Business Intelligence - Modélisation des entrepôts de données

DR. Sofiane AOUAG
Université de Batna II
Faculté des Mathématique et de l'Informatique
Département Informatique
Cours L3 ISIL

Plan du cours

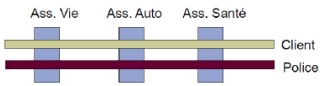
- Qu'est ce que signifie un entrepôt de données
- Métaphore de cube de données
- Modèles en étoile
- Modèles en flocons de neige
- Représentation des données
- Les différentes approches (ROLAP, MOLAP, HOLAP)
- Alimentation des entrepôts de données
- Processus d'ETL

Qu'est ce que signifie un entrepôt de donnés ?

Définition : Un Entrepôt de données est une collection de données orientées sujet , intégrées, non volatile et historisées, organisées pour le support d'un processus d'aide à la décision

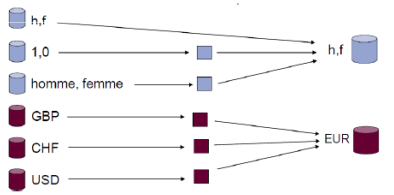
Orientées sujet

- Agrégation des informations de différents métiers
- Pas de prise en compte de l'organisation fonctionnelle des données



Lydie Soler, AgroParisTech/INRA

Intégrées

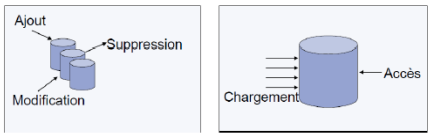


Lydie Soler, AgroParisTech/INRA

Qu'est ce que signifie un entrepôt de donnés ?

Définition : Un Entrepôt de données est une collection de données orientées sujet , intégrées, non volatile et historisées, organisées pour le support d'un processus d'aide à la décision

Non volatiles

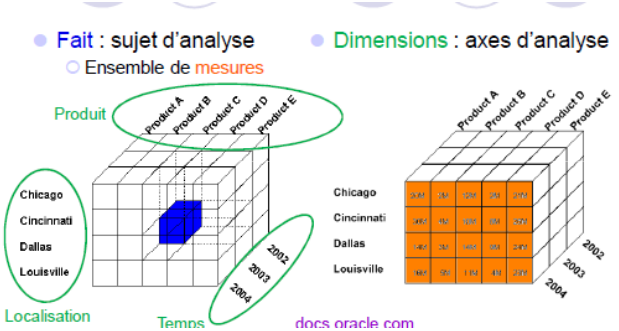


Lydie Soler, AgroParisTech/INRA

Historisées

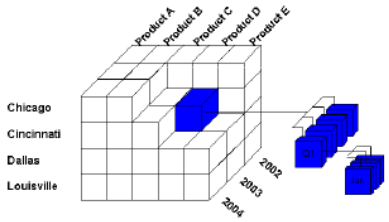
Lydie Soler, AgroParisTech/INRA

Métaphore de cube de données ?



Métaphore de cube de données ?

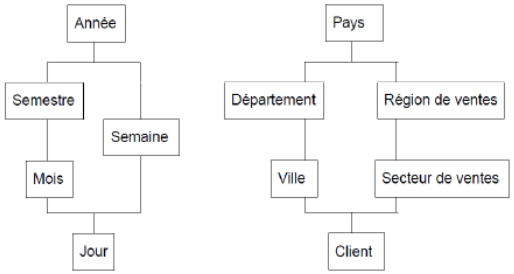
Hiérarchie de dimension



docs.oracle.com

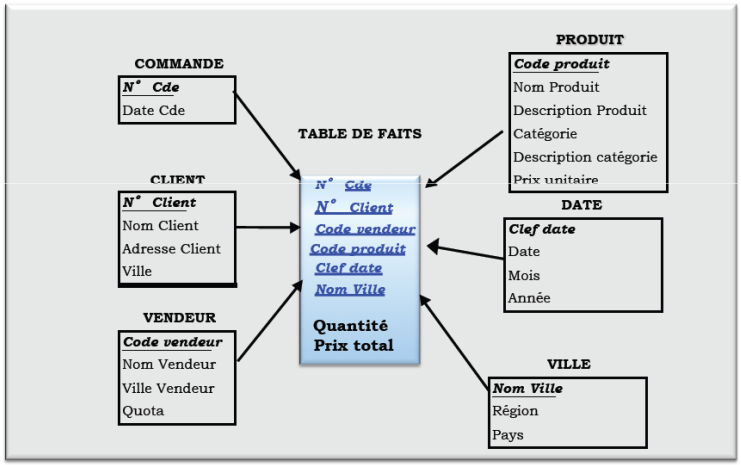
Métaphore de cube de données ?

