

Entrepôts de données Data Warehouse

Collecter, intégrer, manipuler

Dr. Aymen GAMMOUDI

Contenu du cours (1)

- **Définition et architectures fonctionnelles d'un ED**
 - Data Warehouses versus Data Marts
 - Différentes architectures
- **Modélisation d'un ED**
 - Approche multidimensionnelle
 - Cube de données
- **Implémentation d'un ED**
 - ROLAP, MOLAP, HOLAP
 - Schéma en étoile ou flocon

Contenu du cours (2)

- **Alimentation d'un ED**
 - Processus ETL (Extraction, Transformation, Load)
 - *Extraction des données*
 - *Nettoyage et transformation des données*
 - *Chargement des données*
- **Exploitation d'un ED**
 - Reporting et dashboards
 - Analyse en ligne OLAP
 - Langage MDX

Ressources

- **Quelques ressources utilisées**
 - Cours de Ludovic Denoyer (LIP6 - Paris)
 - Cours de Bernard Espinasse (EPU de Marseille)
 - Cours de Jérôme Darmont (ERIC – Lyon 2)
 - Livre de Wilfried Grossmann and Stefanie Rinderle-Ma
« Fundamentals of Business Intelligence » (Springer)

Entrepôts de données Data Warehouse

Définition et architectures
fonctionnelles

Dr. Aymen GAMMOUDI

Data Warehouse (1)

- **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 »

Data Warehouse (2)

- **Caractéristiques principales**
 - **Données thématiques / orientées sujets** : données pertinentes pour un sujet ou thème et nécessaires aux besoins d'analyse
 - **Intégrées** : données résultant de l'intégration de données provenant de différentes sources pouvant être hétérogènes
 - **Historisées** : données représentent l'activité d'une entreprise durant une certaine période (plusieurs années)
 - **Non-volatiles** : données essentiellement utilisées en interrogation (consultation), ne pouvant pas être modifiées

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

- **Avant l'entreposage des données**
 - Avant d'être chargées, les données sélectionnées doivent être :
 - **Extraites de sources**
 - **Internes** (BD opérationnelles) ou
 - **Externes** (données notamment issus du Web)
 - **Netoyées**
 - Afin d'éliminer des erreurs
 - **Intégrées**
 - Afin de réconcilier les sémantiques associées aux différentes sources

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

- **Après – Quelle exploitation des données de l'ED ?**
 - A partir des données d'un ED diverses **analyses** peuvent être faites, notamment par :
 - Des techniques **OLAP** (On-Line Analytical Processing)
 - Des techniques de **fouille de données** (Data Mining)
 - De techniques de **visualisation de données** multi-dimensionnelles
 - Notons que les informations et connaissances obtenues par exploitation d'un entrepôt ont généralement pour objectif
 - *Un impact direct sur les résultats d'une entreprise ou d'un établissement (augmentation des ventes par un marketing plus ciblé, amélioration de la rotation des stocks, amélioration des résultats des étudiants, etc.)*

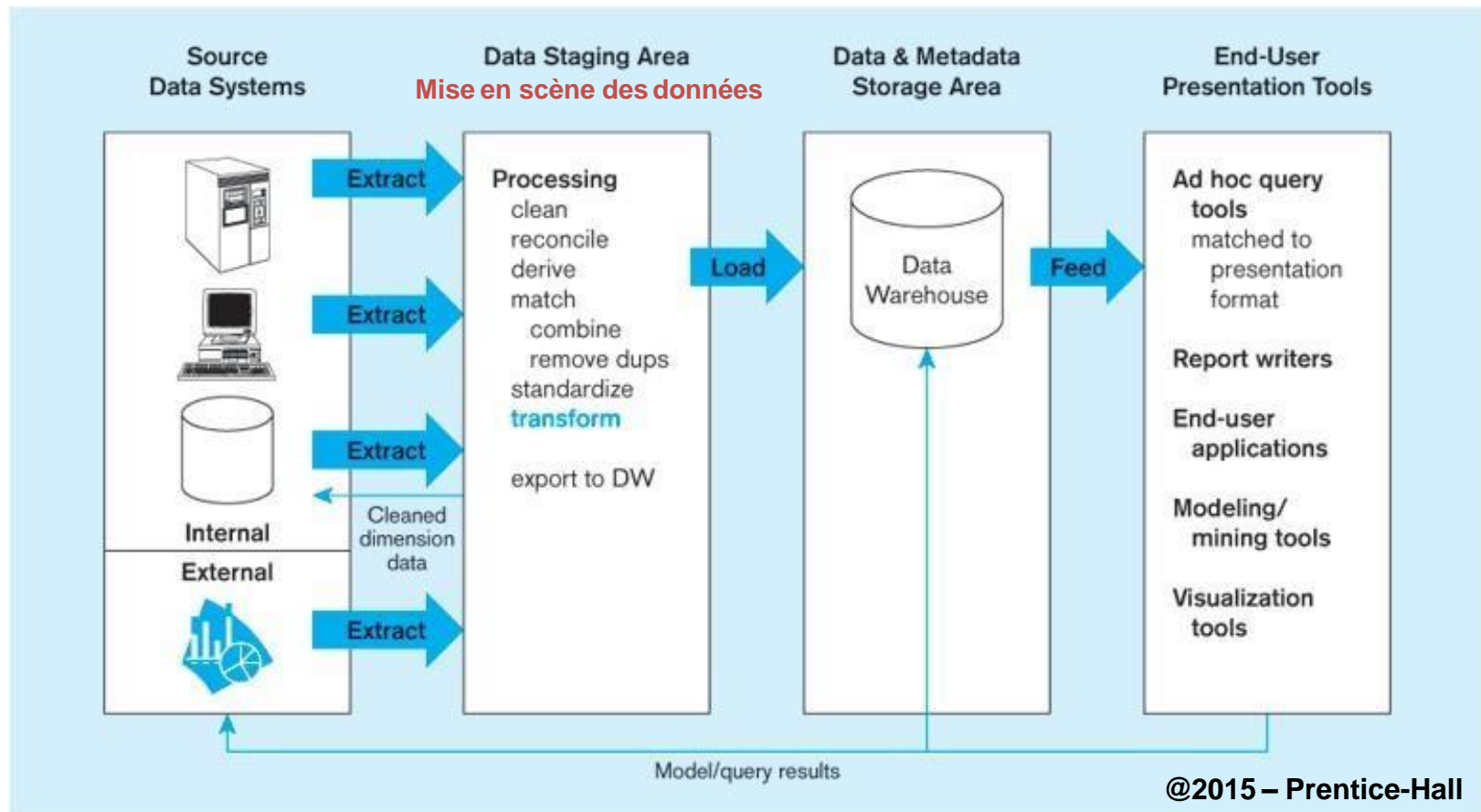
Entrepôts vs Magasins de données (1)

