UNIVERSITÉ DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE

LABORATOIRE DES SYSTÈMES INFORMATIQUES

LES ENTREPOTS DE DONNEES
Data Warehouse

Dr. Kamel Boukhalfa

Boukhalk@gmail.com

Présentation de l'enseignant

- **Dr. Kamel Boukhalfa**
  - Docteur de l'USTHB et de l'ENSMA France
  - Professeur à l'USTHB
- **Emails**
  - Email de l'USTHB : kboukhalfa@usthb.dz
  - Emails souvent utilisés
    - boukhalk@gmail.com
    - Kamelboukhalfa@yahoo.fr
- **Site Web personnel**
  - http://boukhalfa.jimdo.com
  - CV
  - Production scientifique
  - Téléchargement de cours

Domaines de Recherches

- **Entrepôts de données**
  - Optimisation de la couche physique, Fragmentation Horizontale, Indexation, Tuning
- **Méta-heuristiques**
  - AG, RS, HC, RT, CF, …etc.
  - Optimisation multi-objectifs
- **Data mining**
  - Motifs fréquents, classification, règles d'association, data mining spatial
- **Big Data**
- **Cloud Computing**

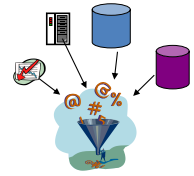
Plan

Le contexte

- **Besoin:** prise de décisions stratégiques et tactiques
- **Pourquoi:** besoin de réactivité
- **Qui:** les décideurs (non informaticiens)
- **Comment:** répondre aux demandes d'analyse des données, dégager des informations qualitatives nouvelles

Les données utilisables par les décideurs

- **Données opérationnelles (de production)**
  - Bases de données (Oracle, SQL Server)
  - Fichiers, …
  - Paye, gestion des RH, gestion des commandes…
- **Caractéristiques de ces données:**
  - **Distribuées**: systèmes éparpillés
  - **Hétérogènes**: systèmes et structures de données différents
  - **Détaillées**: organisation des données selon les processus fonctionnels, données surabondantes pour l'analyse
  - **Peu/pas adaptées à l'analyse** : les requêtes lourdes peuvent bloquer le système transactionnel
  - **Volatiles**: pas d'historisation systématique



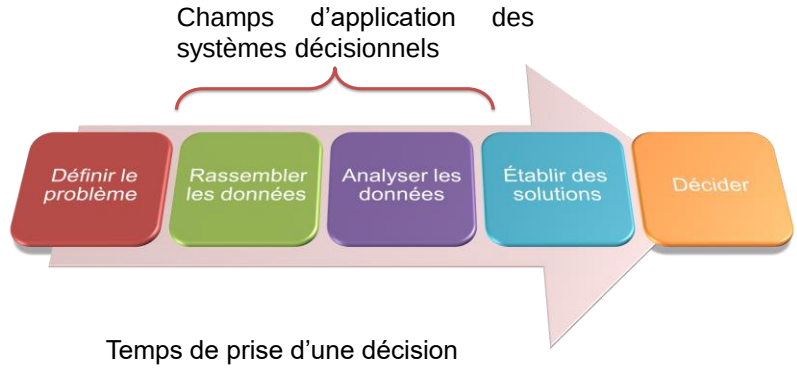
Problématique

- Comment répondre aux demandes des décideurs?
  - En donnant un accès rapide et simple à l'information stratégique
  - En donnant du sens aux données

➡ Mettre en place un système d'information dédié aux applications décisionnelles:

Un Data Warehouse

Le processus de prise de décision



Introduction

- Pourquoi le data warehouse ?
  - Améliorer les performances décisionnelles de l'entreprise
- Comment?
  - En répondant aux demandes d'analyse des décideurs
- Exemples
  - **Clientèle:** Qui sont mes clients? Pourquoi sont-ils mes clients?, Comment les conserver ou les faire revenir (préférence d'achat, habitudes, …)? Ces clients sont-ils vraiment intéressants pour moi?
  - **Marketing, actions commerciales**: où placer ce produit dans les rayons? Comment cibler plus précisément le mailing concernant ce produit?