Hiérarchies multiples



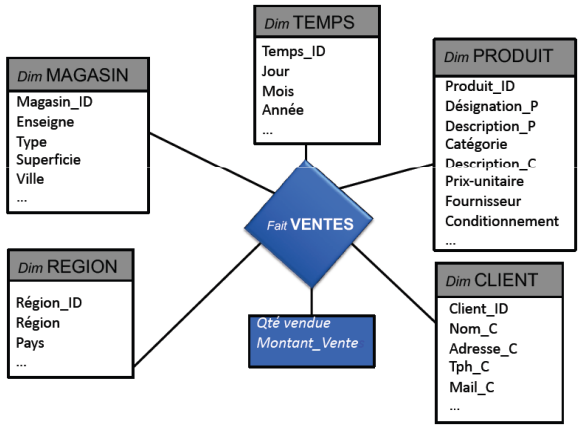
Elsa Nègre, Université Paris Dauphine

Modèle en étoile



Modèle en étoile

- Schéma en étoile



Modèle en étoile

Un modèle en étoile est constitué de :
Une **table de faits : identifiants des tables de dimension ; une ou** plusieurs mesures .
-Plusieurs **tables de dimension : descripteurs des dimensions.**
- Une **granularité définie par les identifiants dans la table des faits.**

Avantages :

-Facilité de navigation
- Performances : nombre de jointures limité ; gestion des données creuses.
-Gestion des agrégats
- Fiabilité des résultats

Inconvénients :

-Toutes les dimensions ne concernent pas les mesures
- Redondances dans les dimensions
- Alimentation complexe.

Modèle en étoile

Propriétés des mesures

Additivité : somme sur toutes les dimensions

Exemple : CA ; Quantité vendue, ...

Semi-additivité : somme sur certaines dimensions :

Exemple : nbre de contacts clients, Etats des stocks...

Non-additivité : pas de somme, recalculer

Exemple: encours moyen fin de mois,
plus grand CA pour l'ensemble des magasins

Modèle en étoile

Exemples de modèles

Dans la grande distribution :

Quelques tables de faits : détaillées et volumineuses
Tables de dimensions :
Classiques : produits, fournisseurs, temps, établissements (structure géographique, fonctionnelle)…
Stratégiques : Clients, Promotions, ….

Rq : Obtenir le plus d'enregistrements possibles.

Dans le secteur des banques :

Tables de faits : nombreuses, dédiées à chaque produit, peu détaillées et peu volumineuses.
Tables de dimensions :
Classiques : produits, temps, établissement (structure géographique, fonctionnelle), …
Stratégiques : Clients, ….

Rq : Obtenir le plus de données (champs) possibles.

Modèle en flacon de neige

Le modèle de l'ED doit être simple à comprendre.
On peut augmenter sa lisibilité en regroupant certaines dimensions.
On définit ainsi des hiérarchies.

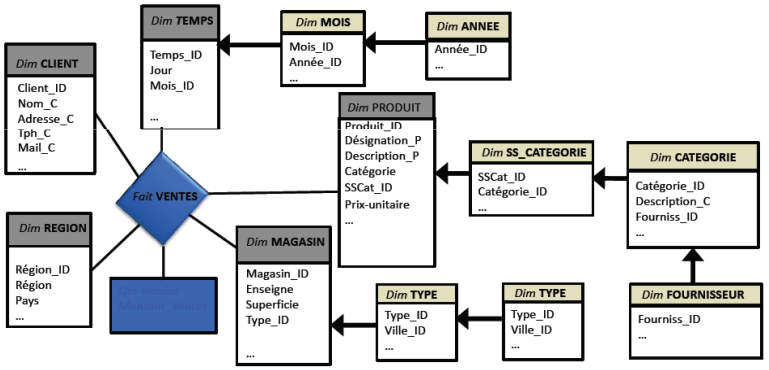
Celles-ci peuvent être géographiques ou organisationnelles.

