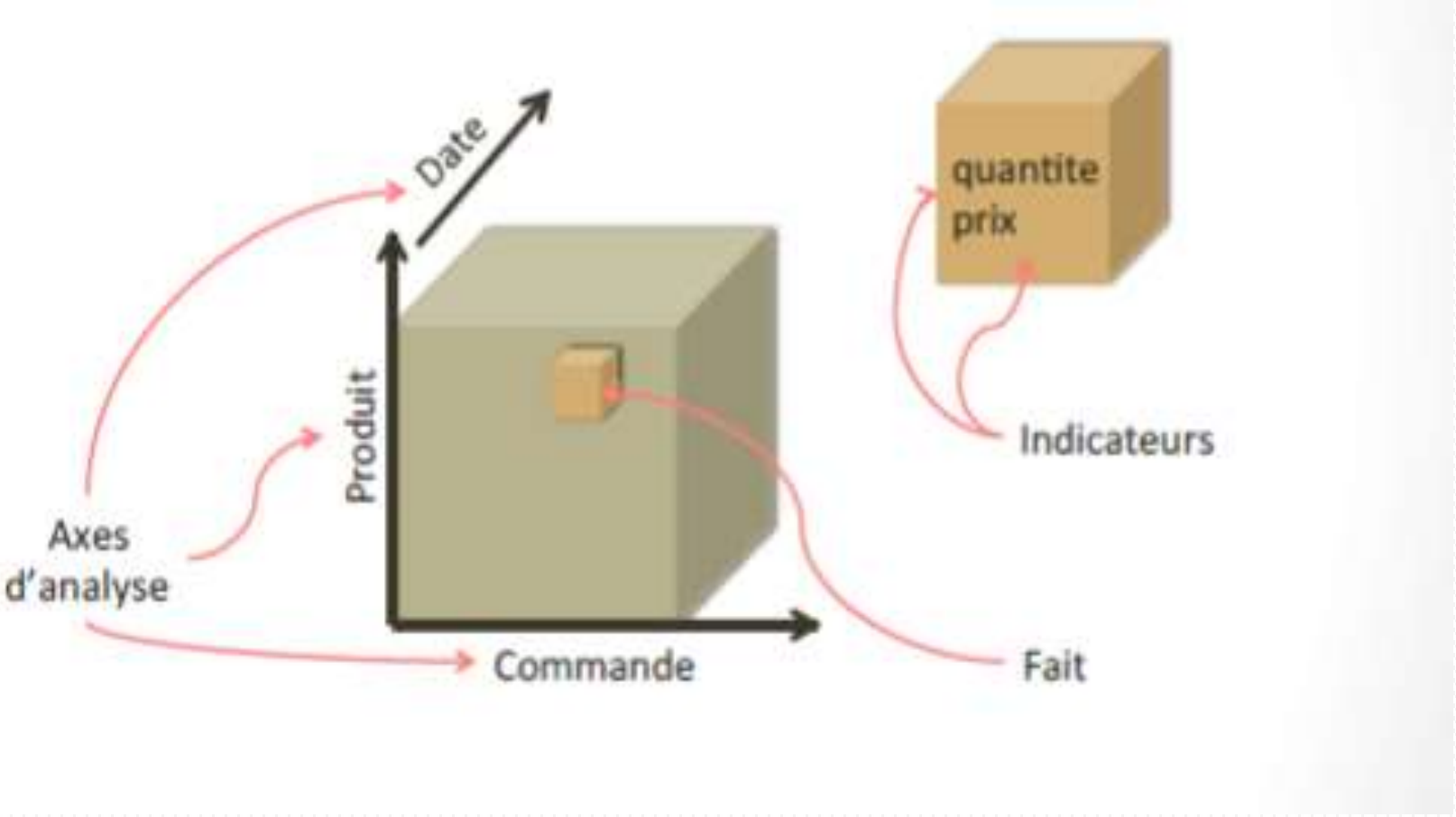


(Hyper)Cube de données 2

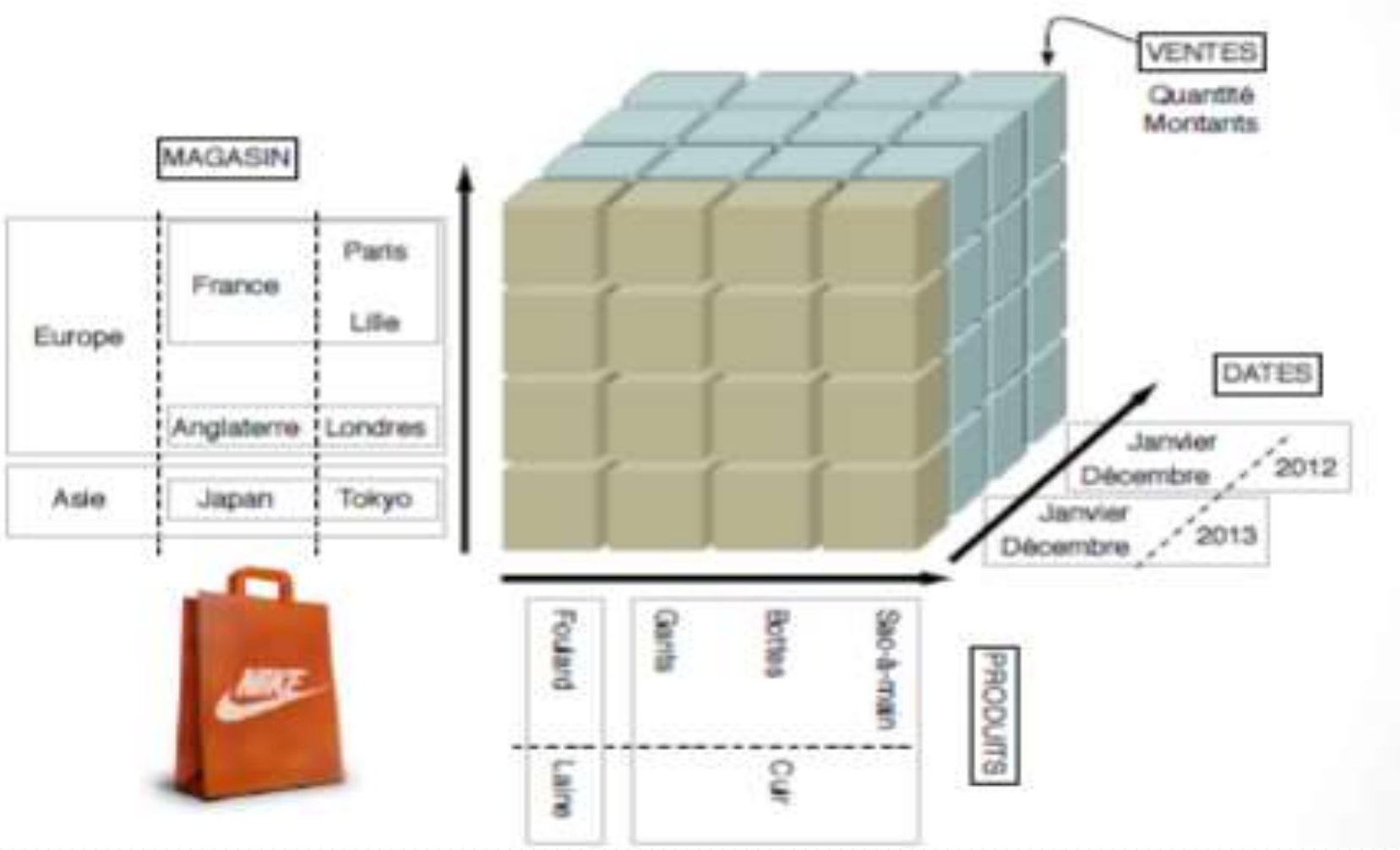
Composantes d'un cube

- Chaque cellule du cube correspond à une occurrence du fait
- Chaque cellule contient des indicateurs (variables, métriques ou mesures)
- Les axes d'analyse, également appelés dimensions, contiennent un ensemble de valeurs
- Des hiérarchies sont spécifiées sur les dimensions afin de Permettre une consolidation des indicateurs
- Chaque indicateur a une foncıon d'agrégat afin d'être exploité sur la hiérarchie