Introduction

Raisons d'être d'un entrepôt de données

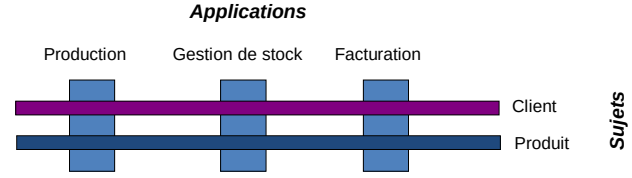
- Rassembler les données de l'entreprise dans un même lieu sans surcharger les BD (systèmes opérationnels)
- Permettre un accès universel à diverses sources de données et assurer la qualité des données
- Extraire, filtrer, et intégrer les informations pertinentes, à l'avance, pour des requêtes ultérieures
- Dégager des connaissances et faire un apprentissage sur l'entreprise, le marché et l'environnement

C'est quoi un entrepôt de données?

- Industrie (Inmon 1992)
  - Collection de données orientées sujets
  - Consolidées dans une base de données unique
  - Non volatiles et historisées variant dans le temps
  - organisées pour le support d'un processus d'aide à la décision
- Recherche (Stanford 1995)
  - Dispositif de stockage d'informations intégrées de sources distribuées, autonomes, hétérogènes

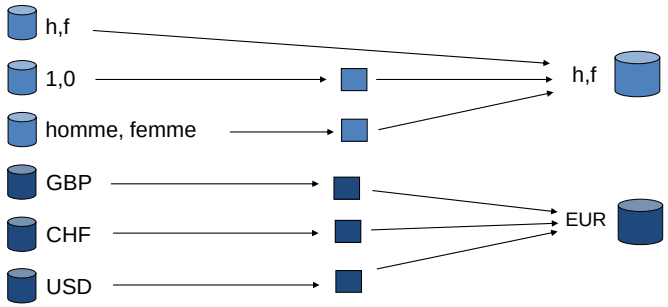
Les 4 caractéristiques des data warehouse

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

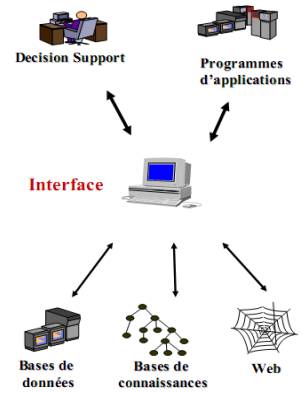


Les 4 caractéristiques des data warehouse

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



Intégration de données



Caractéristiques des sources de données :

- ➔ Hétérogènes (schématique, sémantique),
- ➔ Autonomes
- ➔ Évolutives
- ➔ Réparties

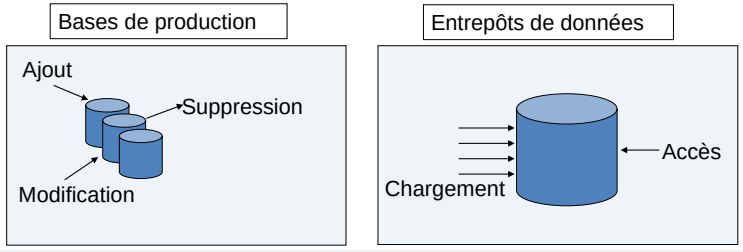
Besoins :

- ➔ Intégration données
- ➔ Gestion de l'évolution des données

Les 4 caractéristiques des data warehouse

3. Données non volatiles

- Traçabilité des informations et des décisions prises
- Copie des données de production



Les 4 caractéristiques des data warehouse

4. Données datées

- Les données persistent dans le temps
- Mise en place d'un référentiel temps

Base de production

Image de la base en Mai 2005

Répertoire

Nom	Ville
Karim	Alger
Salim	Oran

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Karim	Laghouat
Salim	Oran

Entrepôt de données

Calendrier

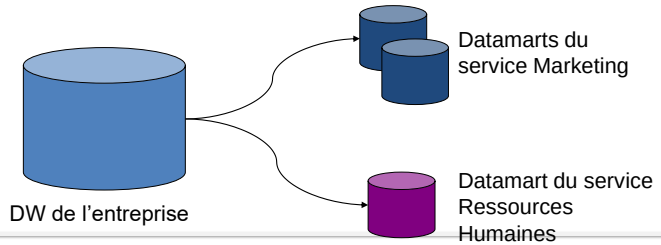
Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