Exemple : Commune, Département, Région, Pays, Continent

Client	Commune	Département	Region	Pays	Continent
Pepone	Lyon 1°	Rhône	Rhône-Alpes	France	Europe
Testut	Lyon 2°	Rhône	Rhône-Alpes	France	Europe
Soinin	Lyon 3°	Rhône	Rhône-Alpes	France	Europe
Vepont	Paris 1°	Paris	Ile-de-France	France	Europe
Martin	Paris 2°	Paris	Ile-de-France	France	Europe
Elvert	Versailles	Yvelines	Ile-de-France	France	Europe

Modèle en flocon de neige

- Après normalisation
- Avant normalisation



Modèle en flacon de neige

Modèle en flocons de neige =

Modèle en étoile + normalisation des dimension

Lorsque les tables sont trop volumineuses

Avantages :

- réduction du volume,
- permettre des analyse par pallier (drill down) sur la dimension hiérarchisée.

Inconvénients :

- navigation difficile ;
- nombreuses jointures.

Plan du cours

- Qu'est ce que signifie un entrepôt de données
- Métaphore de cube de données
- Modèles en étoile
- Modèles en flocons de neige
- Représentation des données
- Les différentes approches (ROLAP, MOLAP, HOLAP)
- Alimentation des entrepôts de données
- Processus d'ETL

Représentation des données

Les données sont perçues à travers plusieurs dimensions. Elles sont qualifiées de **multidimensionnelles,** indépendamment de leur support (tables relationnelles ou tableaux multidimensionnels)

Produit	Region	Ventes
Clous	Est	50
Clous	Ouest	60
Clous	Centre	100
Vis	Est	40
Vis	Ouest	70
Vis	Centre	80
Boulons	Est	90
Boulons	Ouest	120
Boulons	Centre	140
Nettoyeurs	Est	20
Nettoyeurs	Ouest	10
Nettoyeurs	Centre	30

Représentation des données dans une table relationnelle

	Est	Ouest	Centre
Clous	50	60	100
Vis	40	70	80
Boulons	90	120	140
Nettoyeurs	20	10	30

Représentation des données dans un tableau multidimensionnel

Représentation des données

les requêtes décisionnelles sont de type :

" Quelle est le total des ventes dans la région Est ?"

On peut calculer divers totaux.

- **Tables relationnelles** : on peut traiter quelques centaines de tuples par seconde.
- **Tableau multidimensionnel** : on peut rajouter en lignes et en colonnes plus de 10 000 valeurs par seconde.

Pour accélérer les temps de réponses, il est préférable de pré-calculer des sous totaux.

Représentation des données

Produit	Region	Ventes
Clous	Est	50
Clous	Ouest	60
Clous	Centre	100
Clous	**Total**	**210**
Vis	Est	40
Vis	Ouest	70
Vis	Centre	80
Vis	**Total**	**190**
Boulons	Est	90
Boulons	Ouest	120
Boulons	Centre	140
Boulons	**Total**	**350**
Nettoyeurs	Est	20
Nettoyeurs	Ouest	10
Nettoyeurs	Centre	30
Nettoyeurs	**Total**	**60**
Total	**Est**	**200**
Total	**Ouest**	**260**
Total	**Centre**	**350**
Total	**Total**	**810**

Pour le calcul de ces totaux : 28 accès en lecture et 8 accès en écriture.

Un SGBDR lit 200 enregist/s et en écrit environ 20/s.

	Est	Ouest	Centre	Total
Clous	50	60	100	210
Vis	40	70	80	190
Boulons	90	120	140	350
Nettoyeurs	20	10	30	60
Total	200	260	350	810

OLAP consolide entre 20 et 30000 cellules/s

Représentation des données

La valeur **ALL** remplace une colonne ou une valeur d'agrégats.

