Introduction aux entrepôts de données

Bernard ESPINASSE
Professeur à Aix-Marseille Université (AMU)
Ecole Polytechnique Universitaire de Marseille

Mars 2021

- **Les entrepôts de données**
- **Modélisation et implantation d'un entrepôt de données**
- **Alimentation d'un entrepôt de données (ETL)**
- **Exploitation d'un entrepôt de données (OLAP)**
- **Domaines d'application et exemples d'entrepôts de données**

Plan

1. **Définition d'un entrepôt de données**
   - Définition d'un entrepôt de données
   - De l'entrepôt à l'aide à la décision
   - Entrepôts de données Versus Magasins de données
   - Architecture fonctionnelle d'un entrepôt de données
2. **Modélisation et implantation d'un entrepôt de données**
   - Modélisation multidimensionnelle
   - De la table au cube
   - Stratégies d'implantation : ROLAP, MOLAP et HOLAP
   - Schéma en étoile, en flocon et en constellation
3. **Alimentation d'un entrepôt de données**
   - Processus général : ETL
   - Préparation des données
   - Intégration des données
4. **Exploitation d'un entrepôt de données**
   - Problématique de l'OLAP, OLAP versus OLTP
   - Les opérations élémentaires et langages de l'OLAP
   - Reporting, tableaux de bords et visualisation autour de l'OLAP.
5. **Domaines d'application et exemples d'entrepôts de données**

1 – Définition d'un entrepôt de données

- **Définition d'un entrepôt de données**
- **De l'entrepôt à l'aide à la décision**
- **Entrepôts de données Versus Magasins de données**
- **Architecture fonctionnelle d'un entrepôt de données**

Définition d'un entrepôt de données (Data Warehouse)

Définition de Inmon (1992) :

« une **collection** de **données thématiques**, **intégrées**, **non volatiles** et **historisées**, **organisées** pour le support d'un processus d'aide à la décision»

Données :

- **Thématique ou orientées sujet :** un ED **rassemble** et **organise** des données associées **aux différentes structures fonctionnelles de l'entreprise**, pertinentes pour un **sujet** ou **thème** et nécessaire aux **besoins d'analyse**
- **Intégrées :** les données résultent de **l'intégration de données** provenant de **différentes sources** pouvant être **hétérogènes**
- **Historisées :** les données d'un ED représentent **l'activité d'une entreprise durant une certaine période** (plusieurs années) permettant de **d'analyser les variations d'une donnée dans le temps**
- **Non-volatiles :** les données de l'ED sont **essentiellement utilisées en interrogation** (consultation) et ne peuvent pas être modifiées (sauf certain cas de rafraîchissement).

De l'entrepôt à l'aide à la décision (1)

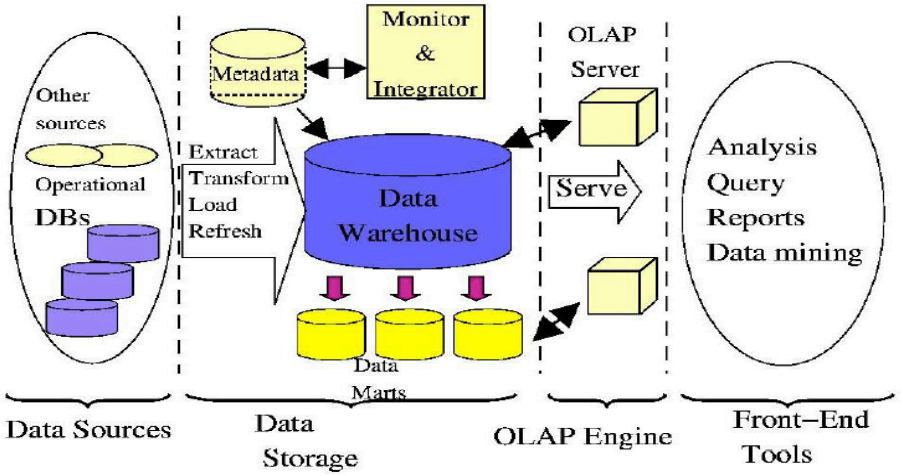
Entreposage des données : avant d'être chargées dans l'entrepôt, les données sélectionnées doivent être :

- **extraites des sources** (internes : BD opérationnelles, externes : BD et fichiers notamment issus du Web)
- **soigneusement épurées** afin d'éliminer des erreurs et réconcilier les différentes sémantiques associées aux sources)

Exploitation des données de l'ED : **systèmes décisionnels**

- A partir des données d'un ED **diverses analyses** peuvent être faites, notamment par des techniques « On-Line Analitical processing » (**OLAP**) ou de **fouille de données** (Data Mining) et de **visualisation**.
- Notons que les informations et connaissances obtenues par exploitation de l'ED ont un **impact direct** **sur les bénéfices de l'entreprise** (augmentation des ventes par un marketing plus ciblé, amélioration de la rotation des stocks, …)

De l'entrepôt à l'aide à la décision (2)