Code	Nom	Ville
1	Karim	Alger
1	Salim	Oran
2	Karim	Laghouat

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

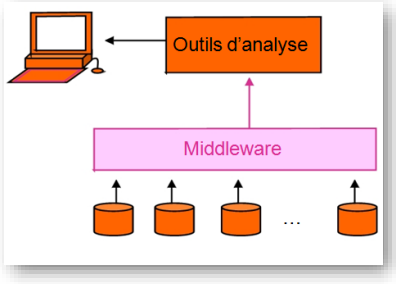


Intérêt des datamart

- Nouvel environnement structuré et formaté en fonction des besoins d'un métier ou d'un usage particulier
- Moins de données que DW
  - Plus facile à comprendre, à manipuler
  - Amélioration des temps de réponse
- Utilisateurs plus ciblés

Architecture d'un entrepôt de données

Approche virtuelle (ou le non-entrepôt)

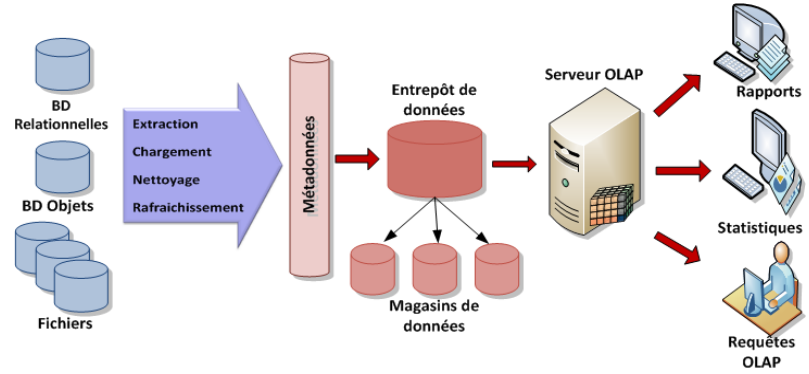


Inconvénients

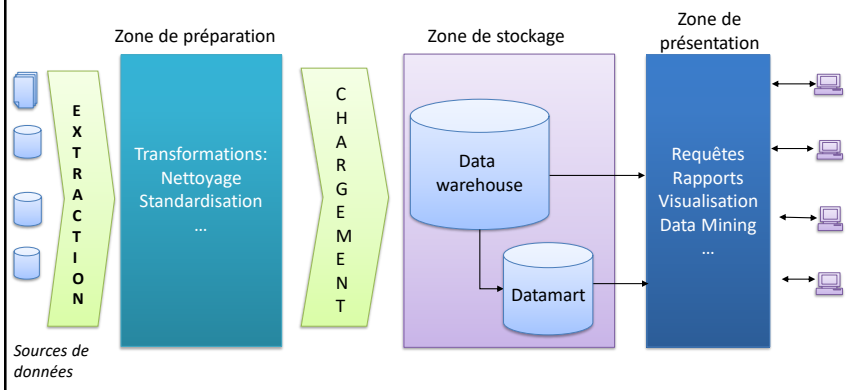
- ❑ pas de réelle intégration des données
- ❑ différentes vues non-réconciliées
- ❑ pas de vues dans le temps
- ❑ les requêtes peuvent facilement bloquer les transactions en cours

Architecture d'un entrepôt de données: physique

Approche entrepôt



Architecture générale



Les flux de données

- **Flux entrant**
  - Extraction: multi-source, hétérogène
  - Transformation: filtrer, trier, homogénéiser, nettoyer
  - Chargement: insertion des données dans l'entrepôt
- **Flux sortant**
  - Mise à disposition des données pour les utilisateurs finaux

Les différentes zones de l'architecture

- **Zone de préparation (Staging area)**
  - Zone temporaire de stockage des données extraites
  - Réalisation des transformations avant l'insertion dans le DW:
    - Nettoyage
    - Normalisation…
  - Données souvent détruites après chargement dans le DW
- **Zone de stockage (DW, DM)**
  - On y transfère les données nettoyées
  - Stockage permanent des données
- **Zone de présentation**
  - Donne accès aux données contenues dans le DW
  - Peut contenir des outils d'analyse programmés:
    - Rapports
    - Requêtes…

Exploitation de l'entrepôt

❑ Business Intelligence:

- Possibilité de visualiser et d'exploiter une masse importante de données complexes

❑ Trois principaux outils:

- OLAP :On-Line Analytical Processing
- Data mining: fouille de données
- Formulation de requêtes et visualisation des résultats

Architecture d'un entrepôt de données

- Souvent une architecture trois-tiers
  - Serveur d'entrepôt ("Warehouse Database Server")
    - Très souvent un système relationnel (ex. Oracle)
  - Serveur OLAP ("OLAP Server") de type ROLAP, MOLAP, ou HOLAP
  - Clients
    - Outils de requêtes et de production de rapports
    - Outils d'analyse et de prospection de données