- **L'entrepôt de données (Data Warehouse)**
 - Collecte **l'ensemble de l'information** utile aux décideurs à partir des sources de données (BD opérationnelle, BD externes, Web ...)
 - **Centralise** l'information décisionnelle en assurant **l'intégration** des données extraites et leur **pérennité** dans le temps
- **Les magasins de données (Data Marts)**
 - Sont (davantage) **orientés sujet** en extrayant pour chaque data mart une partie de l'information décisionnelle de l'entrepôt à partir **d'une partie des données utiles**
 - Sont conçus pour une classe d'utilisateurs ou pour un besoin d'analyse spécifique

Entrepôts vs Magasins de données (2)

- **L'entrepôt de données (Data Warehouse)**
 - Nécessitent de **puissantes machines** pour gérer de très grandes bases de données contenant des données de détail historisées
 - L'organisation des données privilégie une **gestion efficace des données et de leur historisation**
- **Les magasins de données (Data Marts)**
 - Sont de petits entrepôts **nécessitant une infrastructure plus légère** et peuvent être mis en œuvre plus rapidement
 - L'organisation des données privilégie **les traitements décisionnels**

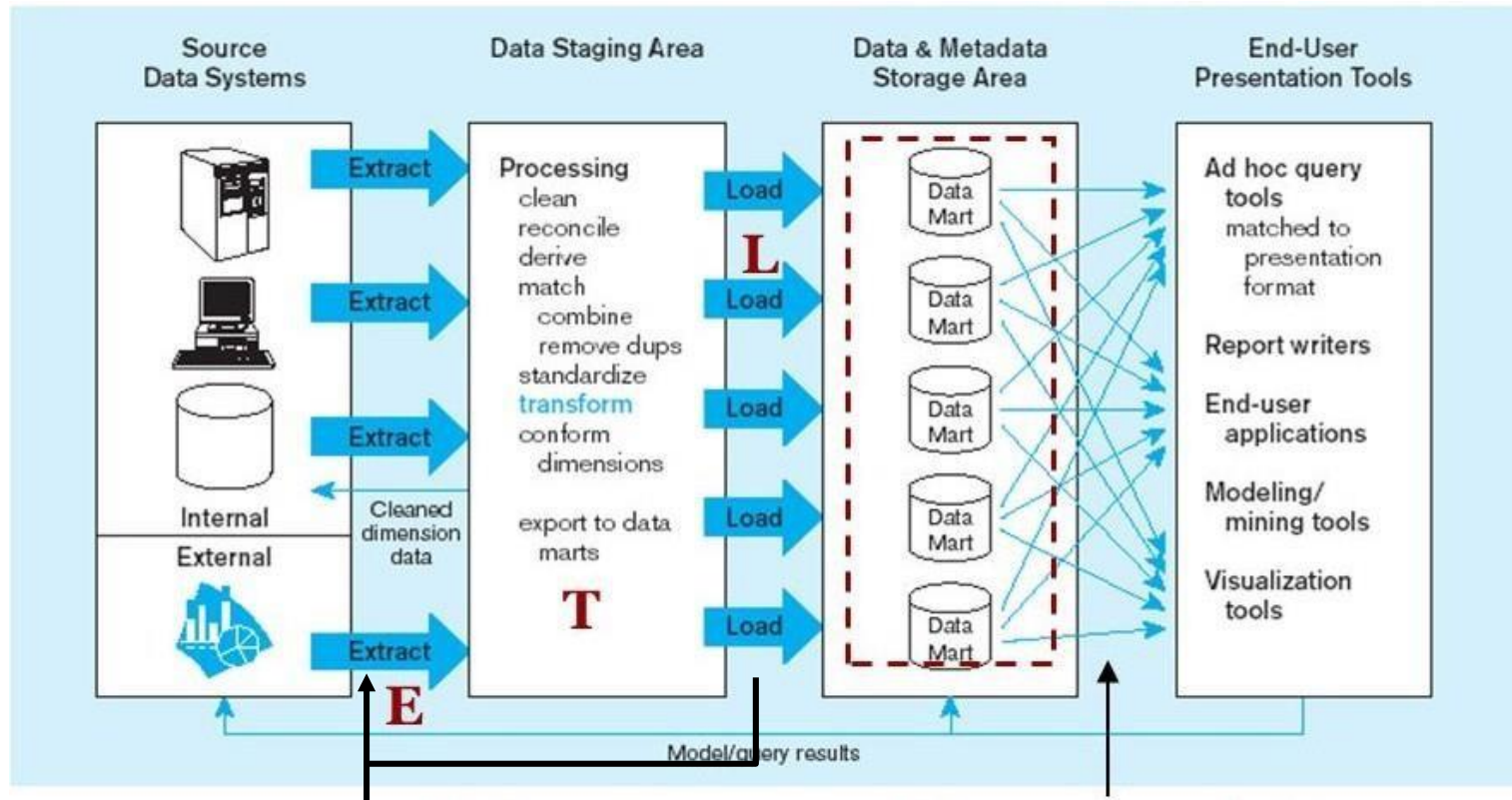
Différentes architectures (1)