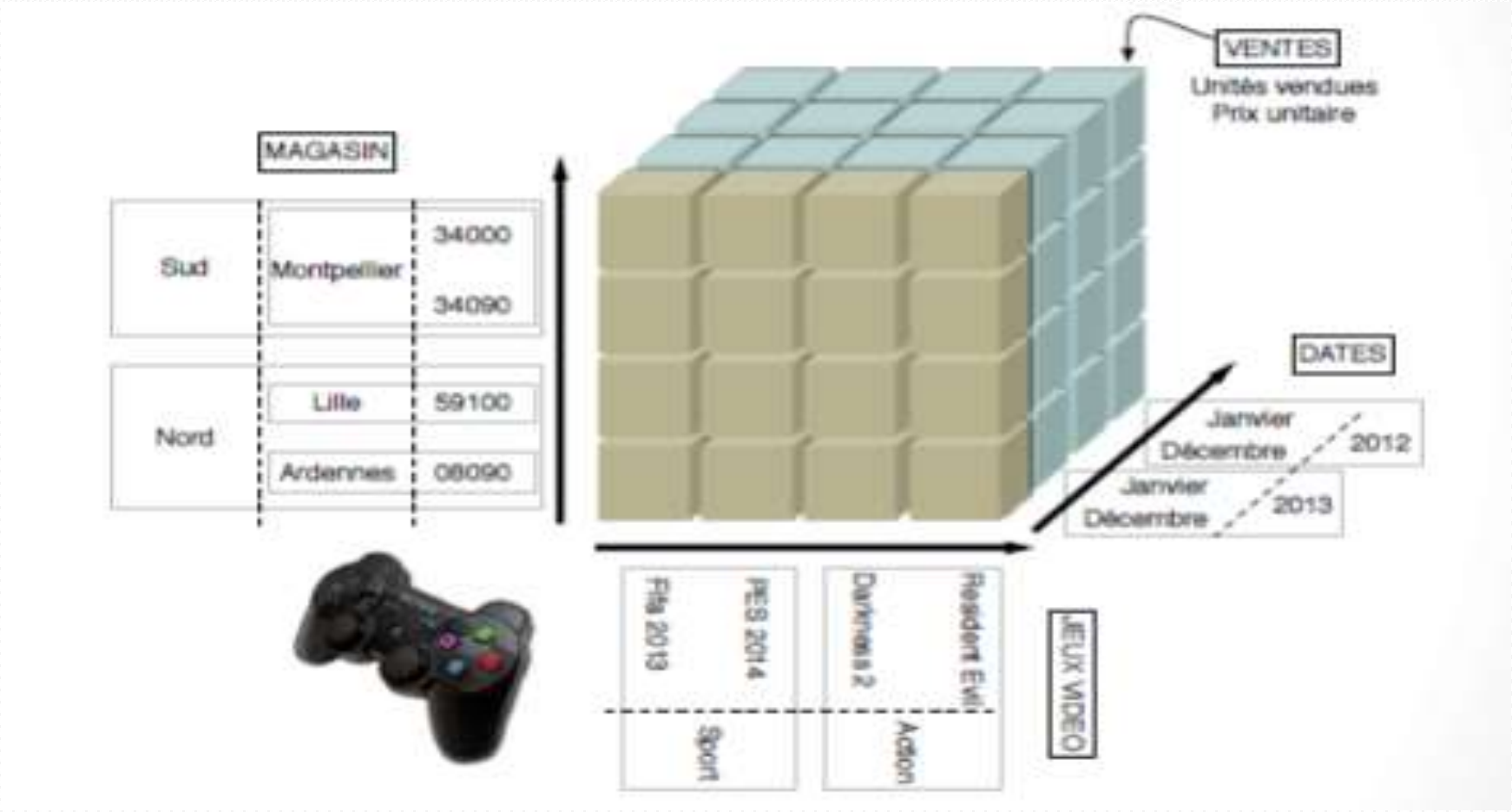
(Hyper)Cube de données 3



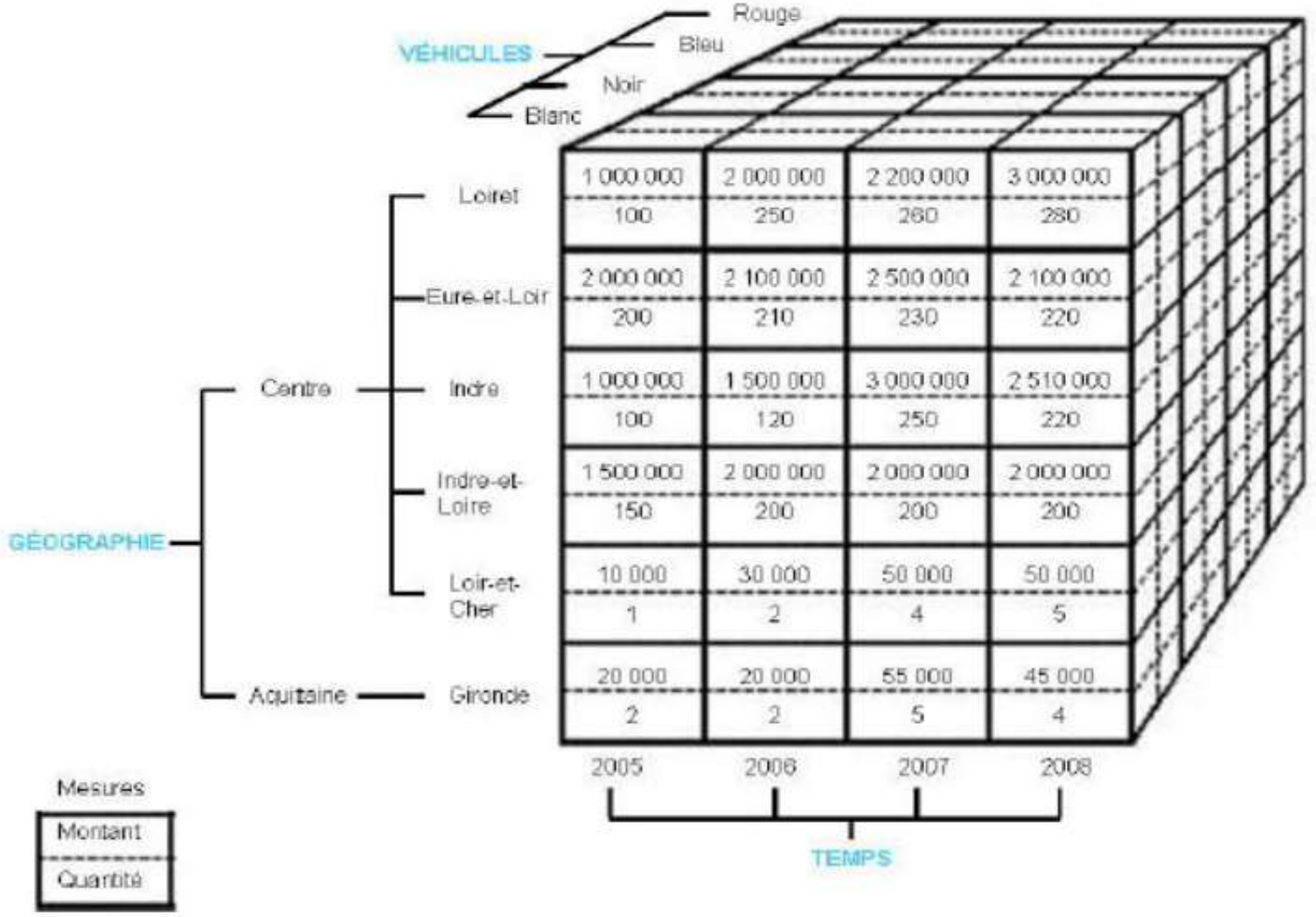
Exemple1


VENTES
Quantité
Montants
MAGASIN
Europe
France
Paris
Lille
Angleterre
Londres
Asie
Japan
Tokyo
DATES
Janvier
Décembre
2012
Janvier
Décembre
2013
Foulard
Gants
Bottes
Sac-à-main
Laine
Cuir
PRODUITS

Exemple 2


VENTES
Unités vendues
Prix unitaire
MAGASIN
Sud
Montpellier
34000
34090
Nord
Lille
59100
Ardennes
08090
DATES
Janvier
Décembre
2012
Janvier
Décembre
2013
JEUX VIDEO
Sport
Fifa 2013
PES 2014
Action
Darkness 2
Resident Evil

Exemple 3


VÉHICULES
Rouge
Bleu
Noir
Blanc
GÉOGRAPHIE
Centre
Loiret
Eure-et-Loir
Indre
Indre-et-Loire
Loir-et-Cher
Aquitaine
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