Domaines d'applications

- Banque, Assurance
  - Détermination des profils client (prêt, …)
- Commerce
  - Ciblage de clientèle
  - Compagnies de grande production
  - Aménagement des rayons (2 produits en corrélation)
- Compagnies téléphoniques
- Santé

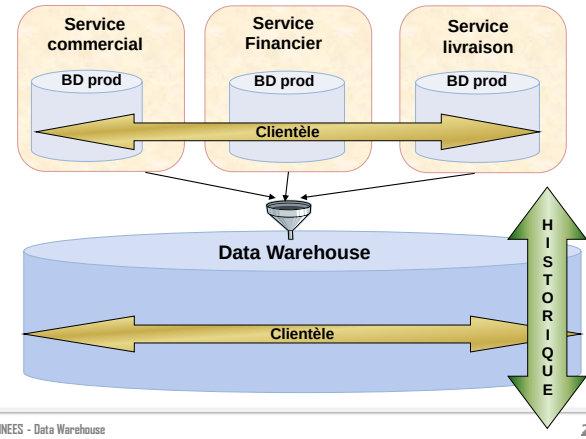
Base de données vs. Entrepôt de donnée

Pourquoi dissocier une BD d'un ED?

- Les objectifs de performances dans les BD ne sont pas les mêmes que ceux dans les EDs :
  - BD : requêtes simples (OLTP), méthodes d'accès et indexation
  - ED : requêtes OLAP souvent complexes!!!
- La nécessité de combiner des données provenant de diverses sources, d'effectuer des agrégations dans un ED et d'offrir des vues multidimensionnelles
- Les données d'un ED sont souvent non volatiles et ont donc une plus longue durée de vie que celles d'une BD

SGBD et DW

OLTP: On-Line Transactional Processing

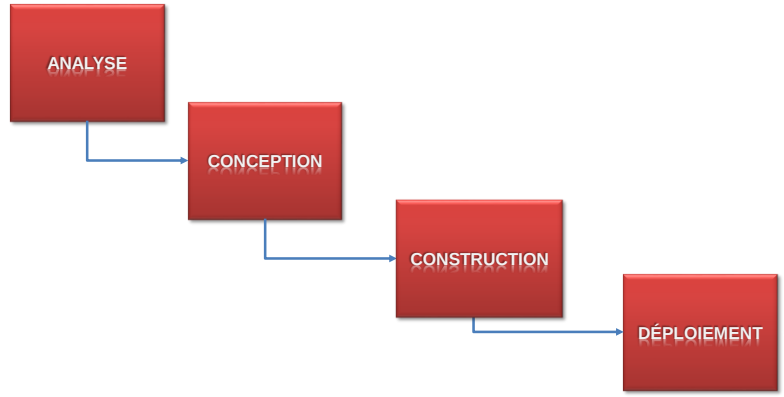


OLAP: On-Line Analitical Processing

OLTP VS DW

OLTP	DW
Orienté transaction	Orienté analyse
Orienté application	Orienté sujet
Données courantes	Données historisées
Données détaillées	Données agrégées
Données évolutives	Données statiques
Utilisateurs nombreux, administrateurs/opérationnels	Utilisateurs peu nombreux, manager
Temps d'exécution: court	Temps d'exécution: long

Cycle de vie de l'entrepôt de données



Cycle de vie

❑ Spécification des besoins

- Rassembler clairement et fidèlement les besoins des utilisateurs (décideurs)
- Clarifier les objectifs spécifiques
  - Comportement de la clientèle, analyse de tendances de prévisions, etc.
- Énumérer les dimensions
- Définir l'architecture du système (modèle de données), l'usage final (rapports, requêtes, outils d'analyse et visualisation)

Cycle de vie

❑ Analyse

- Développer le schéma de l'entrepôt
- Définir les processus nécessaires à la mise en place de l'entrepôt (extraction de données à partir des sources, transformations)

❑ Conception (3 niveaux)

- Conceptuel : mise au point du schéma, définition des méta-données
- Logique : adapté aux particularités du serveur de l'entrepôt (ROLAP, MOLAP, etc.)
- Physique: choix d'index, vues matérialisées, fragmentation