En général, l'extraction des données est **uniquement périodique**, et les données dans l'entrepôts **ne représentent pas l'état courant de l'entreprise**

Différentes architectures (2)

Avec des Data Marts indépendants

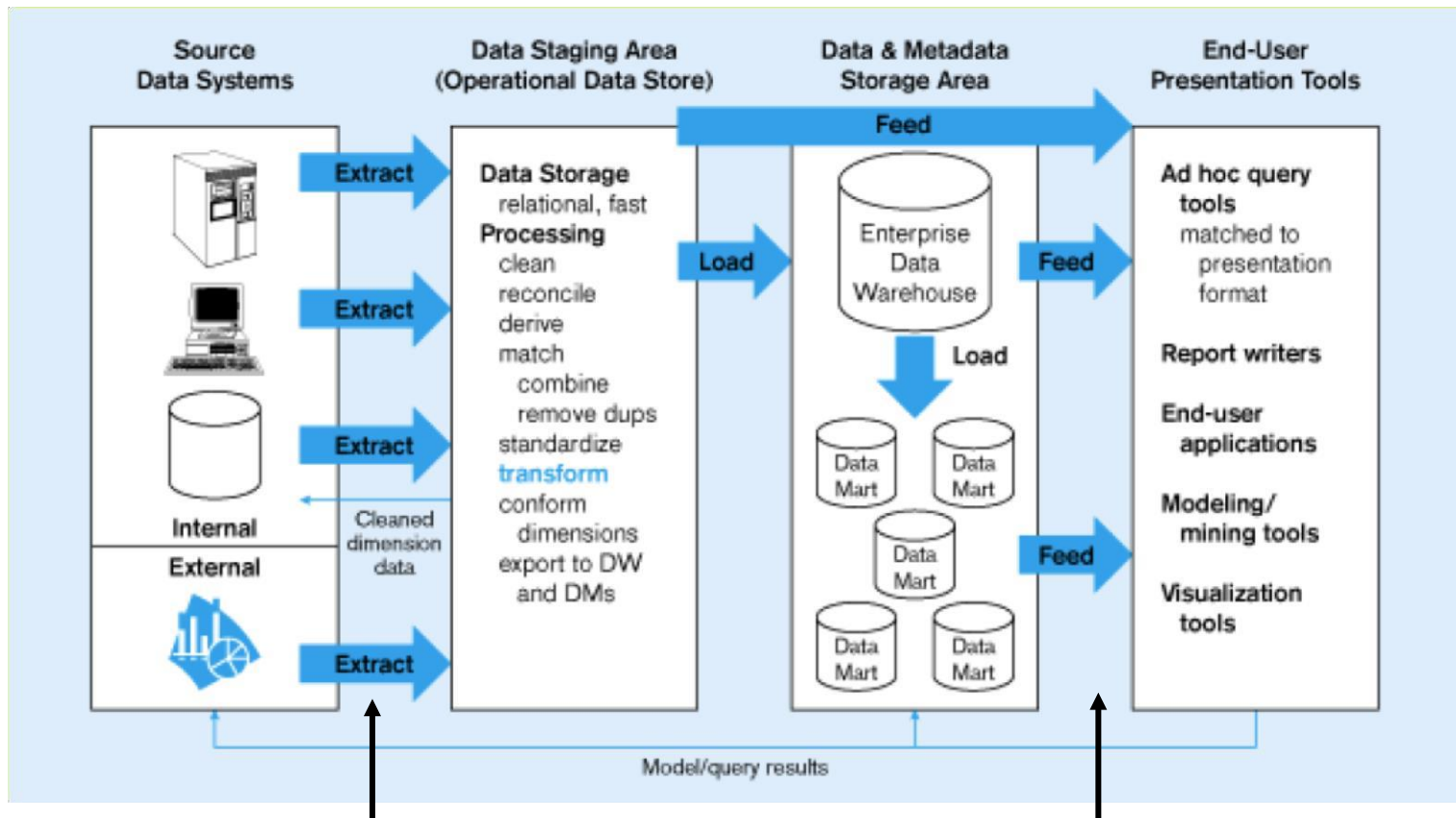


Implique un processus
ETL par Data Mart

Peut rendre complexe
l'accès aux données
par les décideurs

Différentes architectures (3)

Avec des Data Marts dépendants



Nécessite un seul processus ETL

Facilite l'accès aux données par les décideurs

Modélisation d'un entrepôts de données

Modélisation multidimensionnelle,
cubes de données

Dr. Aymen GAMMOUDI

Modélisation multidimensionnelle (1)

- **Objectif**
 - Proposer une modélisation des données **proche de la perception qu'en a un analyste**
 - Basée sur une **vision multidimensionnelle** des données
- **Principe de base**
 - Un **sujet** analysé (**ventes**) est considéré comme un point dans un espace à plusieurs **dimensions** (**quoi, quand, où**)
 - Ces dimensions offrent différentes perspectives d'analyse
 - *Comment les ventes se répartissent par catégorie de produit ?*
 - *Comment les ventes ont évolué dans le temps ?*
 - *Comment les ventes se répartissent dans les différentes régions ?*

Modélisation multidimensionnelle (2)

- Exemple : ventes en 2000 d'une entreprise de distribution

Catégorie de produit	Régions	Montant des ventes
Electroménager	Centre – Val de Loire	50
Electroménager	Nouvelle Aquitaine	30
Electroménager	Pays de La Loire	40
Alimentation	Centre – Val de Loire	10
Alimentation	Nouvelle Aquitaine	20
Alimentation	Pays de La Loire	30
Bricolage	Centre – Val de Loire	10
Bricolage	Nouvelle Aquitaine	40
Bricolage	Pays de La Loire	20