Entrepôt de données VS Bases de données opérationnelles

	BD opérationnelles	Entrepôt de données
Niveau de détail des informations	▪ Très détaillé	▪ Données agrégées, métadonnées
Homogénéité des informations	▪ Informations homogènes	▪ Information pas nécessairement homogènes, ▪ intégration de données souvent nécessaire
Fonctions de l'entreprise concernées par les données	▪ Données organisées par processus fonctionnel	▪ Données orientées sujet
Comparaison de données sur plusieurs années	▪ Non : Archivage ou mise à jour des données	▪ Oui : Données non volatiles, données historisées
Opérations réalisées sur les données	▪ Consultation, mais surtout mise à jour et ajout de données	▪ Consultation de données uniquement

Entrepôts de données Versus Magasins de données

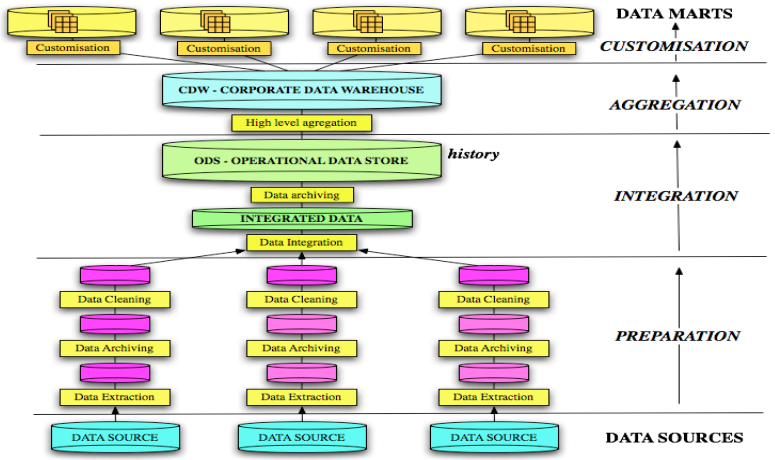
L'entrepôt de données - ED (Data Warehouse - DW) :

- nécessitent de **puissantes machines** pour gérer de **très grandes bases de données** contenant des **données de détail historisées**
- lieu de **stockage centralisé** d'un extrait des bases de production.
- **l'organisation des données** est faite selon un modèle **facilitant la gestion efficace des données et leur historisation**.

Les magasins de données – MD (Data Marts - DM) :

- **petits entrepôts** nécessitant une **infrastructure plus légère** et sont mis en œuvre plus **rapidement** (6 mois environs)
- **conçus pour l'aide à la décision** à partir de **données extraites d'un ED** plus conséquent ou de BD sources existantes
- les **données extraites sont adaptées pour l'aide à la décision** (pour classe de décideurs, usage particulier, recherche de corrélation, logiciel de statistiques,...)
- **l'organisation des données** est faite selon **un modèle facilitant les traitements décisionnels**

Entrepôts et magasins de données (3)



- **ODS Operational Data Store** : regroupe les données intégrées récupérées des sources
- **CDW Corporate Data Warehouse** : regroupe les vues agrégées

Architecture fonctionnelle d'un entrepôt de données

Niveau extraction :

Extraction de données des BD opérationnelles (SGBD traditionnel en OLTP) et de l'extérieur :

- ***approche « push »*** *: détection **instantanée** des mises à jour sur les BD opérationnelles pour intégration dans l'ED*
- ***approche « pull »*** ***:*** *détection **périodique** des mises à jour des BD opérationnelles pour intégration dans l'ED*

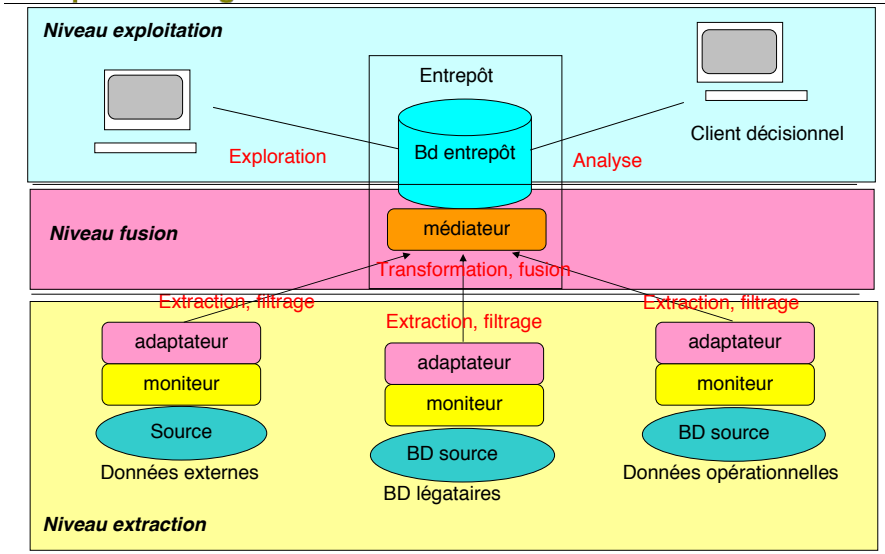
Niveau fusion :

- **Intégration**, **chargement** et **stockage** des données dans la BD entrepôt organisée par sujets
- **Rafraîchissement** au fur et à mesure des mises à jour

Niveau exploitation :

- **Rapports, tableaux de bords, visualisation graphiques diverses, …**
- **Analyse** et **l'exploration** des données entreposées (OLAP)
- **Requêtes complexes** pour analyse de tendance, extrapolation, découverte de connaissance, … (Fouille de données)

Composants logiciels d'un ED



3 – Modélisation et implantation d'un entrepôt de données (ED)

- **Modélisation multidimensionnelle**
- **De la table au cube**
- **Stratégies d'implantation**
- **Schéma en étoile, en flocon et en constellation**

Modélisation multidimensionnelle (1)

- Les analyses décisionnelles (**OLAP)** sont directement reliées à une **modélisation de l'information spécifique :**
  - **proche de la perception qu'en a l'analyste**
  - basée sur une **vision multidimensionnelle des données**
- **Modélisation « multidimensionnelle »** :
  - considère un sujet analysé comme un **point dans un espace à plusieurs dimensions**
  - les **données y sont organisées** de façon **à mettre en évidence le sujet analysé** et les différentes **perspectives de l'analyse**.