Cycle de vie

- **Construction**
  - Développer des programmes d'extraction, d'épuration et de transformation de données
- **Déploiement**
  - Fournir une installation initiale incluant la connexion aux données sources, la synchronisation et la réplication de données
  - Permettre des extensions futures
  - Offrir la formation pour les groupes d'intervenants
  - Offrir les divers mécanismes d'administration de l'ED (reprise, sécurité, performances)
  - Offrir les outils nécessaires à l'exploitation des données et à la consultation des méta-données

Outils requis

- Des outils d'accès aux BD
- Des outils d'intégration des données des BD vers l'ED
- Des outils de nettoyage de données
- Le gestionnaire de l'entrepôt de données
- Des outils de réplication

Modélisation classique (OLTP)

- **Le modèle relationnel**
  - Table, attributs, tuples, vues, …
  - Normalisation (redondance)
  - Requêtes simples (sélection, projection, jointure, …)
- **Le critère temps**
  - Représentation du passé
    - Un fardeau pour les systèmes OLTP

Requêtes décisionnelles plus complexes !!

- **Exemples**

- Combien de clients âgés entre 20 et 30 ans et résidant à Alger ont-ils acheté une caméra vidéo au cours des 5 dernières années ?
- Quelle est la répartition des ventes par produit, ville et par mois au cours de la présente année?
- Quelles sont les composantes des machines de production ayant eu le plus grand nombre d'incidents imprévisibles au cours de la période 1992-97 ?

➡ Critère temps est la base de l'analyse décisionnelle

Récapitulatif

	Base de données	Entrepôt de données
Opération	▪ Gestion courante	▪ Support à la décision
Modèle	▪ Entité association	▪ Etoile, flocon de neige
Normalisation	▪ Plus fréquente	▪ Rare
Données	▪ Actuelle, brutes	▪ Historiques, agrégées
Mise à jour	▪ immédiate	▪ Plus différée
Perception	▪ Bidimensionnel	▪ Multidimensionnelle
Opérations	▪ Lecture/écriture	▪ Lecture et rafraichissement
Taille	▪ Des giga-octets	▪ Vers des téra, péta-octets

OLAP

- ❑ Traitement analytique interactif (Codd) typique dans les systèmes informationnels
- ❑ Catégorie de traitements dédiés à l'aide à la décision
- ❑ Analyses diverses (multidimensionnelles)
- ❑ Information : surtout dérivée et sommaire
- ❑ Aide à la prise de décision

Modélisation Multidimensionnelle

❑ Dimension:

- Présente le point de vue selon lequel on veut voir le données décrites par un ensemble d'attributs· Axe de l'analyse
  - Exemple: Commandes, achats, réclamations, produits, clients,...

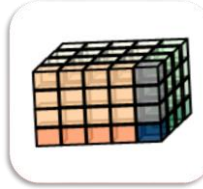
❑ Mesures/faits

- Fonction numérique qui peut être évaluée en tout point du data cube en agrégeant les données correspondant à ce point
- Mesure d'activité (critère d'analyse)
  - Exemple: Chiffre d'affaire, nombre de ventes, gain

Hyper cube OLAP

❑ Objectifs

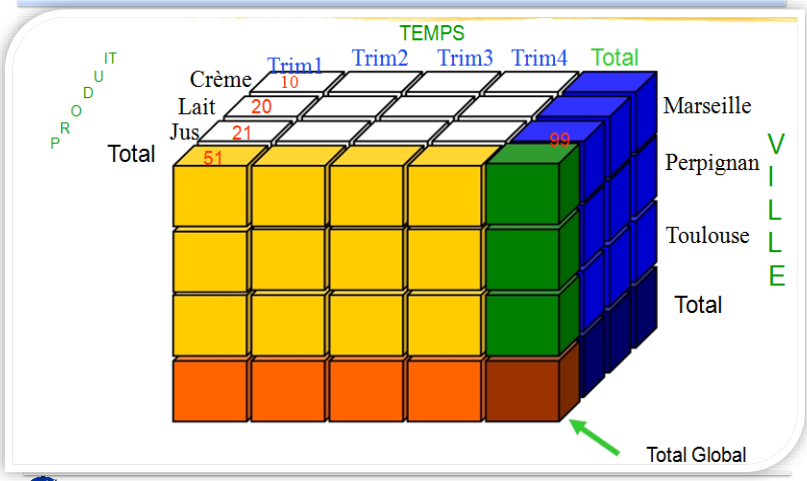
- Obtenir des informations déjà agrégées selon les besoins des utilisateurs
- Représentation de l'information dans un hyper cube à N dimensions