Deux dimensions d'analyse

Sujet étudié

Autre représentation

- Sous la forme d'un tableau croisé

Régions

	Centre – Val de Loire	Nouvelle Aquitaine	Pays de la Loire
Electroménager	50	30	40
Alimentation	10	20	30
Bricolage	10	40	20

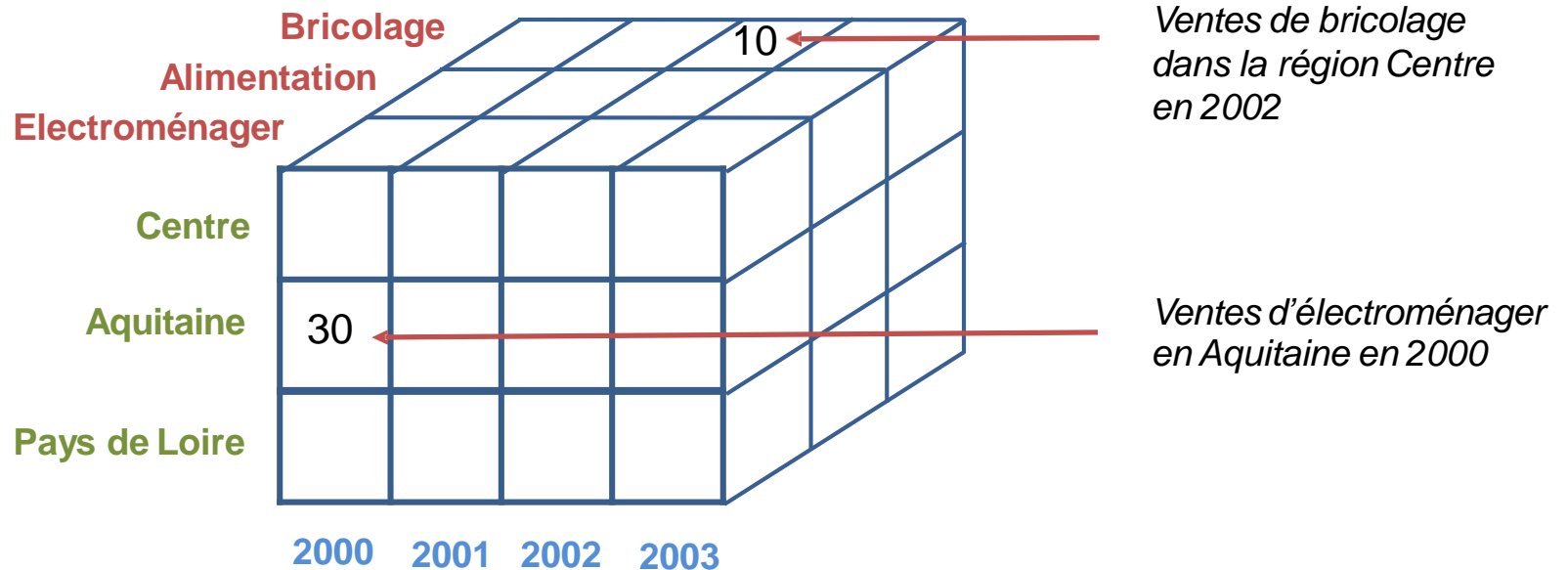
Catégorie de produits

Une vente

Cube de données

- **Poursuite de l'exemple**

- Avec le **temps** comme dimension d'analyse supplémentaire, ajout d'une troisième dimension d'analyse



Note : le contenu des cellules à l'intérieur du cube n'est pas visible

Autre représentation

- Sous la forme d'un tableau **avec imbrication**

		Centre – Val de Loire	Nouvelle Aquitaine	Pays de la Loire
2000	Electroménager	30	20	100
	Alimentation	10	20	30
	Bricolage	10	40	20
2001	Electroménager	20	10	30
	Alimentation	10	20	30
	Bricolage	10	40	20

Faits, mesures et dimensions

- **Notion de fait**

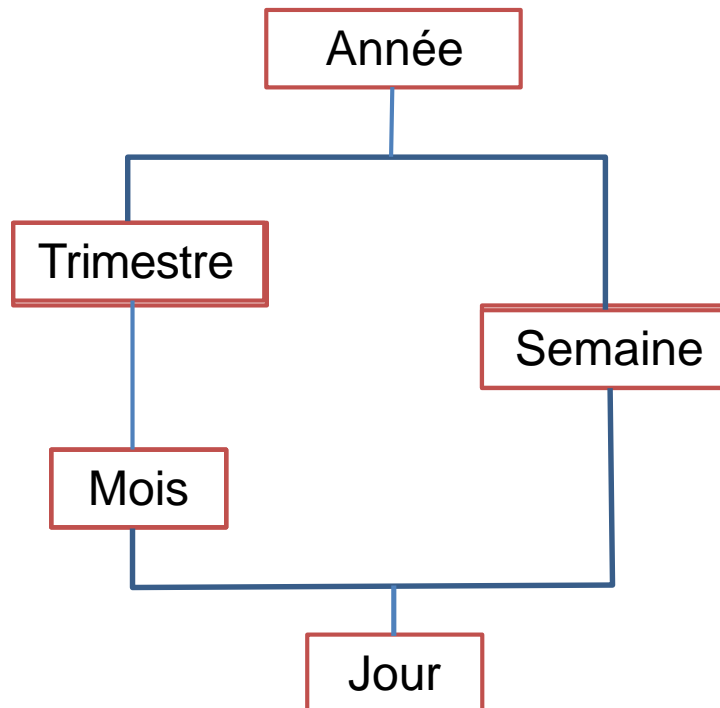
- Un **fait** modélise le sujet d'analyse (les ventes)
- Un fait est formé de **mesures** généralement numériques (montant des ventes, quantités de produits vendues, etc.)
- Les mesures peuvent généralement être **agrégées** (sum, avg, etc.)

