BI = Business Intelligence
Master Data-Science
Cours 3 - Datawarehouse

Ludovic DENOYER - ludovic.denoyer@lip6.fr

UPMC

8 février 2015

Rappel

L'Informatique Décisionnelle (ID), en anglais Business Intelligence (BI), est l'informatique à l'usage des décideurs et des dirigeants des entreprises. Les systèmes de ID/BI sont utilisés par les décideurs pour obtenir une connaissance approfondie de l'entreprise et de définir et de soutenir leurs stratégies d'affaires, par exemple :

- d'acquérir un avantage concurrentiel,
- d'améliorer la performance de l'entreprise,
- de répondre plus rapidement aux changements,
- d'augmenter la rentabilité, et
- d'une façon générale la création de valeur ajoutée de l'entreprise.
- **...et à créer de nouveaux services...**

Plan

- Vision générale
- ETL
- Datawarehouse
- OLAP
- Reporting
- Data Mining

Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise.
Source : Wikipedia

Collecter : Récupérer l'information produite pr l'entreprise

Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise.
Source : Wikipedia

Ordonner : Structurer l'information dans le but de la prise de décision (structure différente des BDs opérationnelles)

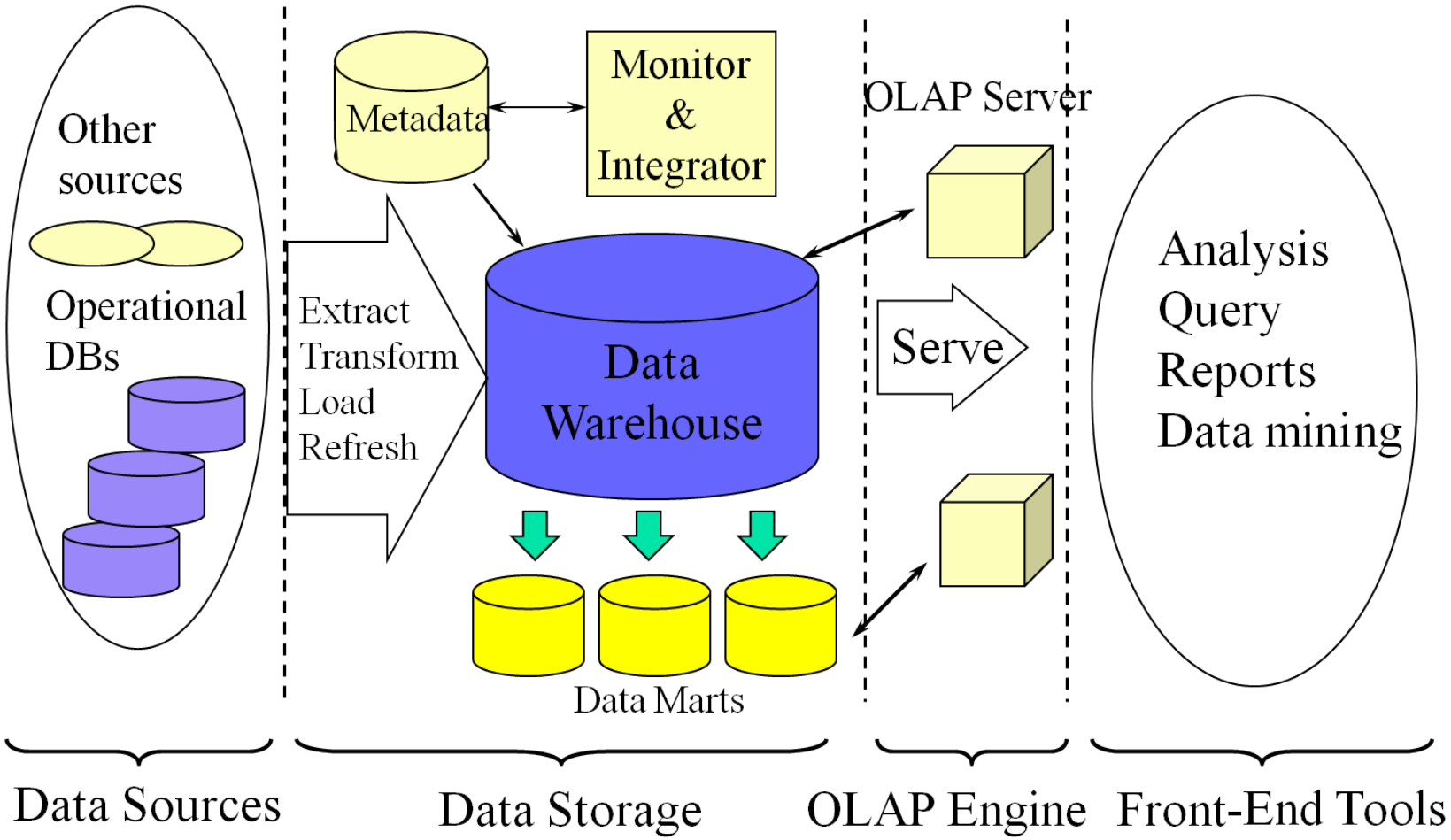
Définition

Datawarehouse

Le terme entrepôt de données (ou base de données décisionnelle, ou encore data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi un socle à l'aide à la décision en entreprise.
Source : Wikipedia

Journaliser : Stocker l'historique des données

Datawarehouse



Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Rassembler des données provenant de sources multiples en une seule base de données afin qu'un moteur de requête unique puisse être utilisé pour présenter des données.
- Permettre l'éxécution de requête longues, bloquantes, sur des données opérationnelles
- Maintenir l'historique des données, même si les systèmes de transaction source ne le font pas
- Intégrer des données provenant de multiples systèmes sources, permettant une vue centrale dans l'entreprise. Cet avantage est particulièrement valable lorsque l'organisation est issue de fusions successives
- Améliorer la qualité des données

Datawarehouse

Un entrepôt de données conserve une copie des informations des systèmes de transaction opérationnels. Il offre la possibilité de :

- Présenter l'information de l'organisation
- Fournir un seul modèle de données commun pour toutes les données d'intérêt, indépendamment de la source de données
- Restructurer les données de sorte qu'elles prennent sens (décisionnel)
- Ajouter de la valeur aux applications métiers opérationnels, notamment la gestion de la relation client (CRM).
- Faire des requêtes d'aide à la décision plus faciles à écrire.