❑ Opérations OLAP

- Fonctionnalités qui servent à faciliter l'analyse multidimensionnelles: opérations réalisables sur l'hyper cube

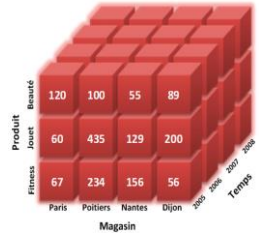
Exemple d'un cube de données



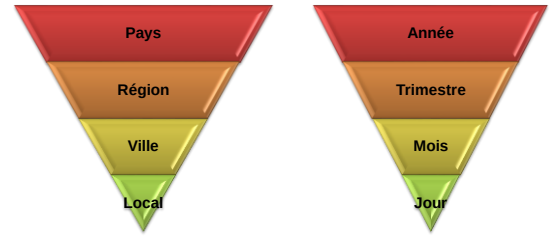
Vue multidimensionnelle des données

- Le volume des ventes (mesure) en fonction de : Produit, Temp et Localisation (Dimension)

Dimensions: Produit, Région, Temps



Hiérarchie des dimensions



Comment stocker le cube de données

- ROLAP: Relational On-Line Analytical Processing
  - Les données sont stockées comme des tables relationnelles: une table de faits et des tables de dimension
- MOLAP: Multidimensional On-Line Analytical Processing
  - le cube de données est stocké sous forme d'un tableau multi-dimensionnel.

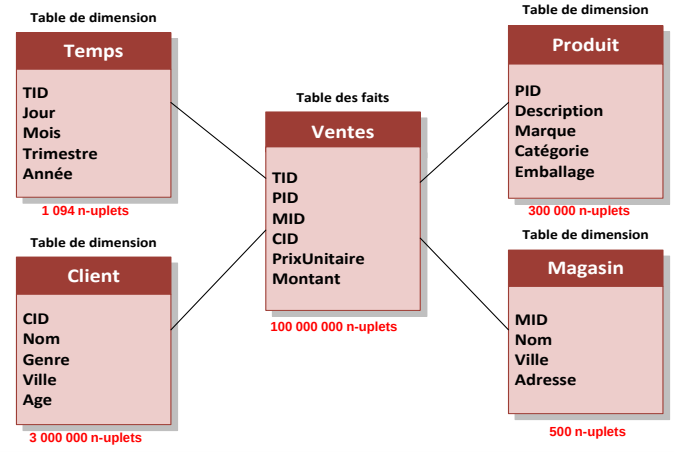
Modèle ROLAP

- Exploiter l'expérience des modèles relationnels (un grand succès!!)
- Il faut des modèles bien adaptés aux ED!
  - Schéma en étoile (star schema)
  - Schéma en flocon de neige (snowflake schema)
  - Schéma en constellation

Modèle en étoile

- Autant de tables de dimension qu'il existe de dimensions
  - **Exemple**
    - Temps, Produit, Client
- Une table de faits contenant la clé de chaque dimension et des mesures
  - **Exemple**
    - Montant en dollars, nombre d'unités vendues

Schéma en étoile



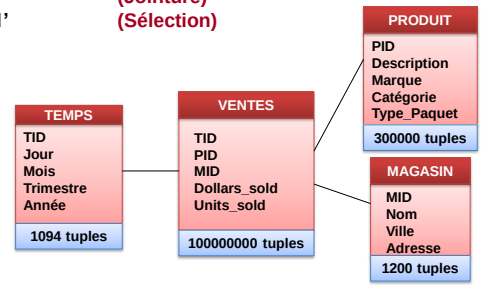
Requête type

Requête de Jointure en étoile :

```
SELECT    P.Marque, sum(dollars_sold),
sum(units_sold)
FROM      Ventes V, PRODUIT P, TEMPS T
WHERE     V.PID = P.PID              (Jointure)
AND       V.TID = T.TID              (Jointure)
AND       T.Trimestre = 'T1'         (Sélection)
GROUP BY  P.Marque
ORDER BY  P.Marque
```

Requêtes de jointure en étoile

- Plusieurs jointures
- Suivies par des sélection



Avantages & Inconvénients

+ Simple

+ Le plus utilisé !!!