L'implémentation du OLAP

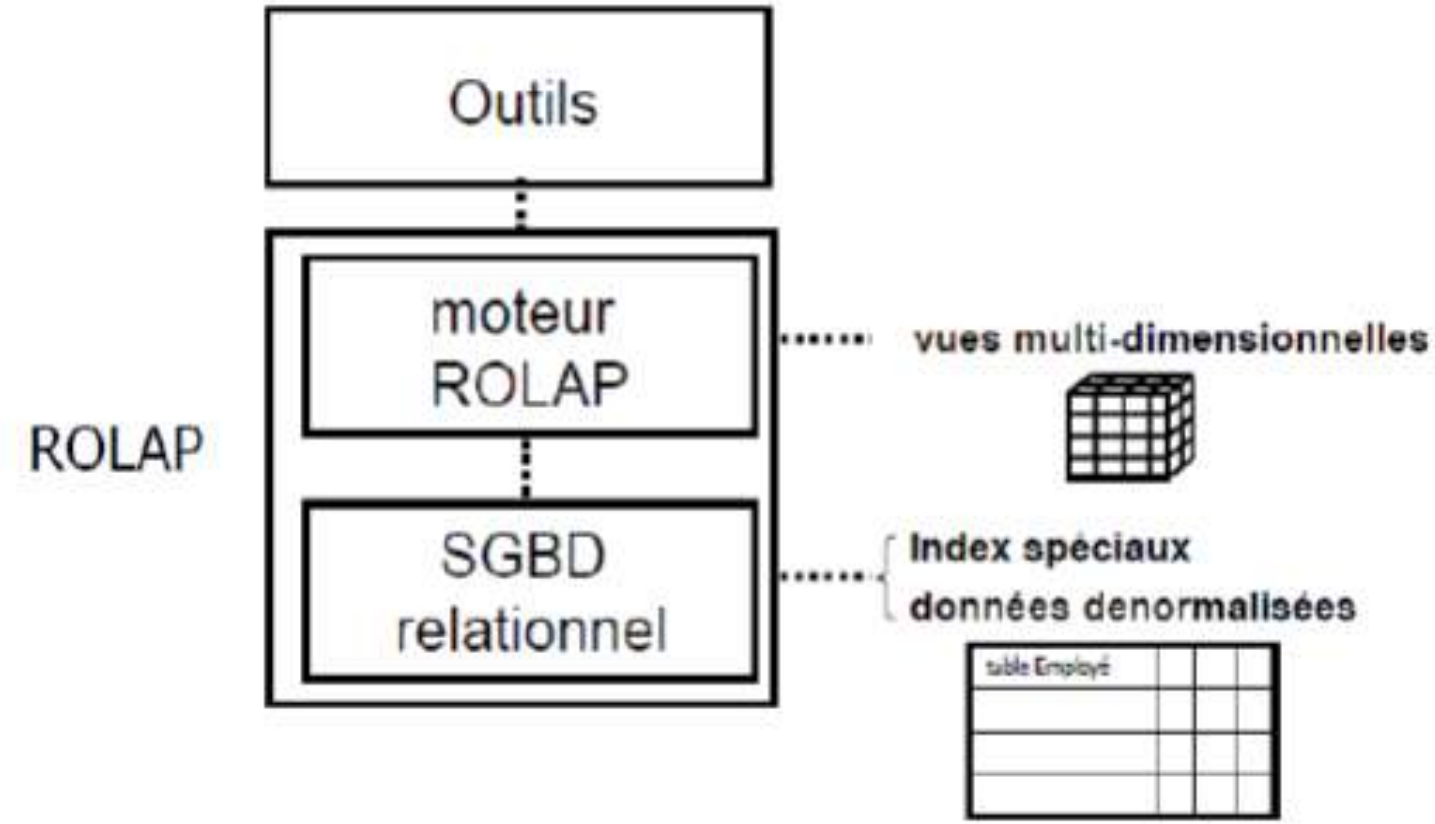
- Description de la base multidimensionnelle suivant la technologie utilisée :
  - ROLAP (Relational-OLAP)
  - MOLAP (Multidimensional-OLAP)
  - HOLAP (Hybrid-OLAP)

ROLAP (1)

- Les données sont stockées dans une BD relationnelle
- Le cube est stocké selon le modèle en étoile (flocon ou constellation)
- 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

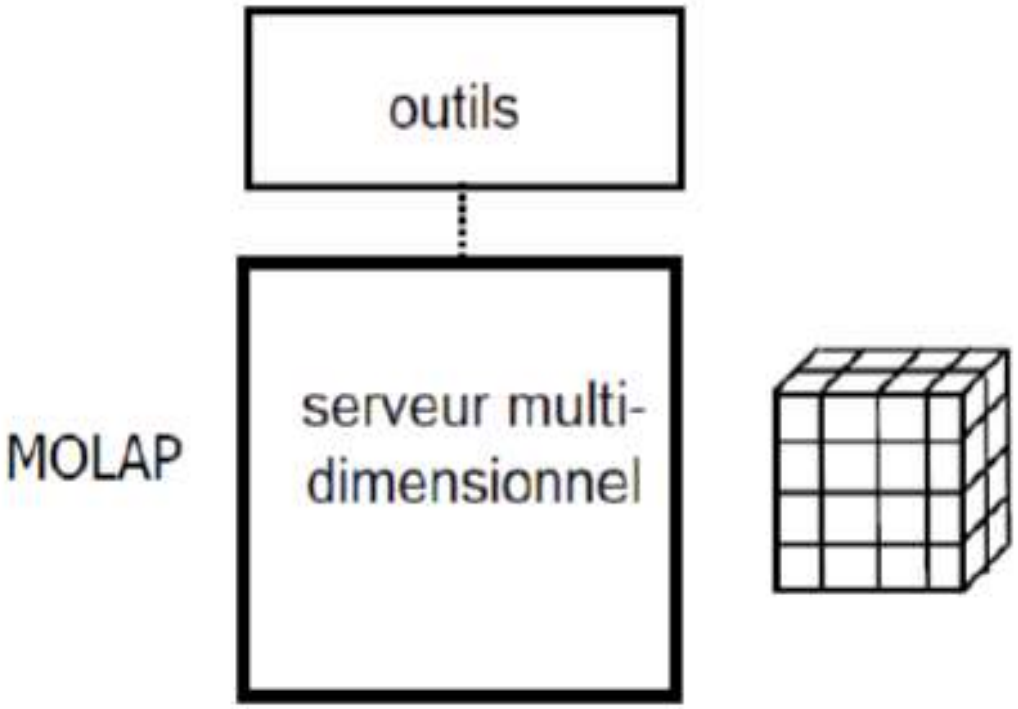
ROLAP (2)