- **Notion de dimension**

- Une **dimension** modélise un axe d'analyse (temps, lieu, catégorie, etc.)
- Une dimension est généralement organisée en une ou plusieurs **hiérarchies** correspondant à **différents niveaux de détail**
 - **Dimension temps** : H1 (jour → mois → trimestre → année), H2 (jour → semaine → année)
 - **Dimension lieu** : ville → département → région → pays
 - **Dimension catégorie** : produit → gamme → catégorie

Hiérarchies multiples

- **Exemple sur la dimension Temps**
 - Permet d'avoir les montants totaux des ventes à différents niveaux de détail / différents niveaux de granularité



Implémentation d'un entrepôts de données

Schémas en étoile ou flocons, ROLAP, MOLAP, HOLAP

Dr. Aymen GAMMOUDI

Stratégie d'implémentation

- **Trois stratégies au niveau physique**
 - Usage d'un SGBD Relationnel (systèmes **ROLAP**)
 - *Avantages* : faible coût de mise en œuvre
 - *Inconvénients* : performance (pour le calcul des jointures & agrégats)
 - Usage d'un SGBD Multidimensionnel (systèmes **MOLAP**)
 - *Avantages* : adapté aux analyses multidimensionnelles
 - *Inconvénients* : difficulté de mise en œuvre (systèmes propriétaires), problème **d'éparité** des cubes, etc.
 - Usage d'un SGBD Hybride (systèmes **HOLAP**)
 - *Pour tirer profit des avantages des technologies ROLAP et MOLAP :*
 - Un système ROLAP pour stocker, gérer les données détaillées **ET**
 - Un système MOLAP pour stocker, gérer les données agrégées

Schéma d'un entrepôt

- **Niveau logique** : dans un système ROLAP
 - Trois grands types de schémas
 - Schémas *en étoiles* (*star schema*)
 - Schémas *en flocon* (*snowflake schema*)
 - Schémas *en constellation* (*fact constellation*)
 - Le schéma en étoile est souvent celui qui est **aussi utilisé au niveau physique**

Schéma en étoile

- **Caractéristiques**

- Une table centrale, i.e. la **table de faits**, en général de taille très volumineuse
- Des tables périphériques, i.e. **les tables de dimensions**, de tailles peu importantes et **non normalisées**

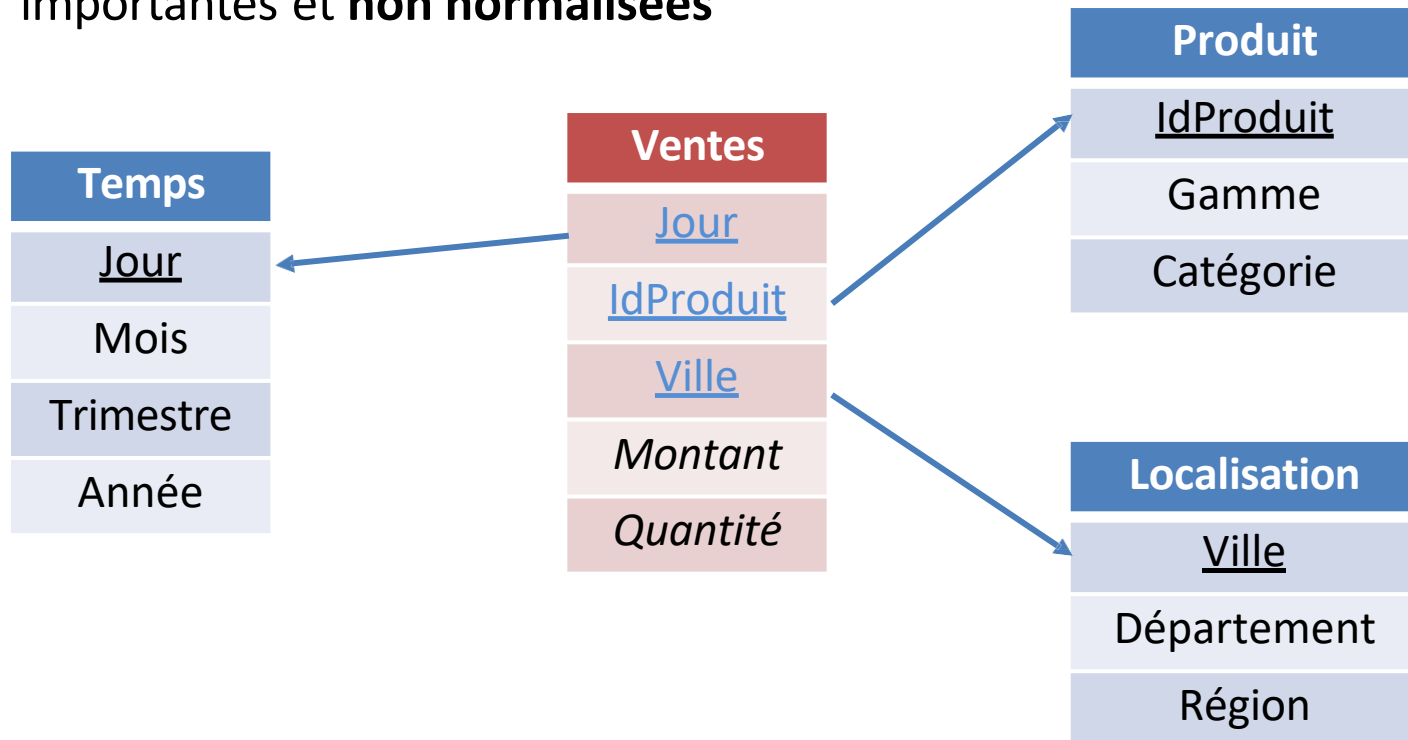


Schéma en flocon

- Objectifs

- **Normaliser les tables de dimensions**, ce qui peut induire de calculs plus coûteux (jointures nombreuses), mais rendre plus facile l'évolution des hiérarchies