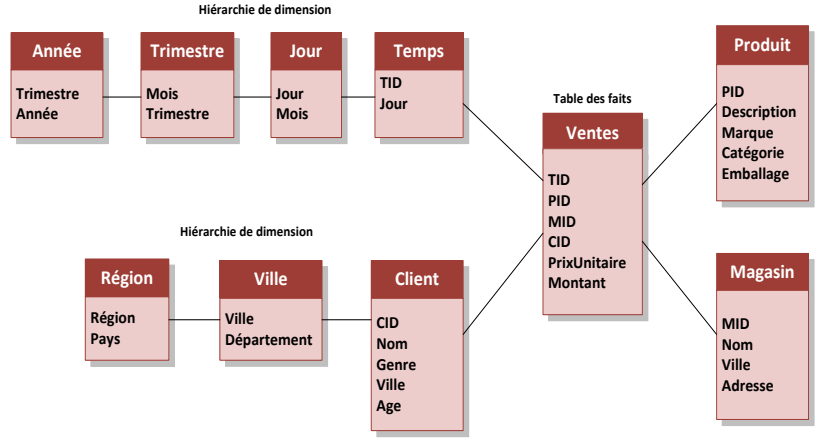
- Possibilité de redondance car les tables de dimension ne sont pas nécessairement normalisées

- Taille de dimensions plus grosse

Modèle en flocon de neige

- Variante du modèle en étoile
- Les tables de dimension sont normalisées
- Réduction de la redondance mais exécution parfois plus lente des requêtes (jointure de tables)
- Modèle adopté par Oracle!!

Exemple d'un modèle en flocon de neige



Définition d'un schéma en étoile avec DMDL

define cube ventes_star [Temps, Produit, Client]

Montant_vente = sum(somme),

moyenne_vente = avg(somme),

unités_vendues = count(*)

Mesures

Dimensions

define dimension Temps as (TID, Année, mois, jour)

define dimension Produit as (PID, nom_item, marque, taille, poids)

define dimension Client as (Cid, Nom, Sexe, Age, Ville)

Définition d'un schéma snowflake avec DMQL

define cube ventes_snowflake [temps, item, branche, lieu]

Montant_vente = sum(somme),

moyenne_vente = avg(somme),

unités_vendues = count(*)

Mesures

Dimensions

define dimension item as (Id_item, nom_item, marque, type,

fournisseur(Id_fournisseur, type_fournisseur))

define dimension branche as (Id_branche, nom_branche, type_branche)

define dimension temps as (Id_temps, jour, jour_semaine, mois, trimestre, année)

define dimension lieu as (Id_lieu, rue, ville(Id_ville, département, pays))

Hiérarchies

MOLAP: Représentation Tableaux

Produit	Temps	Trimestre 1			Trimestre 2			Trimestre 3			Trimestre 4			Total		
	Ville	P	N	Total	P	N	Total	P	N	Total	P	N	Total	P	N	Total
TV LCD		12	34	46	22	36	58	24	37	61	33	55	88	91	162	253
Lecteur DVD		29	66	95	44	50	94	56	55	111	44	39	83	173	210	383
Caméoscope		55	34	89	69	27	96	31	26	57	68	70	138	223	157	380
Total		96	134	230	135	113	248	111	118	229	145	114	309	487	529	1016

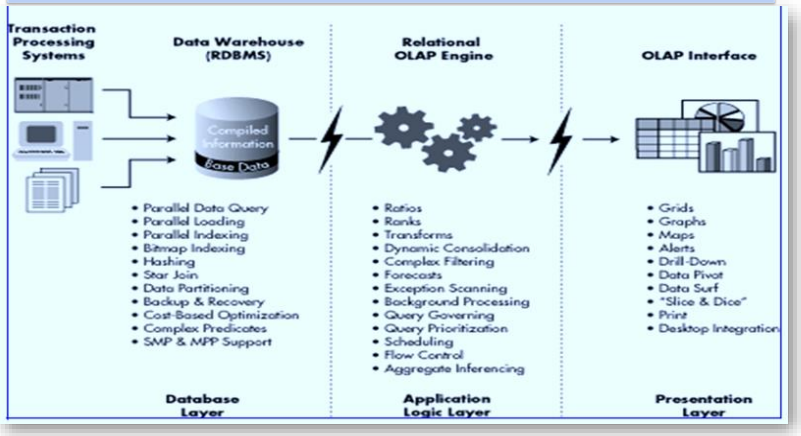
Répartition des ventes par produit, temps et ville