Datawarehouse vs BD opérationnelle

Table 11-1 Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

Datamart

Définition

Un DataMart (littéralement en anglais magasin de données) est un sous-ensemble d'un DataWarehouse destiné à fournir des données aux utilisateurs, et souvent spécialisé vers un groupe ou un type d'affaire. Techniquement, c'est une base de données relationnelle utilisée en informatique décisionnelle et exploitée en entreprise pour restituer des informations ciblées sur un métier spécifique, constituant pour ce dernier un ensemble d'indicateurs utilisés pour le pilotage de l'activité et l'aide à la décision.
Source : wikipedia

Le datawarehouse est **Général**, le datamart est **spécifique** à un métier.

Datamart vs datawarehouse

Table 11-2 Data Warehouse Versus Data Mart

Data Warehouse	*Data Mart*
Scope	*Scope*
• Application independent	• Specific DSS application
• Centralized, possibly enterprise-wide	• Decentralized by user area
• Planned	• Organic, possibly not planned
Data	*Data*
• Historical, detailed, and summarized	• Some history, detailed, and summarized
• Lightly denormalized	• Highly denormalized
Subjects	*Subjects*
• Multiple subjects	• One central subject of concern to users
Sources	*Sources*
• Many internal and external sources	• Few internal and external sources
Other Characteristics	*Other Characteristics*
• Flexible	• Restrictive
• Data-oriented	• Project-oriented
• Long life	• Short life
• Large	• Start small, becomes large
• Single complex structure	• Multi, semi-complex structures, together complex

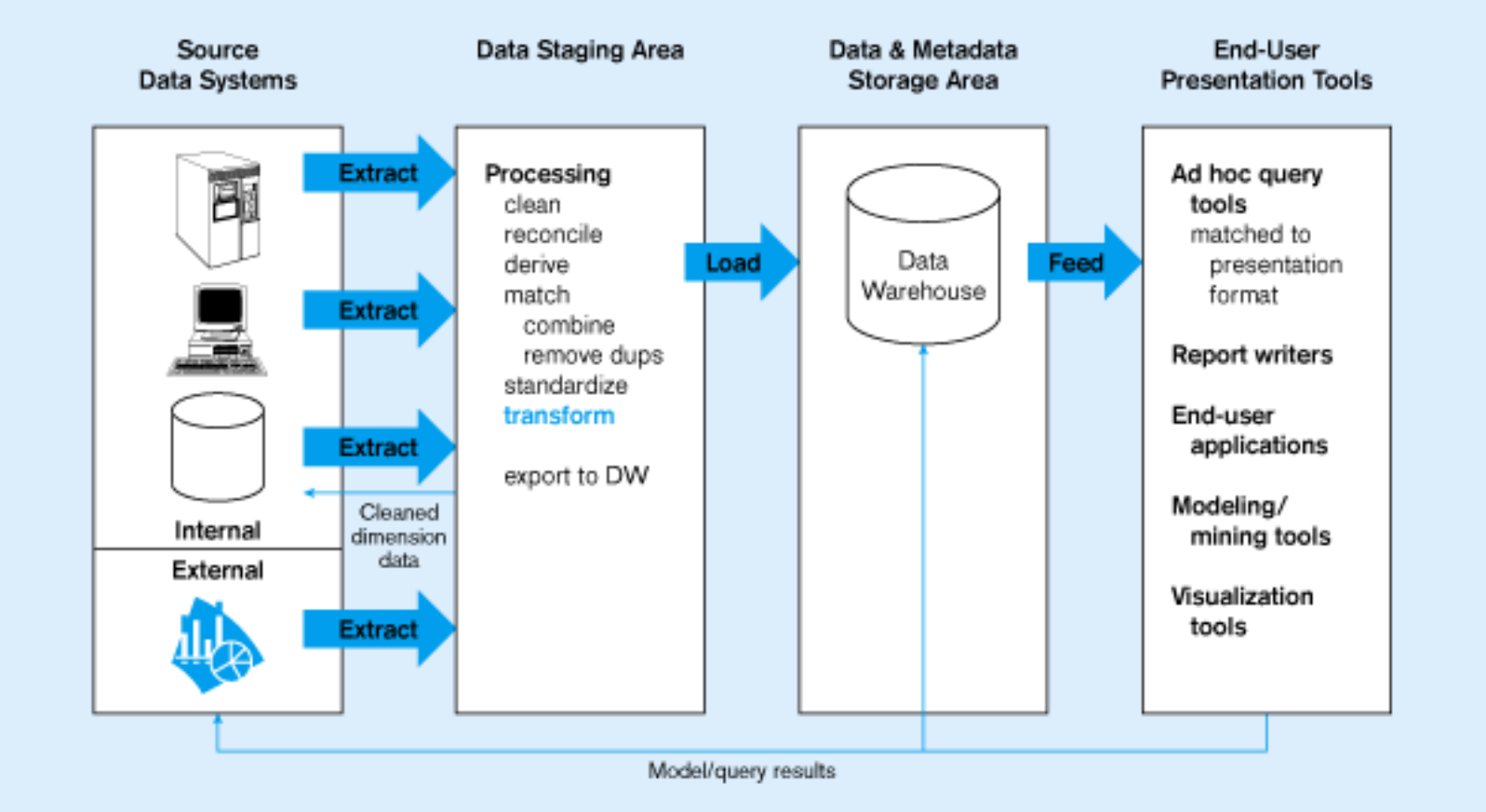
Adapted from Strange (1997)