Magasin	Date	Rayon	CA Ventes
Mag1	1/2/96	010	3500
Mag1	6/2/96	010	2500
Mag1	10/2/96	010	2900
Mag1	ALL	010	8900
Mag2	…	…	…

Représentation des données

Soient N attributs concourant à la construction du cube, il y aura :
$C_1, C_2, \ldots, C_N$ les cardinalités des N attributs, tq : $C_1 = |\mathcal{D}a_1|$; $C_2 = |\mathcal{D}a_2|$; … ; $C_N = |\mathcal{D}a_N|$

Le cube aura : **$\prod(C_i + 1)$ enregistrements**

Dans la tables VENTES si on a **2 * 3 * 3 = 18 enregistrements**

dans le cube on aura **(2+1) * (3+1) * (3+1) = 48 enregistrements.**

Les différentes approches d'OLAP

L'approche relationnelle (ROLAP)

L'ensemble des données est stocké dans une BDR. Les données sont sous forme d'enregistrements (tuples).

VENTES (Magasin, Rayon, Date, CA Ventes, Nb Client)

```
Select Magasin, Date , Sum(CA Ventes)
From VENTES
Group By Magasin, Date
```

Nouveaux opérateurs d'agrégation : **cube** , **rollup**.

☞J.Gray, A. Bosworth, A. Leyman, H. Pirahesh, "*Data Cube : A relationnal Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total*", in Data Mining and Knowledge Discovery Journal, 1(1), 1997]

Les différentes approches d'OLAP

L'approche relationnelle (ROLAP)

L'union de plusieurs group-by donne naissance à un cube :

```
Select ALL, ALL, ALL, Sum(CA Ventes)
   From VENTES
   UNION Select Magasin, ALL, ALL, Sum(CA Ventes)
   From VENTES
   Group-By Magasin ;
   UNION Select Magasin, Date, ALL, Sum(CA Ventes)
   From VENTES
   Group-By Magasin, Date ;
   UNION Select Magasin, Date, Rayon, Sum(CA Ventes)
   From VENTES
   Group-By Magasin, Date, Rayon ;
```

L'opérateur cube est une généralisation N-dimensionnelle de fonctions d'agrégations simples . C'est un opérateur relationnel.

```
Select Magasin, date, Rayon, Sum(CA Ventes)
From VENTES
Group-By Cube Magasin, Date, Rayon ;
```


Les différentes approches d'OLAP

L'approche multidimensionnelle (MOLAP)

Il s'agit de stocker les données dans des tableaux multidimensionnels. Ces tableaux peuvent être éparses.

On y stocke dans les **cellules** les mesures (valeurs à observer), les données représentant les dimensions sont les **coordonnées** de ces valeurs :

$$f = (d_1, d_2, \ldots, d_n, m_1, m_2, \ldots, m_p)$$

[Zhao Yihong, Deshpande Prasad M., Naughton Jeffrey F., «*An Array-Based Algorithm for Simultaneous Multidimensional Aggregates*», in SIGMOD Record n° 26, Vol 2, 1997.]

Les différentes approches d'OLAP

L'approche multidimensionnelle (MOLAP)

BD éparse

- Plus on a de dimensions plus on a de cellules.
  Seulement une partie des produits peut être vendue
  ⇨ des cellules sans valeur : **données éparse.**

Exemple :
On dispose de 100 000 données (eq. tuples)
4 dimensions ayant une cardinalité de 30 modalités chacune:
30 * 30 * 30 * 30 = 810 000 cellules
(dont 710 000 vides : **12,3%** seulement sont pleines)

- Une BD est considérée comme éparse si elle a moins de **40%** de ses cellules " peuplées ".
- Techniques de compression des données

Les différentes approches d'OLAP

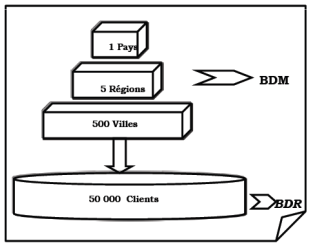
L'approche hybride (HOLAP)

Approche relationnelle : **30%** du temps est consacré aux I/O.
Approche multidimensionnelle : **20%.** (70% calculs et 10% décompression)

La 3° voie préconisée consiste à utiliser les tables comme structure permanente de stockage des données et les tableaux comme structure alors des requêtes.

La démarche consisterait en 3 étapes:

1. Charger les données d'une table vers un tableau.
2. Calculer le cube de ce tableau selon les méthodes initialement présentées.
3. Stocker les résultats (données agrégées) dans un table.