P : Poitiers, N : Nantes

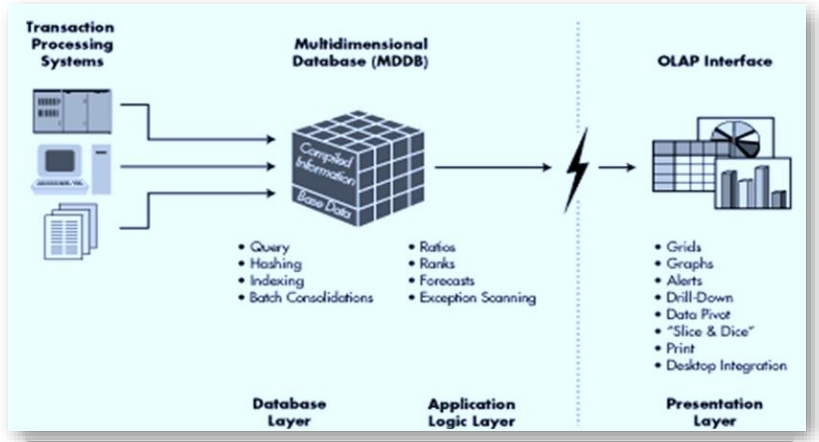
ROLAP vs. MOLAP

	Avantages	Inconvénients
ROLAP	▪ Standard bien établi ▪ Efficace pour le transactionnel ▪ Capacité d'expansion (téra-octet)	▪ Absence de vue conceptuelle ▪ SQL peut être inadéquat pour la formation de tableaux croisés
MOLAP	▪ Implémentation souvent plus performante que ROLAP	▪ Inadéquat pour le transactionnel ▪ Capacité d'expansion limitée (dizaine de géga-octet) ▪ Absence de standard

Serveurs ROLAP (Microstrategy)



Serveurs MOLAP



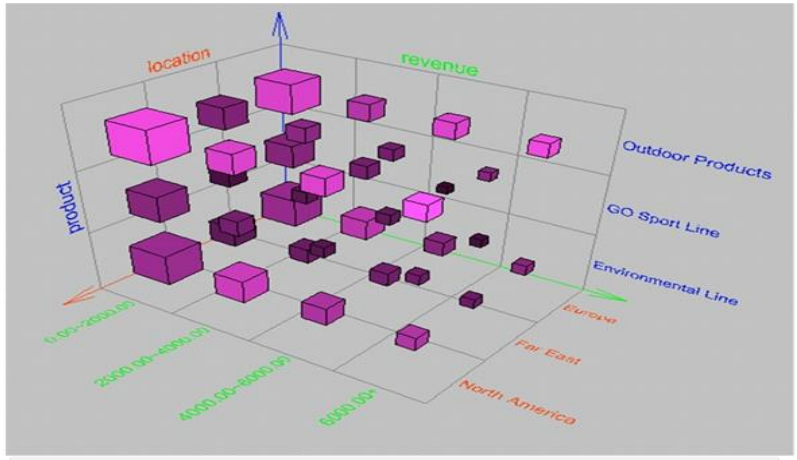
Hybrid OLAP

- HOLAP combine ROLAP et MOLAP
- Les données sont partitionnées en deux sous-ensembles
  - Données non fréquentes sont stockées comme tables relationnelles
  - Données fréquentes sont stockées comme tableaux multidimensionnels
- Données brutes dans ROLAP
- Données agrégées dans MOLAP
- Cette séparation est transparente pour l'utilisateur.

Opérations de base OLAP

- Group By
- Roll-up (ou drill-up)
- Drill-down
- Slice & dice
- Pivot
- Opérations OLTP classiques

Navigation dans un cube



Opérations multidimensionnelles

- Agrégations (consolidation) de données (pré- calculées)
  - Réduction de temps de calcul
- Définition de hiérarchies dans les dimensions («granularités»)
- Exemples
  - Temps: jours/mois/trimestres/années
  - Géographie: ville/région/pays

ROLL-UP

- Agrégation des mesures en allant d'un niveau particulier de la hiérarchie vers un niveau plus général
- Exemples
  - Ayant le total des ventes par ville, ROLL-UP permet de donner le total des ventes par âys

Syntaxe:

```
SELECT ...
GROUP BY
ROLLUP(Grouping-column-reference_list)
```

Drill-Down: opération inverse de roll-up

Slice & dice

- **Slice et dice**
  - **projection et sélection du modèle relationnel**
- **Pivot (rotate)**
  - **Réoriente le cube pour visualisation**