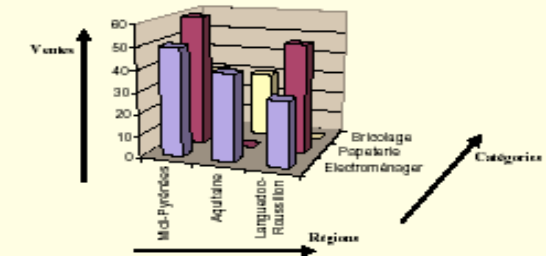
Modélisation multidimensionnelle (2)

Soit les données relatives aux ventes de 1999 d'une entreprise de distribution :

Catégories des produits	Régions	Montant des ventes
Electroménager	Midi-Pyrénées	50
Electroménager	Aquitaine	40
Electroménager	Languedoc-Roussillon	30
Papeterie	Midi-Pyrénées	60
Papeterie	Languedoc-Roussillon	50
Bricolage	Midi-Pyrénées	30
Bricolage	Aquitaine	30

Différentes perspectives pour observer ces données :

- une dimension relative à la **catégorie des produits**
- une dimension relative à la **région**



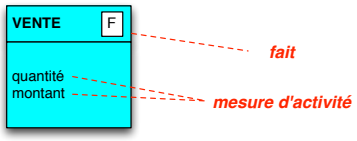
Modélisation multidimensionnelle : concept de fait

Un fait :

- modélise le **sujet** de l'analyse
- est formé de **mesures** correspondant aux informations de l'activité analysée.
- ces mesures sont **numériques** et généralement ***valorisées de façon continue***, on peut les ***additionner***, les ***dénombrer*** ou bien ***calculer*** le minimum, le maximum ou la moyenne.

Exemple : le fait de « Vente » peut être constitué des mesures d'activités suivantes :

- quantité de produits vendus et
- montant total des ventes



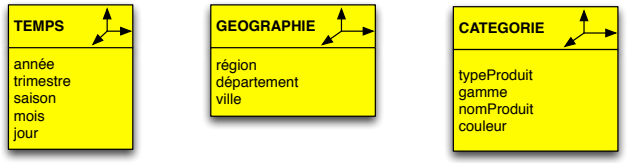
Modélisation multidimensionnelle : concept de dimension

Le sujet analysé, le fait, est **analysé** suivant **différentes perspectives** ou **axes** caractérisant ses mesures de l'activité : on parle de **dimensions**.

Une dimension :

- modélise **un axe d'analyse**
- se compose de **paramètres** correspondant aux informations faisant varier les mesures de l'activité.

Ex: Dans l'exemple précédent, le fait « ***Vente*** » peut être analysé suivant différentes perspectives correspondant à trois dimensions : la dimension ***Temps***, la dimension ***Geographie*** et la dimension ***Categorie*** :



Modélisation multidimensionnelle : hiérarchie de dimension

- Les faits sont **analysées selon les dimensions qui les caractérisent**
- Nécessaire de définir pour **chaque dimension** ses **différents niveaux hiérarchiques de détail** (d'agrégation),
- **Les hiérarchies de dimensions** définissent des **niveaux de détail de l'analyse sur les dimensions**

  Ex:

  - ***Dimension « temps» :***
    - H1 : **jour -> mois -> année** ;
    - H2 : **jour -> mois -> trimestre -> année** ;
    - H3 : **jour -> mois -> saison -> année** ;
  - ***Dimension « géographie » :*** **ville -> département -> région** (chaque ville appartient à un département qui est situé dans une région)
  - ***Dimension « catégorie» :*** **couleur -> nomProduit -> gamme -> typeProduit** (chaque produit appartient à une gamme de produit qui appartient à un type de produit)

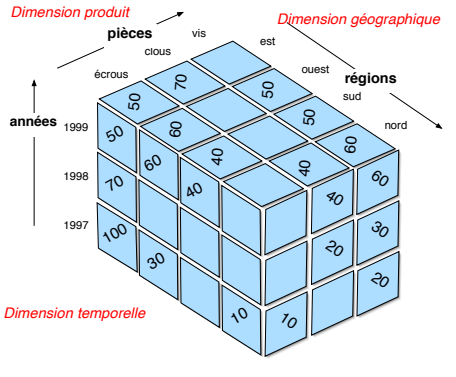
De la table … au cube

De la table …
Table Ventes :

VENTES	pièces	Régions	Années	quantités
	écrous	est	1999	50
	clous	est	1997	100
	vis	ouest	1998	50
	…	…	…	…
	écrous	est	total	220
	…	…	…	…
	écrous	total	total	390
	…	…	…	…
	total	total	total	1200

(pièce, région, année) ⊡ quantité

… au cube
Cube Ventes :



Stratégies d'implantation d'un ED

3 stratégies :

1 - Usage d'un SGBD Relationnel (systèmes ROLAP)

- les **SGBDR représentant plus de 80% des SGBD** : ils sont principalement envisagés pour le développement d'ED mais doivent être adaptés
- Ils doivent cependant **être adaptés** car ils n'ont pas les caractéristiques adéquates pour répondre aux besoins des ED.