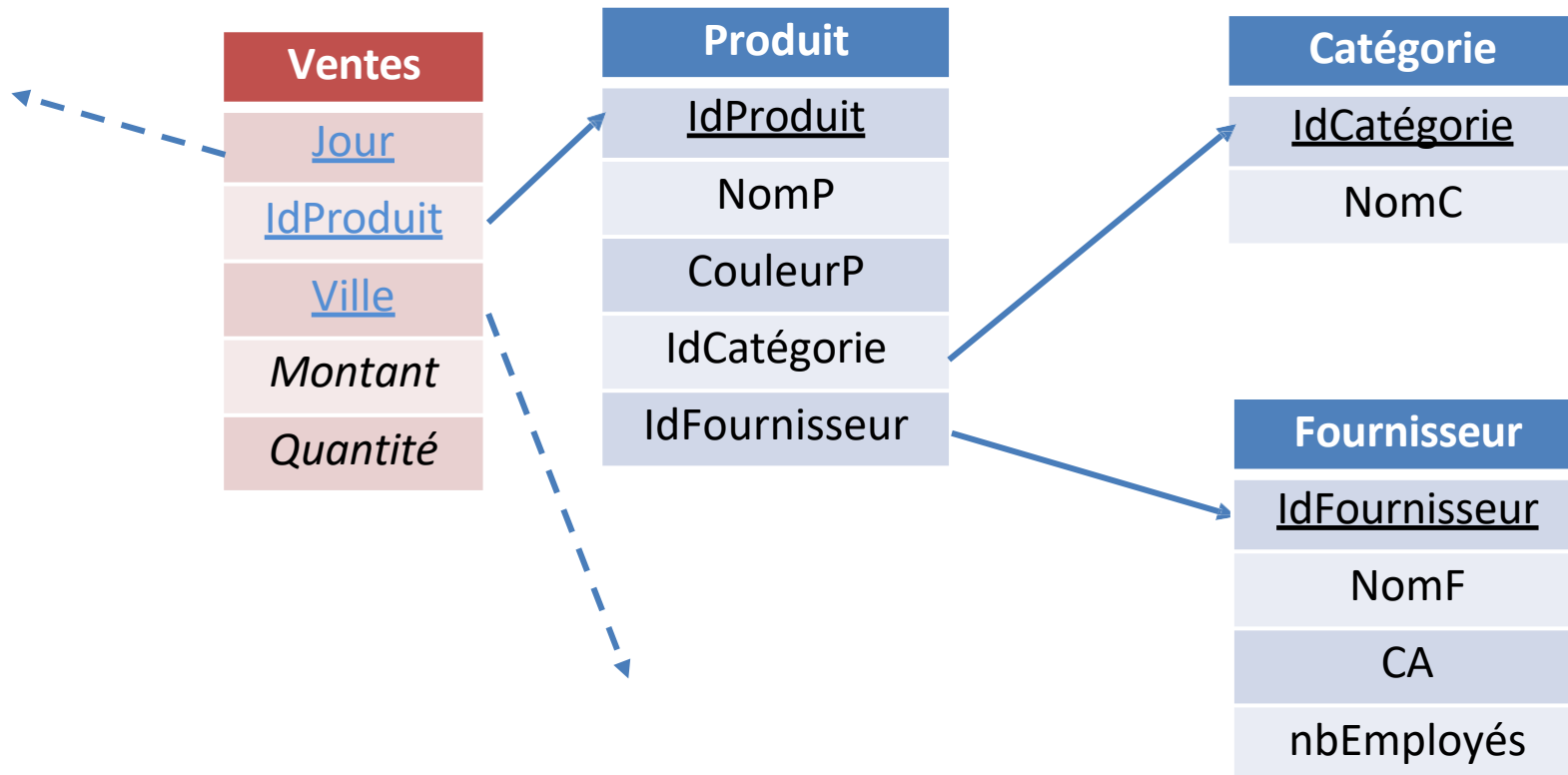
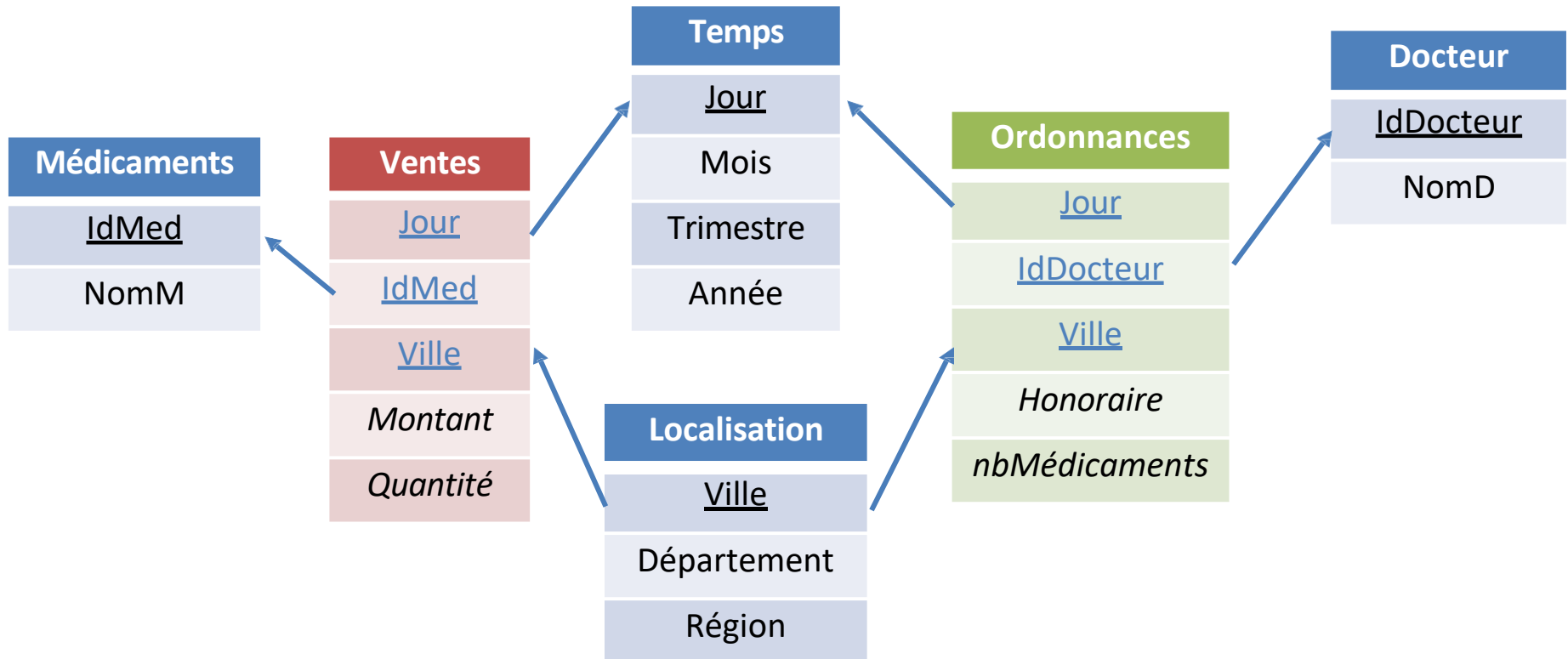