Les différentes approches d'OLAP

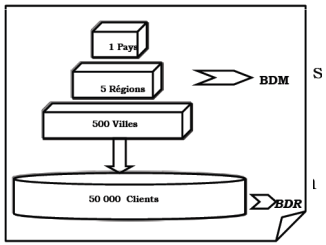
L'approche hybride (HOLAP)

Approche relationnelle : **30%** du temps est consacré aux I/O.
Approche multidimensionnelle : **20%.** (70% calculs et 10% décompression)

La 3° voie préconisée consiste à utiliser les tables comme structure permanente de stockage des données et les tableaux comme structure alors des requêtes.

La démarche consisterait en 3 étapes:

1. Charger les données d'une table vers un tableau.
2. Calculer le cube de ce tableau selon les méthodes initialement présentées.
3. Stocker les résultats (données agrégées) dans un table.



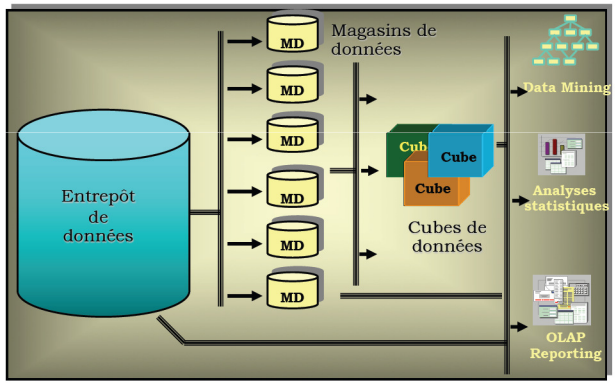
Alimentation des entrepôts de données

Les magasins de données (data marts)

- Simples "**magasins de données**" (***Data Marts***), on y stockera des données portant sur **une seule** des activités de l'entreprise.
- Ceux sont en quelque sorte des vues métier.
- Exemple Data mart Comptabilité, Data mart RH,.....
- Ces mini ED peuvent alors être considérés comme des espaces d'analyse, du fait que les données sont bien moins nombreuses et surtout qu'elles sont thématiques.
- Ils peuvent également servir de bases de construction à des cubes de données.

Alimentation des entrepôts de données

Entrepôts, Magasins et Cubes de données



Plan du cours

- Qu'est ce que signifie un entrepôt de données
- Métaphore de cube de données
- Modèles en étoile
- Modèles en flocons de neige
- Représentation des données
- Les différentes approches (ROLAP, MOLAP, HOLAP)
- Alimentation des entrepôts de données
- Processus d'ETL

Processus d'ETL

Processus d'ETL

(Extracting – Transforming – Loading)

Processus d'ETL

Alimenter un ED…

☑ Le principe de l'entreposage des données est de rassembler de multiples données sources qui souvent sont hétérogènes en les rendant *homogènes* afin de les analyser.

☑ Ce travail d'homogénéisation nécessite des règles précises servant de **dictionnaire** (ou de **référentiel**) et qui seront mémorisées sous forme de **métadonnées** (information sur les données).

☑ Ces règles permettent d'assurer des tâches d'administration et de gestion des donnés entreposées.

Processus d'ETL

Alimenter un ED…

- ☑ Le principe de l'entreposage des données est de rassembler de multiples données sources qui souvent sont hétérogènes en les rendant *homogènes* afin de les analyser.

- ☑ Ce travail d'homogénéisation nécessite des règles précises servant de **dictionnaire** (ou de **référentiel**) et qui seront mémorisées sous forme de **métadonnées** (information sur les données).

- ☑ Ces règles permettent d'assurer des tâches d'administration et de gestion des donnés entreposées.

Processus d'ETL

- Le dictionnaire (ou référentiel) de données est constitué de l'ensemble des métadonnées.

- Il renferme des informations sur toutes les données de l'ED.

- Il renferme également des informations sur chaque étape lors de la construction de l'ED ; sur le passage d'un niveau de données à un autre lors de l'exploitation de l'ED .