2 - Usage d'un SGBD Multidimensionnel (systèmes MOLAP)

- Un SGBD Multidimensionnel (SGBDM) est **un SGBD capable de stocker et traiter des données multidimensionnelles**
- A ce jour **pas encore de cadre technologique commun** pour le développement de tels systèmes : chaque produit est spécifique

3 - Usage d'un SGBD Hybride (systèmes HOLAP)

Tire profit des avantages des technologies ROLAP et MOLAP :

- un **ROLAP** pour stocker, gérer les ***données détaillées ET***
- un **MOLAP** pour stocker, gérer les ***données agrégées***

Schéma d'un entrepôt de données

Niveau logique « ROLAP » :

3 grands types de schémas :

- schéma en **étoile** (star schema)
- schéma en **flocon** (snowflake schema)
- schéma en **constellation** (fact constellation)

⇨ le schéma en **étoile** est souvent utilisé pour **l'implantation physique**

Schéma en étoile (1)

- **table des faits** : normalisée, de taille très importante, avec de nombreux champs
- **tables de dimensions** : dimensions de l'analyse, taille peu importante, avec peu de champs

Ex 1 : Vente de médicaments dans des pharmacies

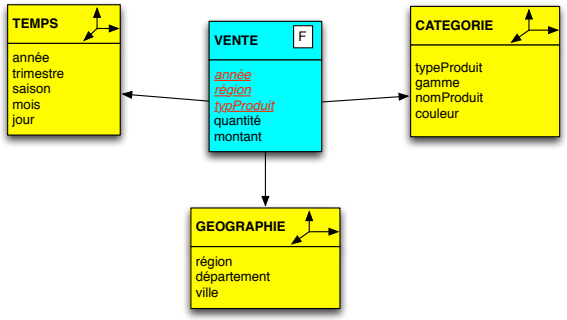


Schéma en étoile modélisant les analyses des quantités et des montants des médicaments dans les pharmacies selon 3 dimensions : le temps, la catégorie et la situation géographique

- Table de **faits** : ***Vente***
- Tables de **dimension** : ***Temps, Catégorie, Géographie***

Schéma en étoile (3)

Ex 2 : Ventes d'articles dans un supermarché

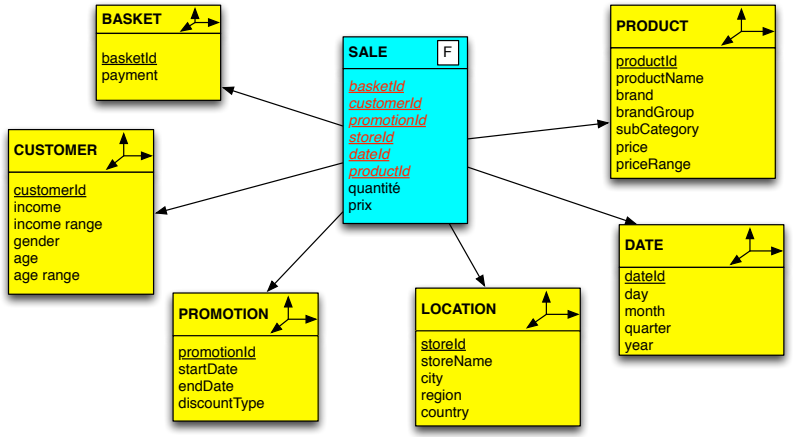


Schéma en étoile (4)

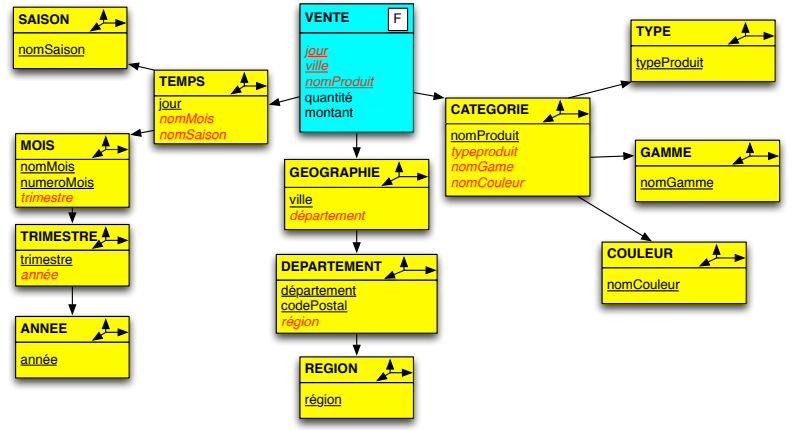
Associé à Ex 2 :