Schéma en constellation

- **Caractéristique**

- Fusionne plusieurs schémas en étoile, (ex : **Ordonnances** et **Ventes**), avec plusieurs tables de faits qui partagent des dimensions communes (ex : Temps et Localisation)



Alimentation d'un entrepôts de données

Processus ETL (Extraction,
Transformation, Load)

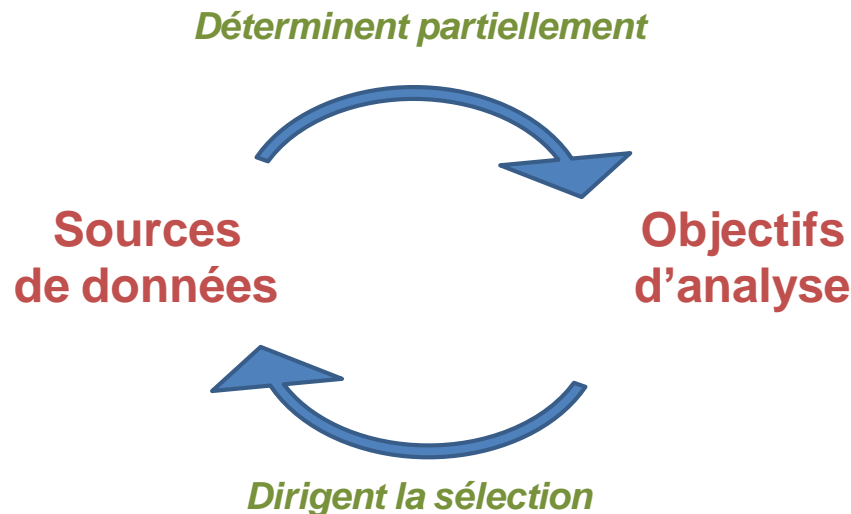
Dr. Aymen GAMMOUDI

Processus d'alimentation

- **Objectif principal**
 - Rassembler des données sources multiples et hétérogènes
 - Homogénéiser (intégrer) ces données selon des règles précises
 - *En général, ces **règles** sont mémorisées sous forme de métadonnées stockées dans des **dictionnaires de données***
 - *Afin de faciliter le travail d'administration et gestion des entrepôts*
- **Un processus ETL en quatre étapes**
 - Sélection des données sources
 - **E**xtraction des données sources
 - Nettoyage et **T**ransformation des données sources
 - Chargement / **L**oading

1 - Sélection des données sources

- **Quelles données faut-il intégrer dans l'entrepôt ?**
 - Des données internes (bases de production) ou externes (données du web)
 - Toutes les données ne sont pas forcément utiles
 - Par contres, l'intégration de données de sources différentes est essentielle pour permettre de nouveaux croisements



2 – *Extraction des données sources*

- **Un extracteur (wrapper) est associé à chaque source**
 - Il sélectionne et extrait des données de types variés : fichiers Excel, logs, BD opérationnelles, tweets, ...
 - Il les formate dans un format cible commun
 - Via l'utilisation d'interfaces comme ODBC, JDBC, ...
 - Le format cible est souvent le modèle relationnel
- **Avec deux stratégies de rafraichissements**
 - **Push** : ce sont les sources qui déclenchent une nouvelle extraction
 - **Pull** : les sources sont interrogées par les extracteurs
 - **Et une contrainte essentielle** : ne pas trop perturber les opérations OLTP (des bases sources)

3a – Nettoyage des données

- **Constat**
 - 5% à 30% des données dans des BD opérationnelles peuvent être erronées
 - Impossible de conduire de bonnes analyses à partir de données de mauvaises qualité
- **Plusieurs sous-tâches**
 - Suppression des **doublons**
 - Traitement des valeurs **manquantes**
 - Détection des valeurs **erronées** (ex : à partir de dictionnaire) ou **incohérentes**

3b – Transformation des données

- **Pour l'intégration des schémas**
 - En identifiant que des **noms d'attributs différents** représentent le **même type d'entité**
 - En identifiant des niveaux d'abstraction / de granularités qui sont différents
- **Pour l'intégration des données**
 - Par l'application de fonctions de normalisation, de conversion
 - *Exemple : valeurs dans des unités de mesure différentes*
 - Via des dictionnaires de synonymes et/ou abréviations
 - *Exemple : pour associer la même valeur M. à Mr, Monsieur, etc.)*

4 – Chargement des données

- **Objectif**
 - Charger les données nettoyées et transformées dans l'entrepôt
- **Différentes politiques**
 - **Complet** ou **incrémental**
 - **En ligne** ou **hors ligne**
- **Quelques problèmes posés**
 - Coût de rafraichissement des structures d'indexation
 - Evolution des dimensions au cours du temps
(**ex** : changement de nom d'un produit ; nouveau niveau d'analyse dans une dimension)

Exploitation d'un entrepôts de données

Reporting, analyse OLAP, langage
de requêtes MDX, etc.