Datamart vs datawarehouse

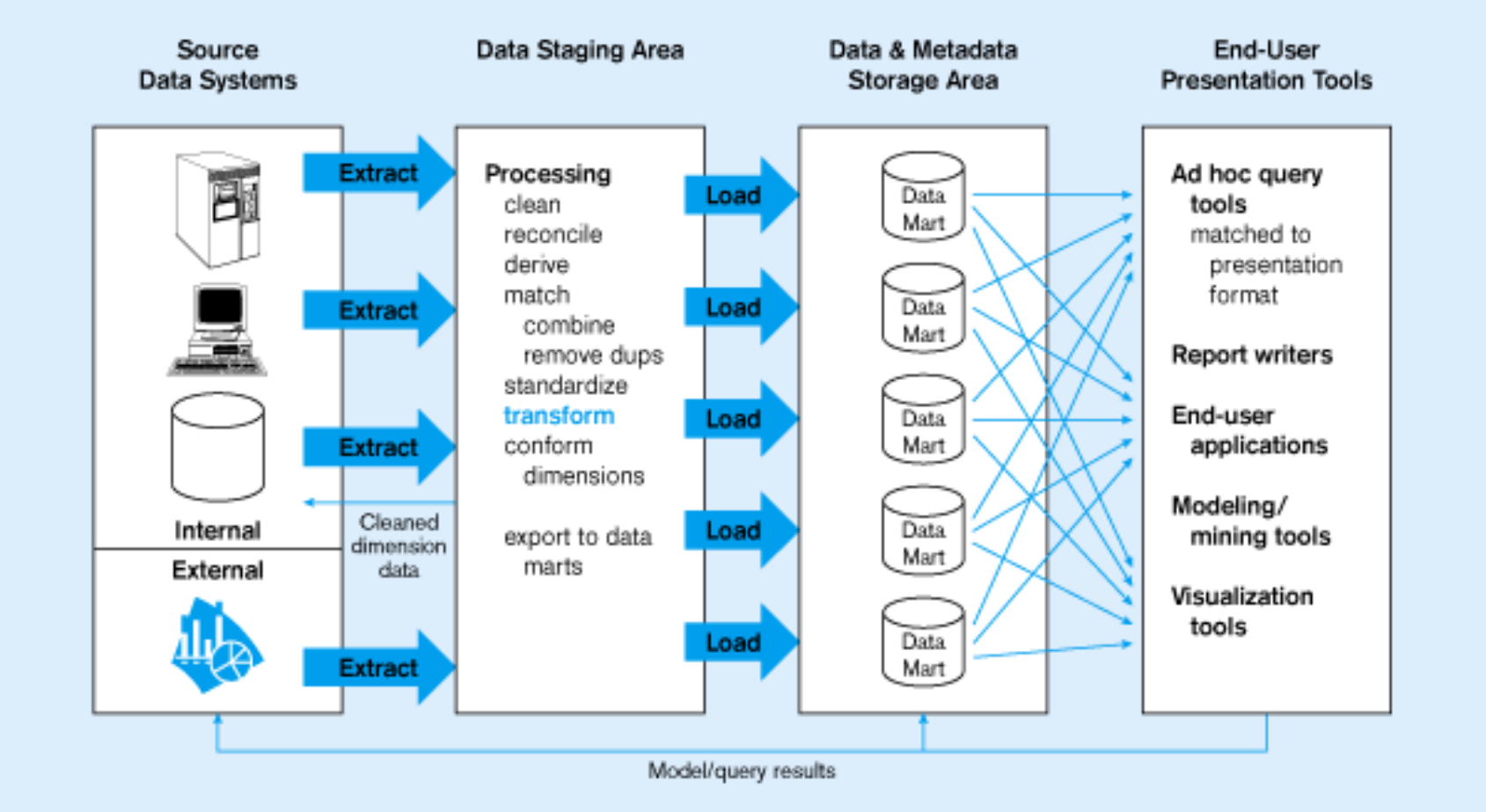
Deux conceptions existantes :

- Définition d'Inmon : Le DataMart est issu d'un flux de données provenant du DataWarehouse. Contrairement à ce dernier qui présente le détail des données pour toute l'entreprise, il a vocation à présenter la donnée de manière spécialisée, agrégée et regroupée fonctionnellement.
- Définition de Kimball : Le DataMart est un sous-ensemble du DataWarehouse, constitué de tables au niveau détail et à des niveaux plus agrégés, permettant de restituer tout le spectre d'une activité métier. L'ensemble des DataMarts de l'entreprise constitue le DataWarehouse.