Outils
moteur ROLAP
SGBD relationnel
ROLAP
vues multi-dimensionnelles
Index spéciaux
données denormalisées
table Employé

MOLAP (1)

- Les données sont stockées comme des matrices à plusieurs dimensions : Cube[1:m,1:n,1:p](mesure)
- On trouve en colonne tous les axes, puis tous les indicateurs
- Chaque cellule du cube est stockée par une ligne dans la matrice
- Accès direct aux données dans le cube
- Avantages :
  - Rapidité
- Inconvénients :
  - Difficile à mettre en place
  - Formats souvent propriétaires
  - Ne supporte pas de très gros volumes de données
- Exemple de moteurs MOLAP :
  - Microsoft Analysis Services, Hyperion

MOLAP (2)

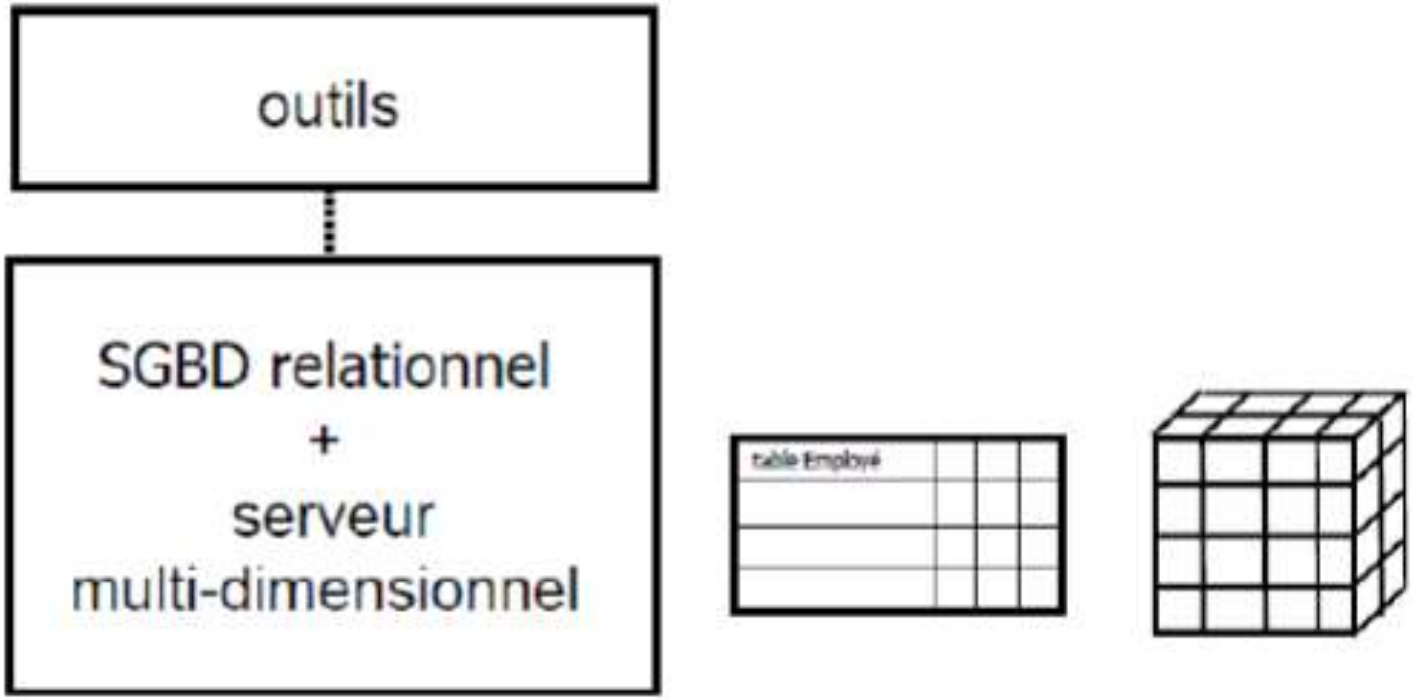


HOLAP (1)

- 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



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).