Dr. Aymen GAMMOUDI

A partir d'un entrepôt


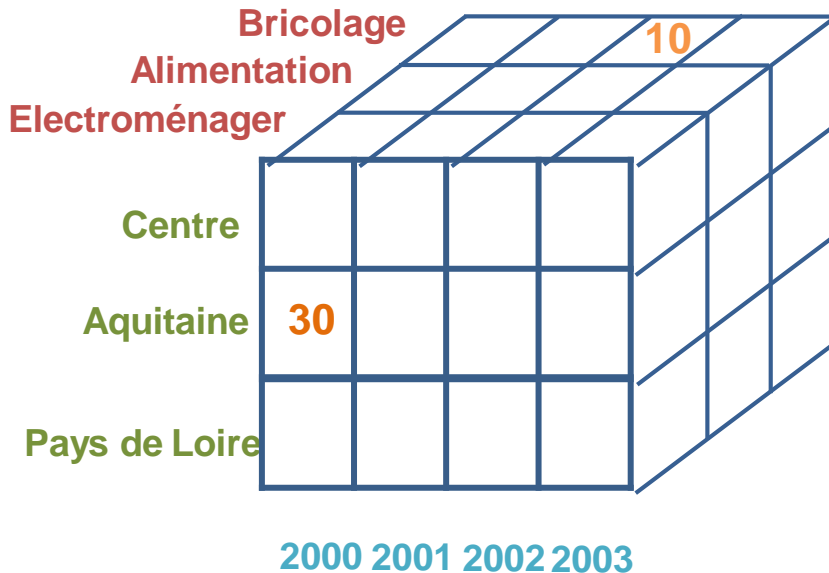
- **Différentes exploitations possibles**
 - Réalisation de rapports (reporting)
 - Réalisation de tableaux de bords (dashboards)
 - **Analyse en ligne OLAP** (OnLine Analytical Processing)
 - Fouille de données (data mining)
 - Visualisations de données
 - Etc.

OLAP : On Line Analytical Processing

- **Objectif**
 - Permettre une navigation / exploration interactive des données dans un cube de données
- **Algèbre OLAP**
 - Opérateurs **ensemblistes** : sélection classiques
 - Opérateurs de **restructuration** : pour changer de point de vue sur les données
 - Opérateurs liés à la **granularité** : pour « zoomer » ou « de-zoomer »

Visualisation de résultats

- En 2D sous la forme de tableaux croisés



		2000	2001	2002
Centre	Bricolage	20	20	10
	Alimentation	20	20	20
	Electroménager	30	20	30
Aquitaine	Bricolage	20	20	20
	Alimentation	20	20	20
	Electroménager	30	20	20

Opérateur : slice (tranche)

		2000	2001	2002
Centre	Bricolage	20	20	10
	Alimentation	20	20	20
	Electroménager	30	20	30
Aquitaine	Bricolage	20	20	20
	Alimentation	20	20	20
	Electroménager	20	20	20

De l'année
2000



		2000
Centre	Bricolage	20
	Alimentation	20
	Electroménager	30
Aquitaine	Bricolage	20
	Alimentation	20
	Electroménager	20

Opérateur : rotate

		2000	2001	2002
Centre	Bricolage	20	20	10
	Alimentation	20	20	20
	Electroménager	30	20	30
Aquitaine	Bricolage	20	20	20
	Alimentation	20	20	20
	Electroménager	20	20	20

Pour changer
de perspectives



		Brico.	Alim.	Electro
Centre	2000	20	20	30
	2001	20	20	20
	2002	10	20	20
Aquitaine	2000	20	20	20
	2001	20	20	20
	2002	20	20	20

Opérateur : rollup / drill down

Rollup sur la dimension « **Produit** » avec **somme**

Drill down sur la dimension « **Produit** »

	95	96
Alimentation	100	130

Rollup sur la dimension « **Temps** » avec **somme**

	95	96
Laitage	20	20
Légume	20	30
Viande	60	80

	S1-95	S2-95	S1-96	S2-96
Laitage	10	10	5	15
Légume	8	12	15	15
Viande	30	30	30	50

Drill down sur la dimension « **Temps** »

Langage de requêtes OLAP

- **MDX : Multi Dimensional eXpression**
 - **Un standard de fait** développé en 1997 par M. Pasumansky au sein de Microsoft
 - Fait pour naviguer dans les bases multidimensionnelles
 - *En définissant des requêtes sur tous leurs objets (dimensions, hiérarchies, niveaux, membres et cellules)*
 - Une requête MDX retourne
 - *Un rapport à plusieurs dimensions consistant en un ou plusieurs tableaux 2D imbriqués*
 - Utilisé aujourd'hui par de **nombreux outils de BI** commerciaux ou non
 - **Langage assez complexe** permettant des **requêtes souvent plus compactes** que les requêtes SQL « équivalentes »

Exemple de requête MDX

- **Construction d'un tableau croisé**

- Avec en **colonne** les familles de produit (de la dimension **Product**)
- Avec en **ligne** les villes (de la dimension **Customer**) croisés avec les trimestres (de la dimension **Time**)
- Et dans les cellules le nombre d'unité de produit vendu (**Unit Sales** : une des mesures du cube **Sales**)
- La fonction d'agrégat utilisée est **par défaut la somme**

SELECT

 [**Product**].[Product Family].MEMBERS **ON COLUMNS**,
 {**CROSSJOIN**([Customer].[City].MEMBERS,
 [Time].[Quarter].MEMBERS)} **ON ROWS**

FROM [Sales]

WHERE (Measures.[Unit Sales])