- **un fait** :
  - il a été acheté **3 exemplaires à 1 euro** (SALE)
  - du **produit pid3**
  - par le **client cid1**
  - à la **date did3**
  - dans le **magasin mid2** (store)
  - dans le **chariot cid8** (basket)
  - correspondant à la **promotion prid1**
- **un élément de la dimension location :**
  - store id **mid2**
  - store name **rondpoint**
  - city **blois**
  - region **centre**
  - country **France**

Schéma en flocon (1)

Evolution du schéma en étoile avec une **décomposition des tables de dimensions** du modèle en étoile **selon leurs hiérarchies** (normalisation des tables de dimensions)

Ex 3: Vente de médicaments dans des pharmacies

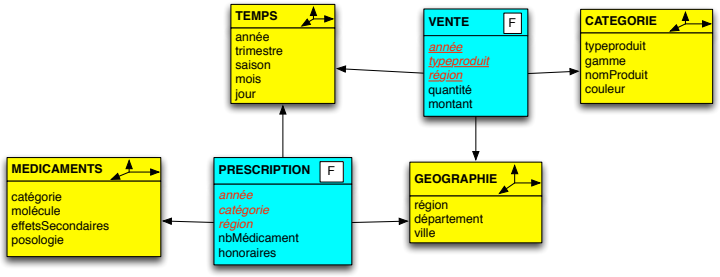


Chaque dimension du schéma en étoile précédent est **dénormalisée**

Schéma en constellation

- **fusionne plusieurs modèles** en étoile qui utilisent des dimensions communes.
- comprend en conséquence **plusieurs faits** et des **dimensions communes ou non**

Ex : Vente de médicaments dans des pharmacies



- une constellation est constituée de **2 schémas en étoile** :
  - l'un correspond aux VENTEs effectuées dans les pharmacies et
  - l'autre analyse les PRESCRIPTIONs des médecins
- les dimensions ***Temps*** et ***Géographie*** sont partagées par les faits PRESCRIPTION et VENTE.

4 - Alimentation d'un Entrepôt de données

- **Processus général d'alimentation d'un ED**
- **Processus et outils ETL**
- **Taches d'un processus ETL**

Processus d'alimentation d'un ED

- Le **processus d'alimentation d'un ED (ou entreposage des données)** consiste à :
  - **rassembler** de multiples données sources souvent ***hétérogènes***
  - les **homogénéiser**
- **Homogénéisation** faite selon des ***règles précises***
- **Ces règles** :
  - sont mémorisées sous forme de **méta-données** (information sur les données) stockées dans le dictionnaire de données
  - permettent d'assurer des **tâches d'administration** et de **gestion des donnés entreposées**.

Processus d'alimentation d'un ED

Après avoir conçu le modèle des données, comment alimenter l'ED ?

→ problématique de l'ETL (Extracting Transforming and Loading)

4 étapes :

1. **Sélection des données sources**
2. **Extraction des données**
3. **Nettoyage et Transformation**
4. **Chargement**

Les processus ETL (Extract Transform Load)

- Sont des **opérations de migration de données** pour les systèmes de Business Intelligence notamment les entrepôts de données
- Ils gèrent toutes les étapes de collecte et de transfert des données dans les ED :
  - **Extraction de données** des applications et des bases de données de gestion et de production (ERP, CRM, SGBDR, fichiers, etc.).
  - **Transformation de ces données** pour les consolider et les mettre en concordance.
  - **Chargement** : distribution des données auprès des applications cibles ou des systèmes décisionnels (Data Warehouse, Data Marts, applications OLAP ou "cubes"…).
- **Différents outils logiciels sont proposés :**
  - **Talend** : outils ETL en mode open source le plus connu
  - Autres solutions : Apatar, Jitterbit et Pentaho.
  - IBM, Informatica ou SAP11 proposent également des solutions ETL

Les outils ETL (Extract Transform Load)

Ils permettent un support et/ou automatisation des tâches suivantes :

Tâches	Support
Extraction	accès aux différentes sources
Nettoyage	recherche et résolution des inconsistances dans les sources
Transformation	entre différents formats, langages, etc.
Chargement	des données dans l'entrepôt
Réplication	des sources dans l'entrepôt
Analyse	Ex : détection de valeurs non valides ou inattendues
Transfert de données haut débit	pour les très grands entrepôts
Test de qualité	Ex : pour correction et complétude
Analyse des méta données	aide à la conception

1 - Tâche de sélection des données sources

Quelles données de production faut-il sélectionner pour alimenter l'ED?

→ Toutes les données sources ne sont forcément pas utiles

Ex : Doit-on prendre l'adresse complète ou séparer le code postal ?

→ Les données sélectionnées seront réorganisées pour devenir des informations.