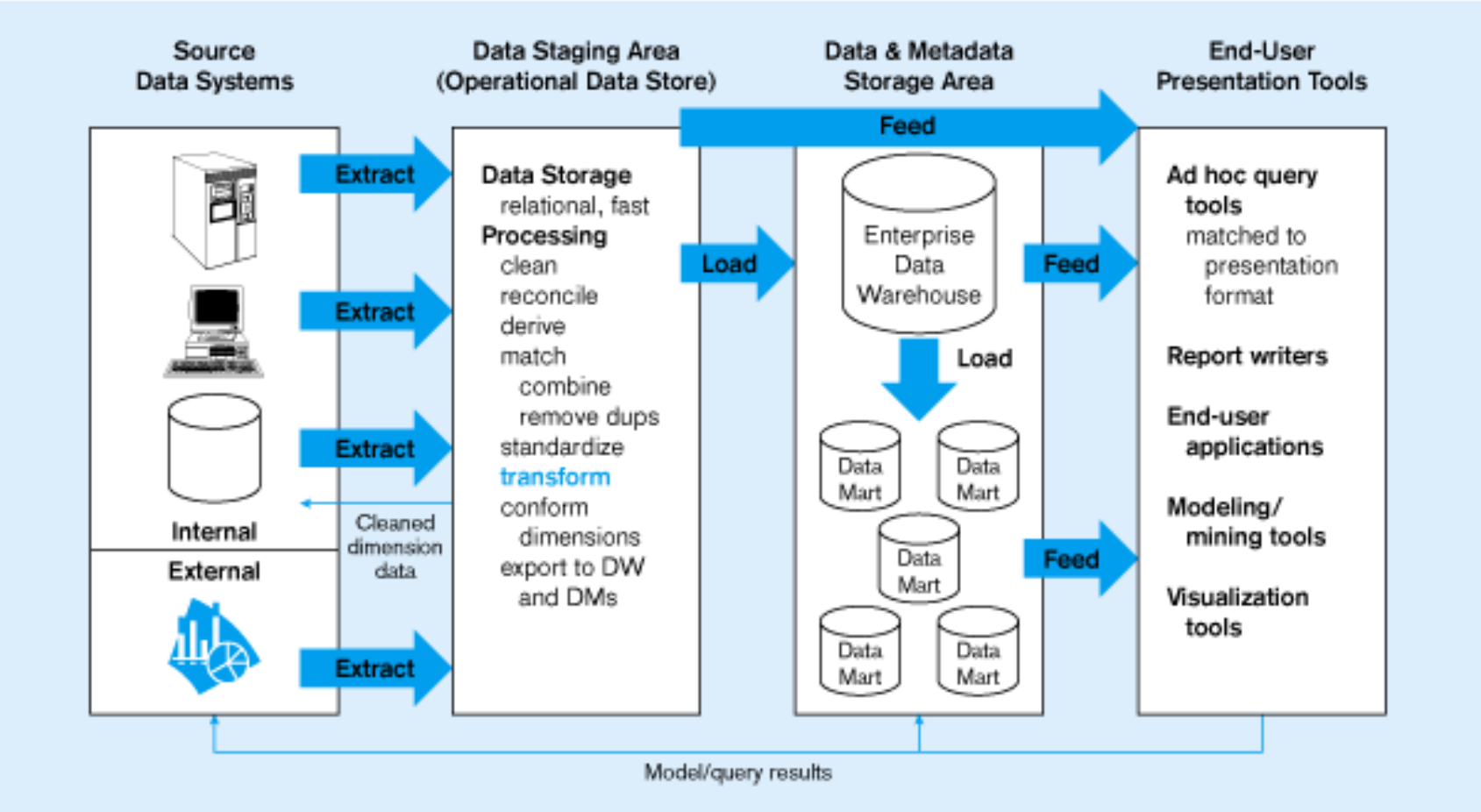
Différentes Architectures



Différentes Architectures

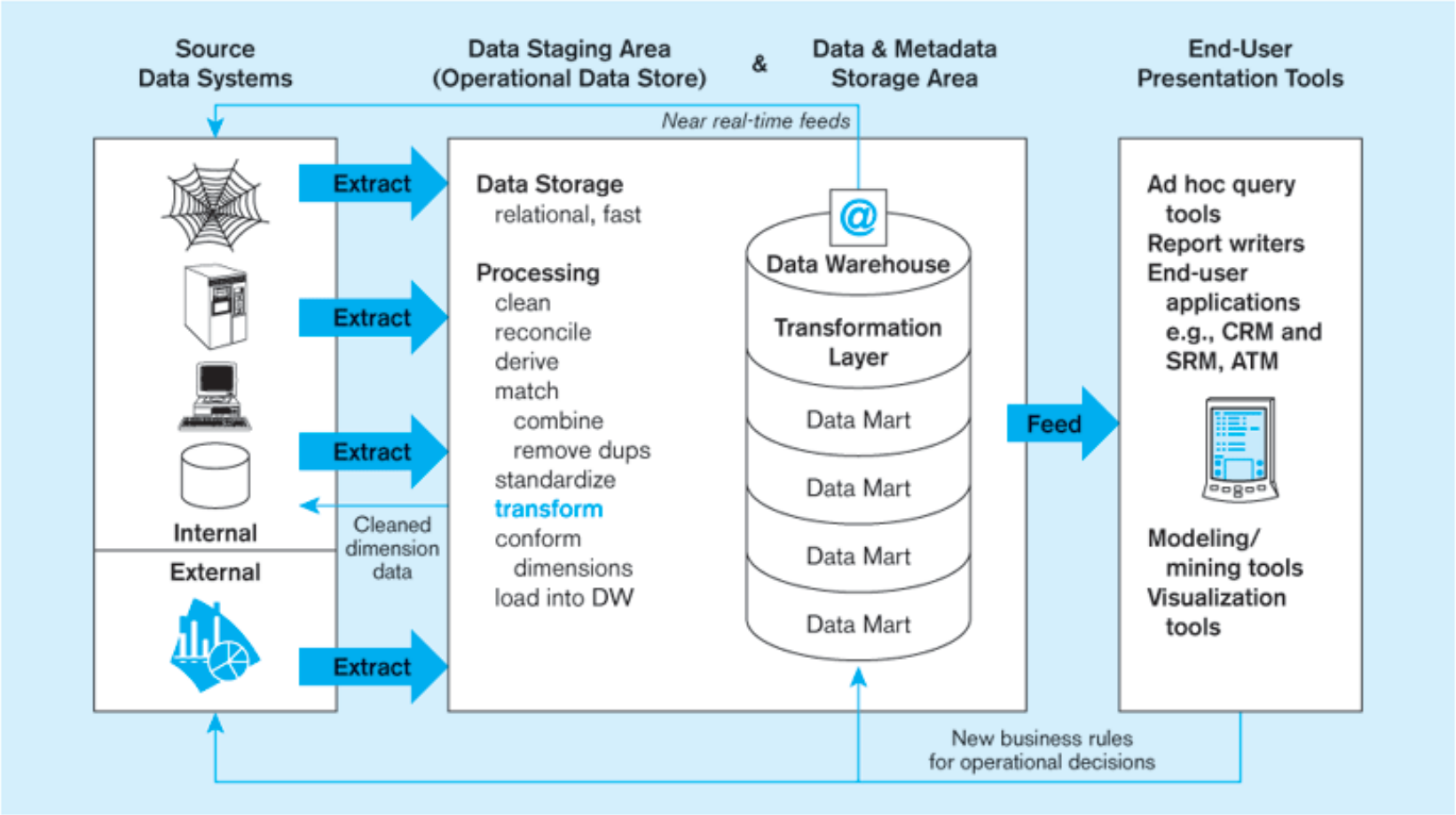


Différentes Architectures



Différentes Architectures

Figure 11-5 Logical data mart and @ctive warehouse architecture



Données orientées sujets

- En production : données organisées par processus fonctionnels
- Datawarehouse : données organisées autour de sujets majeurs
- Données structurées par thème, potentiellement transverses par rapport aux domains fonctionnels et organisationnelles