A savoir (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**

A savoir (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

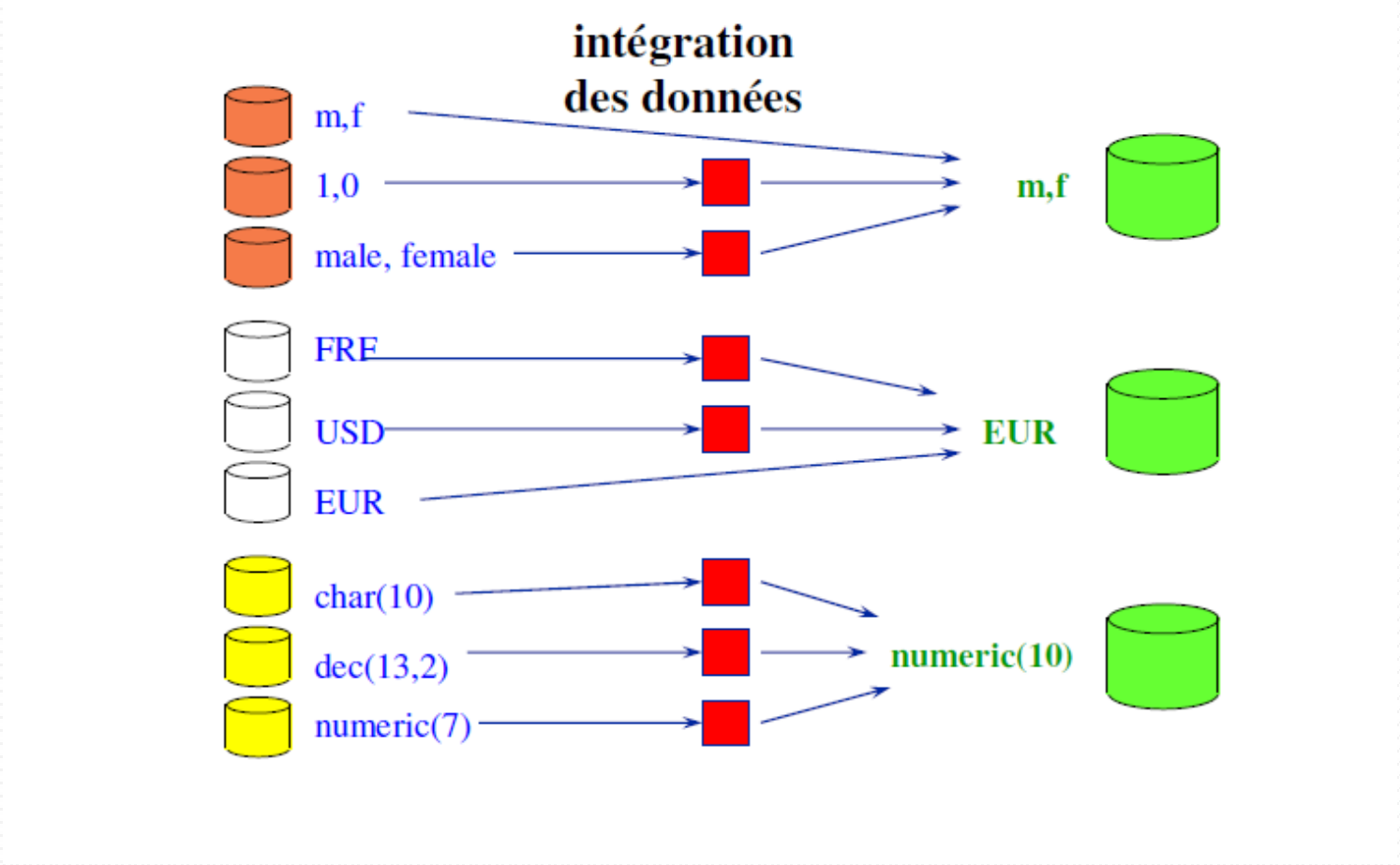
ETL

- 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

ETL

- **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…

ETL


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

ETL

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 datamining

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?)

6- Exploitation d'un entrepôt de données

1. Stratégies d'implantation d'un ED

1. Exploitation d'un ED

1. Visualisation autour d'un ED

Principales applications autour d'un ED

- Réalisation de rapports divers (Reporting)
- Réalisation de tableaux de bords (Dashboards)
- Analyse en ligne diverses (OLAP)
- Fouille de données (Data Mining)
- Visualisations autour d'un ED (Visualizations)
- Etc.