- La **synthèse** de ces données sources a pour but de les enrichir.
- La **dénormalisation** des données crée des liens entre les données et permet des accès différents

2 - Tâche d'Extraction des données

Un extracteur (wrapper) est associé à chaque source de données :

- Il sélectionne et extrait les données
- Il les formate dans un format cible commun
- Utilisation d'interfaces comme **ODB**, **OCI**, **JDBC**.
- Le format cible est en général le **modèle Relationnel**

3 - Tâche de Nettoyage et Transformation des données

Objectifs du nettoyage :

- résoudre le problème de **consistance** des données au sein de chaque source
- *une centaine de type d'inconsistances ont été répertoriées*
- *5 à 30 % des données des BD commerciales sont erronées*

Types d'inconsistances :

- présence de données fausses dés leur saisie :
  - *fautes de frappe*
  - *différents formats dans une même colonne*
  - *texte masquant de l'information (e.g., "N/A")*
  - *valeur nulle*
  - *incompatibilité entre la valeur et la description de la colonne*
  - *duplication d'information, …*
- persistance de données obsolètes
- confrontation de données sémantiquement équivalentes mais syntaxiquement différentes

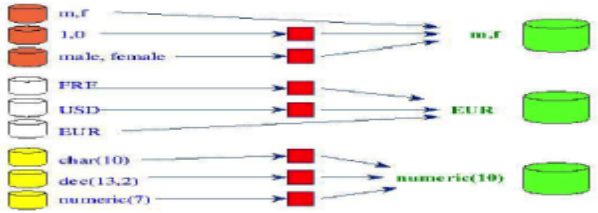
3.1 Tâche de Nettoyage des données

- fonctions de **normalisation**
- fonctions de **conversion**
- usage de **dictionnaires de synonymes ou d'abréviations**

Définition de table de règles :

valeur source	*remplacé par*	Valeur cible
Mr		M
monsieur		M
Masculin		M
M		M
Msieur		M

Exemple de conversions :



nettoyage = jointure + projection

3.2 Tâche de Transformation des données

Objectifs :

Suppression des incohérences sémantiques entre les sources pouvant survenir lors de l'intégration :

- des **schémas** :
  - ***problème de modélisation*** : différents modèles de données sont utilisés
  - ***problèmes de terminologie*** : un objet est désigné par 2 noms différents, un même nom désigne 2 objets différents
  - ***incompatibilités de contraintes*** : 2 concepts équivalents ont des contraintes incompatibles
  - ***conflit sémantique*** : choix de différents niveaux d'abstraction pour un même concept
  - ***conflits de structures*** : choix de différentes propriétés pour un même concept
  - ***conflits de représentation*** : 2 représentations différentes choisies pour les mêmes propriétés d'un même objet
- des **données :**
  - ***Equivalence de champs***
  - ***Equivalence d'enregistrements*** : fusion d'enregistrements

4 - Tâche de Chargement des données

Objectif :

charger les données nettoyées et préparées dans l'ED

C'est une opération :

- qui risque d'être **assez longue**
- plutôt **mécanique** et la **moins complexe**.

Il est nécessaire de définir et mettre en place :

- des **stratégies pour assurer de bonnes conditions à sa réalisation**
- une **politique de rafraîchissement**.

5 – Exploitation d'un entrepôt de données (ED)

- **OLAP, OLAP versus OLTP**
- **Les opérations élémentaires de l'OLAP et langages pour l'OLAP**
- **Réalisation de rapports divers *(Reporting)***
- **Réalisation de tableaux de bords à partir d'un ED *(Dashboards)***
- **Visualisations autour d'un ED *(visualizations)***

OLAP (On Line Analytical Processing)

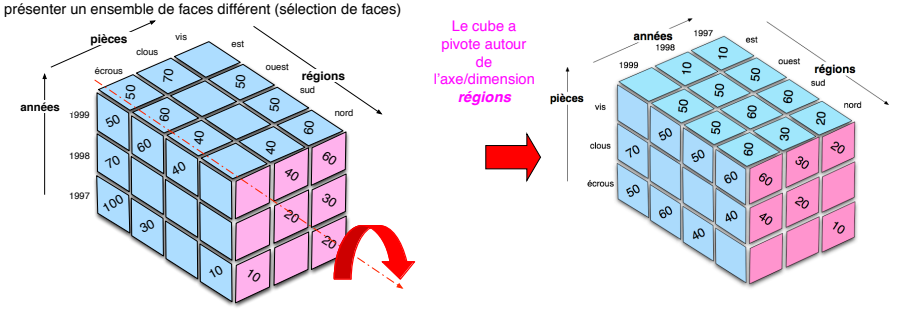
- Apparue dans les **années 90 dans les entreprises**
- **Façon la plus naturelle d'exploiter un ED** du fait de son organisation multidimensionnelle - cube
- Permettent de **réaliser** des **synthèses**, des **analyses** et de la **consolidation dynamique** de **données multidimensionnelles organisé en cubes**
  Exemple d'analyse OLAP : un supermarché ANALYSANT l'ensemble de ses ventes
- **3 types d'opérations OLAP élémentaires** liées à des **transformations du cube** :
  - **Restructuration** : permet un changement de points de vue selon différentes dimensions : opérations liées à la structure, manipulation et visualisation du cube : **Rotate/pivot**, **Switch**, **Split, nest, push, pull**
  - **Granularité** : concerne un changement de niveau de détail : opérations liées au niveau de granularité des données : **roll-up, drill-down**
  - **Ensembliste** : concerne l'extraction et l'OLTP classique : **slice, dice**, **selection**, **projection**, **jointure (drill-across)**