Examples (médecine) : Actes, Séjours vs Bases par services

Architecture

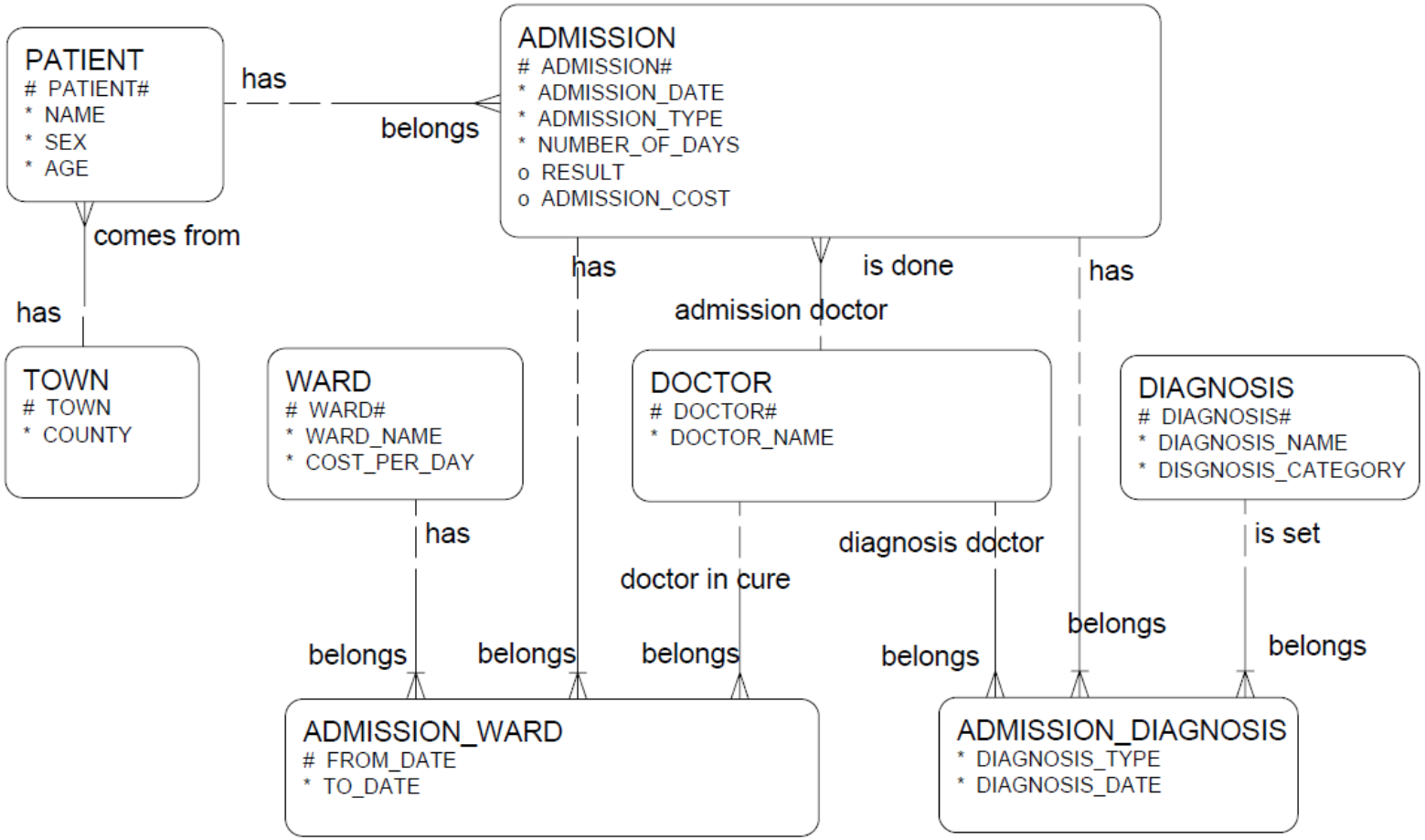
Good DW architecture

"It's not easy to describe a good design, but I'll know it when I see it"

Modèle relationnel

- Normalisation (3NF)
- Répond aux besoins transactionnels (OLTP)
- Avantages :
  - Réduction de l'entrée de données
  - Réduction du nombre d'index
  - Ajouts/destructions/modifications plus rapides
- Désavantages :
  - Peu efficace pour l'extraction de données analytiques
  - Beaucoup de relations
  - Trop complexe pour l'utilisateur BI