Rapports (Reporting)

- Ils sont crées pour les utilisateurs qui ont besoin d'un accès régulier à des informations d'une manière presque statique
  - Ex: les hôpitaux doivent envoyer des rapports mensuels

    à des agences nationales
- Un rapport est défini par une requête (plusieurs requêtes) et une mise en page (diagrammes, histogrammes, etc)
- Les rapports peuvent être exécutés automatiquement ou manuellement

Tableaux de bords (Dashboards)

- Affichent une quantité limitée d'informations dans un format graphique facile à lire
- Fréquemment utilisé par les cadres supérieurs qui ont besoin d'un rapide aperçu des changements les plus importants
  - Ex : un aperçu en temps réel d'évolutions
- Pas vraiment utile pour une analyse complexe et détaillée

Analyse OLAP (On-Line Analytical processing)

- Techniques OLAP apparues en recherche dans les années 70 mais ont été développées dans les années 90 dans l'industrie

- Permettent de réaliser des synthèses, des analyses et de la consolidation dynamique de données multidimensionnelles

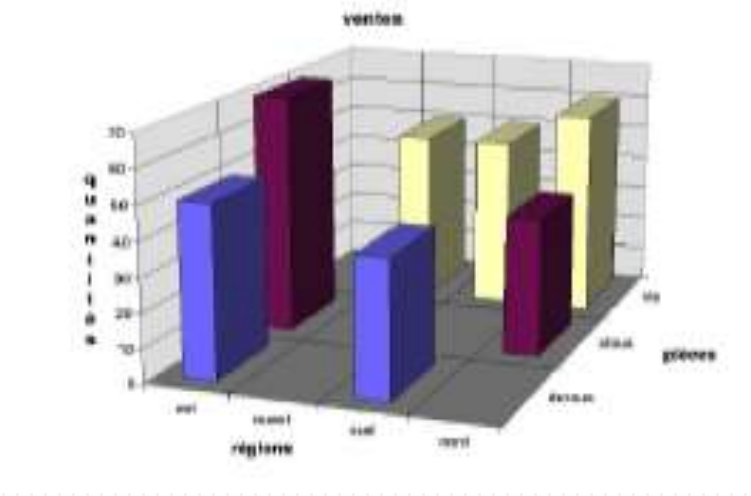
- Constitue la façon la plus naturelle d'exploiter un ED du fait de son organisation multidimensionnelle

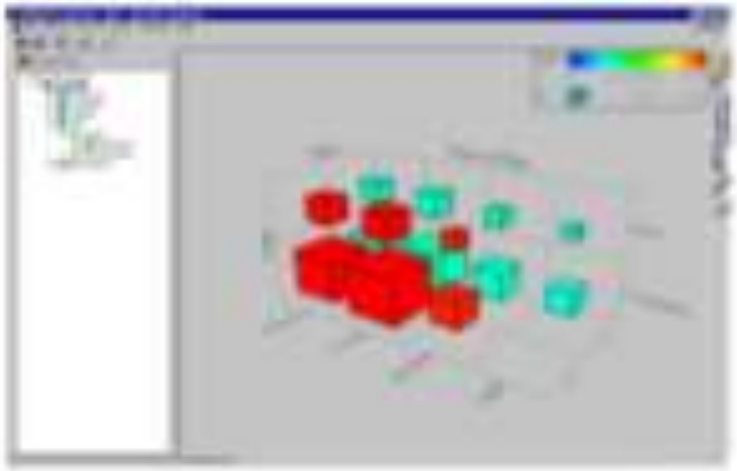
Fouille de données (Data Mining)

- Recherche de connaissance, sous forme de modèle de
- comportement, cachés dans les données
- Domaine jeune à l'intersection de l'Intelligence Artificielle, les Statistiques et les BD
- Nombreuses techniques de fouille :
  - Régression linéaire, induction d'arbres de décision, algorithmes génériques, réseaux de neurones, ...
- Les techniques de fouille sont en pleine évolution et sont de plus en plus intégrées dans les ED

Visualisation autour d'un ED

- Facilitent l'analyse et l'interprétation de données
- Convertissent des données complexes en images, graphiques en 2 et 3 dimensions, voire en animations
- Sont de plus en plus intégrées dans les ED





Domaines d'application des entrepôts et « succès stories »

1. Les domaines privilégiés :

- Domaine bancaire
- Domaine de la grande distribution
- Domaine des télécommunications
- Domaines de l'assurance et de la pharmacie
- Domaine de la santé, …

2. « Succès stories »:

- Casino, Walmart, Camaieu, …
- FranceTélécom, …

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

Fin