OLTP versus OLAP

	Caractéristiques	**OLTP**	**OLAP**
Conception	*Orientation*	Transaction	Analyse
	Conception	Entité-Relation	Star/snowflake
Données	*Granularité*	Détail	Résumées, agrégées
	Nature	Relationnelle	Multidimentionnelle
	Actualisation	Actualisées, mises à jour	Historisées, recalculées
	Taille fichiers	< 100 Mo/Go	> 100 Go/To
Traitements	*Unité de travail*	Transaction simple	Requête complexe
	Accés	Lecture/écriture	Lecture
	Nb de tuples accédés	Dizaines	Millions
	Métrique	Débit de transactions	Temps de réponse
Utilisateurs	*Utilisateur*	Agent opérationnel	Analyste/décideur
	Nombre d'utilisateurs	Milliers	Centaines

Exemple d'opérations OLAP de restructuration : rotate/pivot

Rotate/pivot : effectue au cube une rotation autour d'un de ses 3 axes passant par le centre de 2 faces opposées, de façon à présenter un ensemble de faces différent (sélection de faces)



la visualisation résultante est souvent 2D :

nord	1999	1998	1997
vis	60	30	20
clous	40	20	
écrous			10

vis	1999	1998	1997
est		10	10
ouest	50	50	50
sud	50	60	60
nord	60	30	20

Langages pour l'OLAP

- **2 langages possibles pour faire de l'OLAP :**

1. SQL étendu (Extensions de SQL-3 / SQL-99 pour OLAP) :

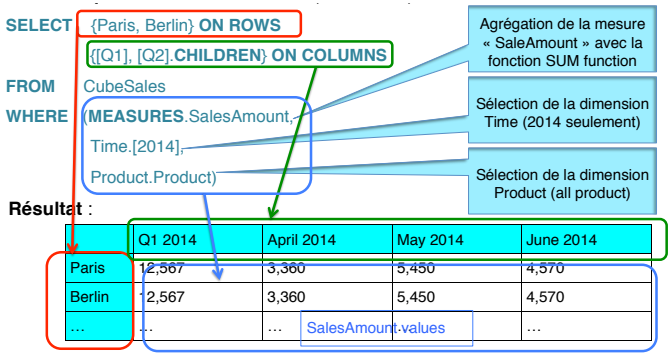
- Nouvelles fonctions SQL d'agrégation: *Rank, N_tile, …*
- Nouvelles fonctions de la clause GROUP BY :
  - *ROLLUP equivalent to "control breaks"*
  - *CUBE equivalent to "cross tabulation"*
  - *GROUPING SETS equivalent to multiple GROUP BYs*
- Fenêtre glissante : *WINDOWS/OVER/PARTITION, …*

2. MDX (Multi Dimensional eXpression) :

- **langage de requêtes inventé pour faire de l'OLAP** par Mosha Pasumansky (Microsoft)
- disponible dans la plupart des serveurs OLAP
- plus puissant que SQL pour faire de l'OLAP

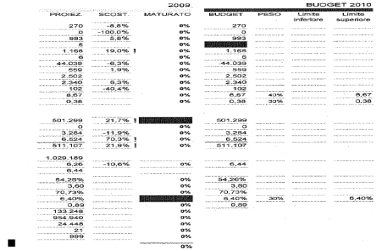
Exemple de requête OLAP en MDX :

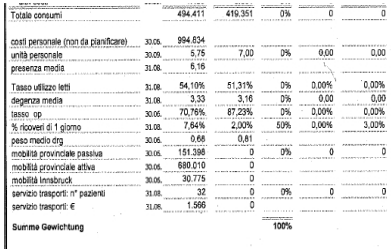
```
SELECT   {Paris, Berlin} ON ROWS
         {[Q1], [Q2].CHILDREN} ON COLUMNS
FROM     CubeSales
WHERE    (MEASURES.SalesAmount,
          Time.[2014],
          Product.Product)
```



OLAP et reporting

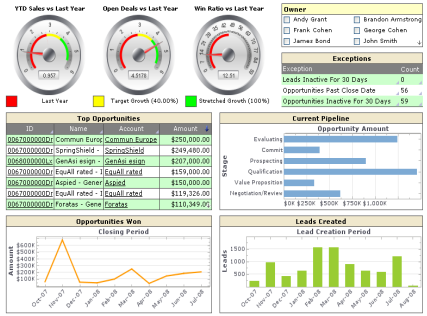
- Pour des utilisateurs qui ont besoin d'un **accès régulier à des informations d'une manière presque statique**
  Ex: les hôpitaux doivent envoyer des rapports mensuels à des agences nationales
- Un rapport est défini par une requête (plusieurs requêtes) et une mise en page (diagrammes, histogrammes, etc)
- Les rapports peuvent être exécutés automatiquement ou manuellement





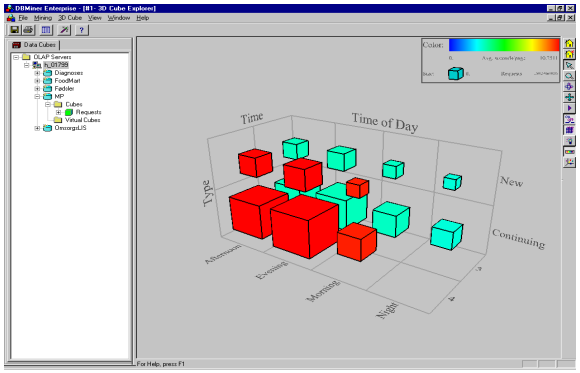
OLAP et tableaux de bords (Dashboards)