Modèle relationnel



Le modèle relationnel n'est pas (très) approprié pour les DWs

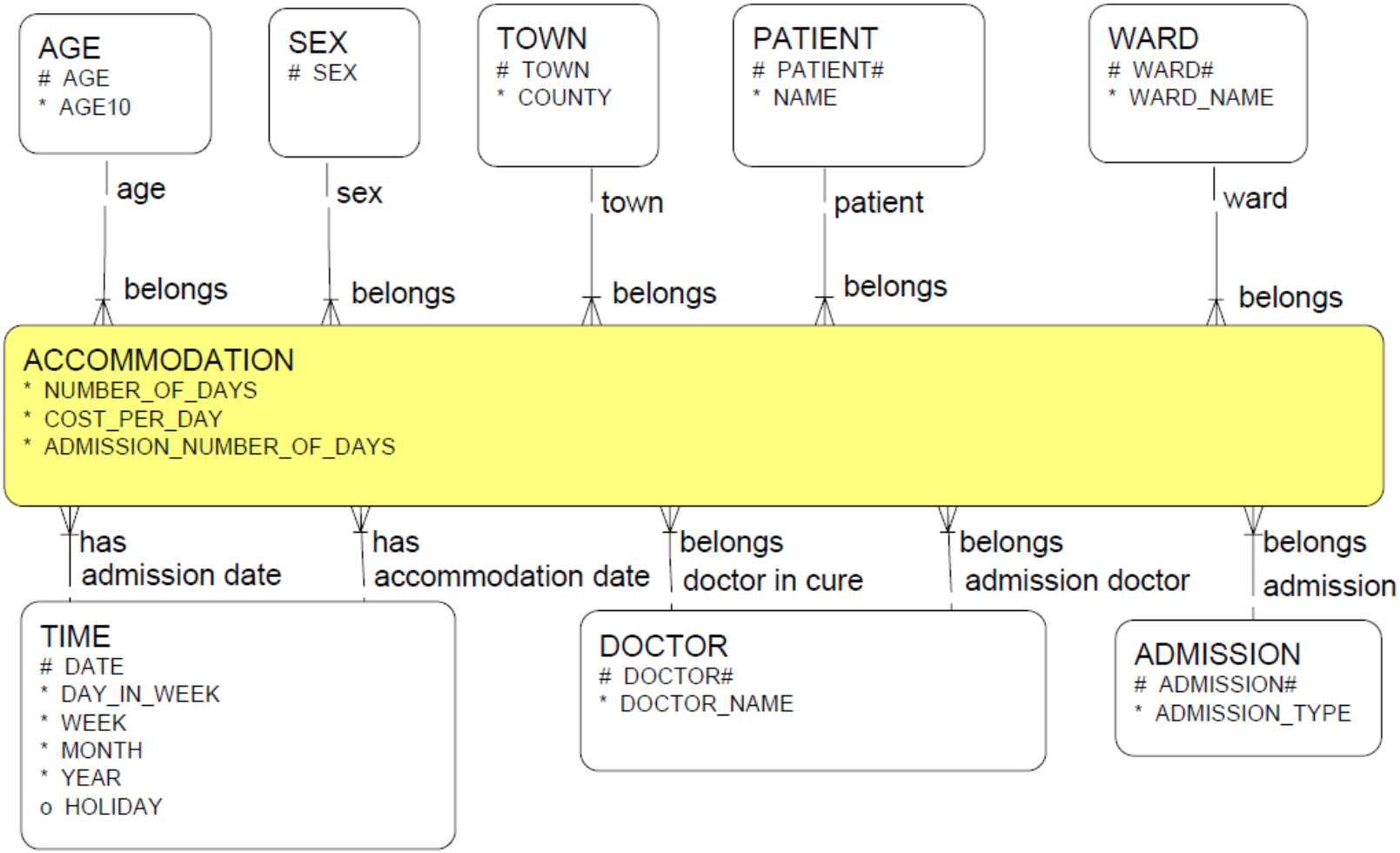
Modèle dimensionnel

Principes

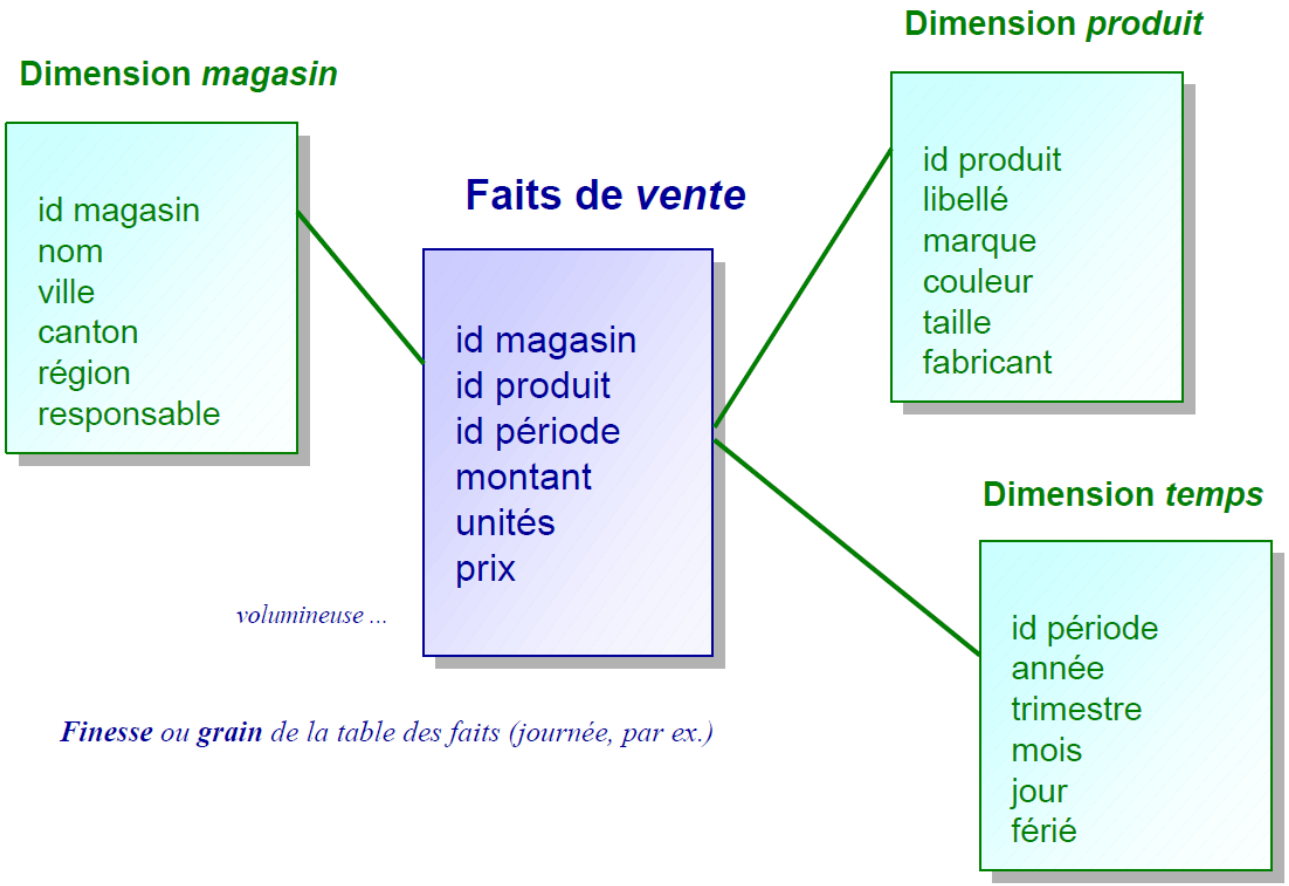
On va partir du besoin "client" (quel analyse ?). On va définir des **faits** et des **dimensions**.