Le rôle des métadonnées est de permettre :

- **La définition des données**
- **La fabrication des données**
- **Le stockage des données**
- **L'accès aux données**
- **La présentation des données.**

Processus d'ETL

Processus d'ETL

L'alimentation d'un ED est un processus qui s'effectue en plusieurs étapes :

- **Sélection des données sources**
- **Extraction des données**
- **Transformation**
- **Chargement**

Processus d'ETL

- Sélection des données sources

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

Toutes les données sources ne sont forcément pas utiles.

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

Les données sélectionnées seront réorganisées pour servir à la fabrication

La synthèse de ces données sources a pour but de les enrichir.

Processus d'ETL

La sélection des données utiles à partir des BD de production n'est pas simple à faire .

Les données sont :

- **hétérogènes** (**différents SGBD et différentes méthodes d'accès**)
- **diffuses** (**différents environnements matériels et différents réseaux interconnectés ou non**)
- **complexes** (**différents modèles logiques et physiques principalement orientés vers les traitements transactionnels**)

La définition de la granularité dépend du niveau de raffinement de l'information qu'on veut obtenir

Processus d'ETL

Extraction des données

- L'extraction peut se faire à travers un outil d'alimentation qui doit travailler de façon native avec les SGBD qui gèrent les données sources.
- Ou alors créer des programmes extracteurs. L'inconvénient de cette approche est le risque de faire des extractions erronées, incomplètes et qui peuvent biaiser l'ED.
- Il faut gérer les anomalies en les traitant et en gardant une trace

Processus d'ETL

- L'extraction doit se faire conformément aux règles précises du référentiel.
- Elle ne doit non plus perturber les activités de production.
- Il faut faire attention aux données cycliques. Celles qu'on doit calculer à chaque période, pour pouvoir les prendre en considération.
- L'extraction peut se faire en interne selon l'horloge interne ou par un planificateur ou par la détection d'une donnée cible (de l'ED) ; ou en externe par des planificateurs externes.
- Les données extraites doivent être marquées par "*horodatage*" afin qu'elles puissent être pistées.

Processus d'ETL

Transformations

C'est une suite d'opérations qui a pour but de rendre les données cibles **homogènes** et puissent être traitées de façon **cohérente**.

Exemple

Donnés sources données cibles	**Donnés sources données cibles**
Appli 1 : male, femelle	Appli 1 : $150,000
Appli 2 : 1, 0	Appli 2 : 16 000 CHF
Appli 3 : Masculin, féminin	Appli 3 : 200.000£

Processus d'ETL

- L'ensemble des données sources, après nettoyage ou transformation d'après des règles précises ou par application de programmes (*pour un contrôle de vraisemblance par des méthodes statistiques*), seront restructurées et converties dans un **format cible**.

- Il faut synchroniser les données pour que les valeurs agrégées obtenues soient cohérentes. Avant de passer à la phase de chargement.

Processus d'ETL

- ☑ C'est l'opération qui consiste à charger les données nettoyées et préparées dans le DW.

- ☑ C'est une opération qui risque d'être assez longue. Il faut mettre en place des stratégies pour assurer de bonnes conditions à sa réalisation et définir la politique de rafraîchissement.

- ☑ C'est une phase plutôt mécanique et la moins complexe.

Références

Omar Boussaid 2019 – cours Introduction aux Systèmes d'Information Décisionnels
Jérome Damont 2015 –cours Introduction aux entreprots de données
Jean-Marie Gouarné, Le Projet décisionnel - Enjeux, Modèles, Architectures du Data Warehouse [archive], Eyrolles, 1997, (ISBN 978-2-212-05012-7) ⍰ Alain Garnier, L'Information non structurée dans l'entreprise - Usages et Outils, Hermes - Lavoisier, 2007, (ISBN 978-2-7462-1605-1) ⍰ R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, Le Data Warehouse : Guide de conduite de projet, Eyrolles, 2005, (ISBN 978-2-212-11600-7) ⍰ Alain Fernandez, Les Nouveaux Tableaux de bord des managers, Le Projet Business Intelligence clés en main, Eyrolles, 6e édition, 2013. (ISBN 978-2-212-55647-6) présentationéditeur [archive] ⍰ Roland et Patrick Mosimann, Meg Dussault, The Performance Manager Faire de la performance le quotidien de chacun [archive], CognosPress, 2007, (ISBN 978-0-9730124-4-6) ⍰ James Taylor, Decision Management System [archive], IBM Press, Pearson Education