- 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



OLAP et visualisation de données

- 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**



5 – Domaines d'application et exemples d'entrepôts de données

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

Domaine bancaire : un des premiers utilisateurs des ED

- Pour une banque, il est important de pouvoir **regrouper les informations relatives à un client** afin de répondre à ses demandes de crédit par exemple
- **Des mailing ciblés doivent aussi être rapidement élaborés** à partir de toutes les informations disponibles sur un client lors de la commercialisation d'un nouveau produit
- **L'utilisation de cartes de crédit nécessite des contrôles à posteriori**, par exemple pour la recherche de fraudes : la mémorisation des mouvements peut rendre de grands services
- **Les échanges d'actions et de conseils de courtages sont facilités** par une mémorisation de l'histoire et une exploitation par des outils décisionnels avancés par exemple pour déterminer des tendances de marchés

Domaine de la grande distribution : fortement demandeur d'ED

- Intéressant de **regrouper les informations de ventes** pour déterminer les **produits à succès**, mieux suivre les **modes**, détecter les **habitudes d'achats**, les **préférences des clients** par **secteur géographique**
- La **fouille de données** (Data Mining) a permis de développer des techniques sophistiquées d'exploitation de données qui aident à **mettre en évidence les règles de consommation**
- **Explorer le panier de la ménagère** est devenu un exercice d'école : il s'agit de trouver à partir de l'enregistrement des transactions quelles sont les habitudes d'achats, plus précisément quels sont les produits achetés en même temps

Apports constatés dans la grande distribution :

- **augmentation des ventes grâce à un meilleur marketing**
- **amélioration des taux de rotation de stocks**
- **élimination des produits obsolètes**
- **réduction des rabais, remises, ristournes**
- **meilleure négociation des achats**

Domaine très concurrentiel des télécommunications : utilise beaucoup les ED

- **Grande masse de données** concernant les abonnés et les appels est **enregistrée**
- **Plusieurs mois de description détaillée des appels** comprenant, pour chaque appel appelant, appelé, heure et durée sont disponibles chez les opérateurs

En respectant les lois de sécurité et liberté, que peut-on faire de telles données ?

Couplées ou non avec des informations comptables, l'exploitation de ces données regroupées en ED par des techniques d'analyse et d'exploration permet :

- D'**analyser le trafic,**
- De **mieux cerner les besoins des clients**,
- **De classer les clients** par catégories,
- **De comprendre pourquoi certains changent d'opérateurs** et mieux répondre à leurs besoins

Domaines de l'assurance et de la pharmacie : très friands de techniques décisionnelles

- L'exercice de base de **l'assureur** est de **déterminer le facteur de risque d'un assuré**
- Celui d'un **producteur pharmaceutique** est de **détecter l'impact d'un médicament**
- Plus généralement, le **suivi des informations relatives à la liaison produit-client** sur un ED est souvent synonyme de **gains importants** : meilleure connaissance des produits, détection des défauts, meilleure connaissance des clients, détection de rejets, ciblage du marketing, etc
- Le couplage aux **technologies du Web** ouvre aussi des horizons nouveaux pour le suivi des produits, des clients, des concurrents : notion émergente de « **Data Webhouse** »

Exemple dans la grande distribution (1)

Exemple du groupe Casino :

Projet :

- un des premiers entrepôts en France
- plusieurs millions de dollars économisés en s'apercevant que les stocks de coca-cola faisaient souvent défaut...
- 1994 : 80 Go et 50 utilisateurs
- 2002 : + de 10 To, 1500 utilisateurs, 25000 requêtes/jour

Solution : Teradata

Exemple du groupe Walmart :

Projet :

- le plus gros entrepôt de données du monde, en 2006 : 0.5 Po de données
- distributeurs, magasins, clients (> 108), produits (> 109)...
- un des plus secret également...

Solution : Teradata

Wal-Mart, for example, discovered that people who buy Pampers often buy beer, so they moved Pampers and beer close together. The result was that sales of both increased (Computer Business Review, October 1996).

Exemple dans la grande distribution (2)

Exemple du groupe Camaieu:

Projet :

- plusieurs systèmes de production (magasin, logistique, comptable, etc.)

Solution :

- ***1996*** : agrégés dans un entrepôt de données, via **l'ETL Sunopsis**
- base **Oracle** découpée en référentiels métier (datamarts achat, marketing...)
- consultation des **datamarts** via le système de reporting de **Business Objects**
- ***2003*** : ajout d'un **cube OLAP** intégré à la base relationnelle Oracle9i :
  - meilleure ergonomie,
  - permet des requêtes complexes avec prise en compte de plusieurs niveaux au sein de la BD (types d'articles, collections, produits, zones géographiques, …)
  - base de composants Java (BI Beans) livrée par l'éditeur au sein de son environnement de développement (JDeveloper).

Exemple dans les télécommunications

Exemple de France Télécom :

Le projet :

- 12 BD sources
- récupération des données : 1,5 année
- données régionales et nationales
- parfois chez des prestataires de services
- parfois au prix d'un intense lobbying
- en 2003 : environ 5 années de travail

Solution :

- entreposage : SQL server
- DW de 3 bimestres, vidé périodiquement
- 1,2 million d'individus
- 1 fait = 1 client
- 250 colonnes
- intégration faite à la main périodiquement

Exploitation : progiciel de DM développé spécifiquement