- **Faits :** les faits représentent un sujet d'analyse. Les faits sont caractérisées par plusieurs informations
- **Dimensions :** les dimensions sont les critères selon lesquels on souhaite faire de l'analyse.

Modèle dimensionnel

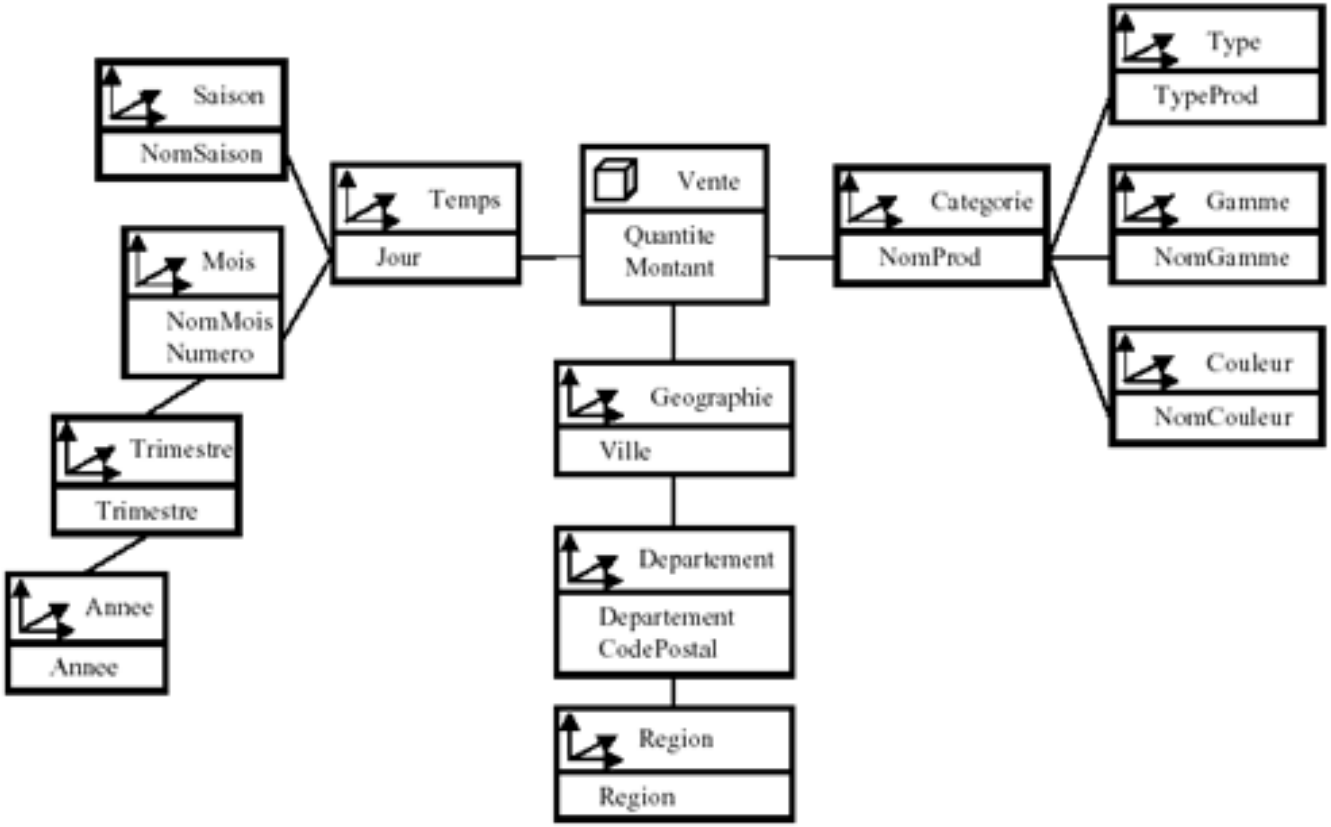


Modèle dimensionnel



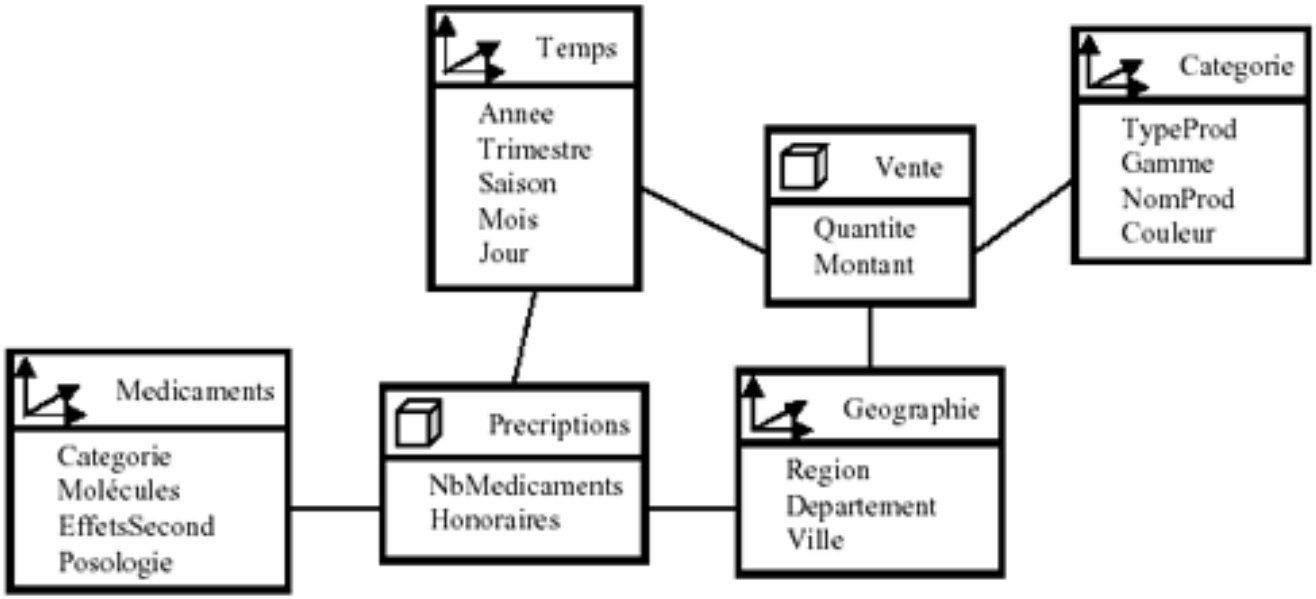
Aussi connu sous le nom de modèle en étoile

Modèle dimensionnel



Aussi connu sous le nom de modèle en flocon

Modèle dimensionnel



Aussi connu sous le nom de modèle en constellation

Dimensions à évolution lente

Slowly Changing Dimensions (SCDs)

On parle d'une dimension à évolution lente (slowly changing dimension) lorsqu'une dimension peut subir des changements de description des membres.

- Un client peut changer dadresse, se marier, ...
- Un produit peut changer de noms, de formulations « Raider » en « Twix »,« Yaourt à la vanille » en « saveur Vanille »

Comment gérer cette situation dans un DW ?

SCD : Type 0

Pas de prise en compte des SCDs

SCD : Type 1

Overwrite

Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Supply Company	CA

Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Supply Company	IL

SCD : Type 2

Utilisation de clefs "surrogate"

(ou clefs de substitution)

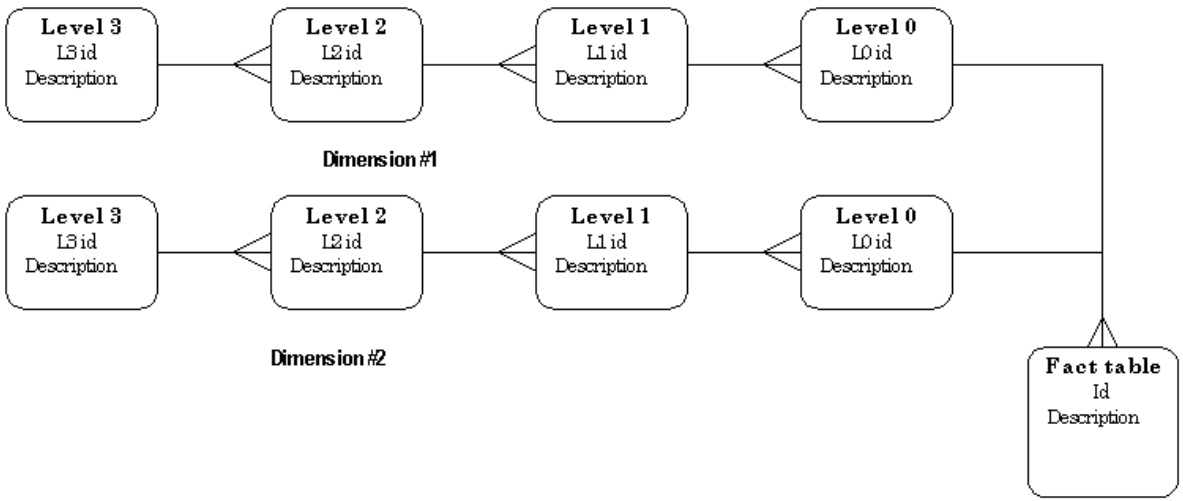
Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Supply Company	CA

Supplier_key	Supplier_Name	Supplier_State
001	Phlogistical Supply Company	IL

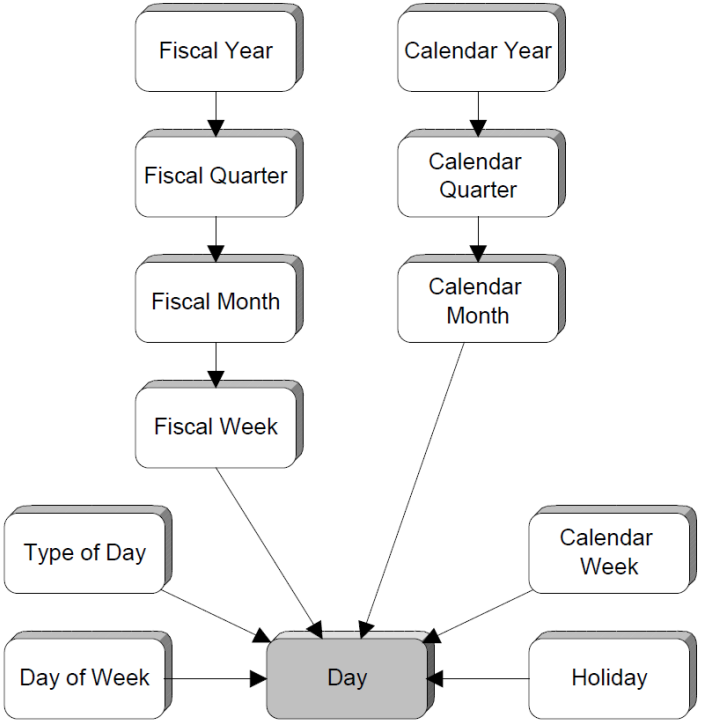
=¿ et d'autres types (voir Kimball)

Dimensions hiérarchiques

Les dimensions peuvent avoir une organisation hiérarchique :



Dimensions hiérarchiques



Conclusion

	Relational modelling	Dimensional modelling
Aim	Data modelling of transactional systems	Data modelling of decision support systems
Analysis subject	Execution of business process → *process flow modelling*	Effects of business process → *process effect modelling* or *informational modelling*
Analysis focus	Discovery of strong entities in the course of business process execution	Discovery of associative entities (relationships of strong entities) that represent the effects of business process
Analysis details	Definition of the strong entities attributes and the relationship between them	Definition of business measures – attributes of associative entities, definition of business dimensions