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

Ludovic DENOYER - ludovic.denoyer@lip6.fr

UPMC

25 janvier 2016

Plan

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

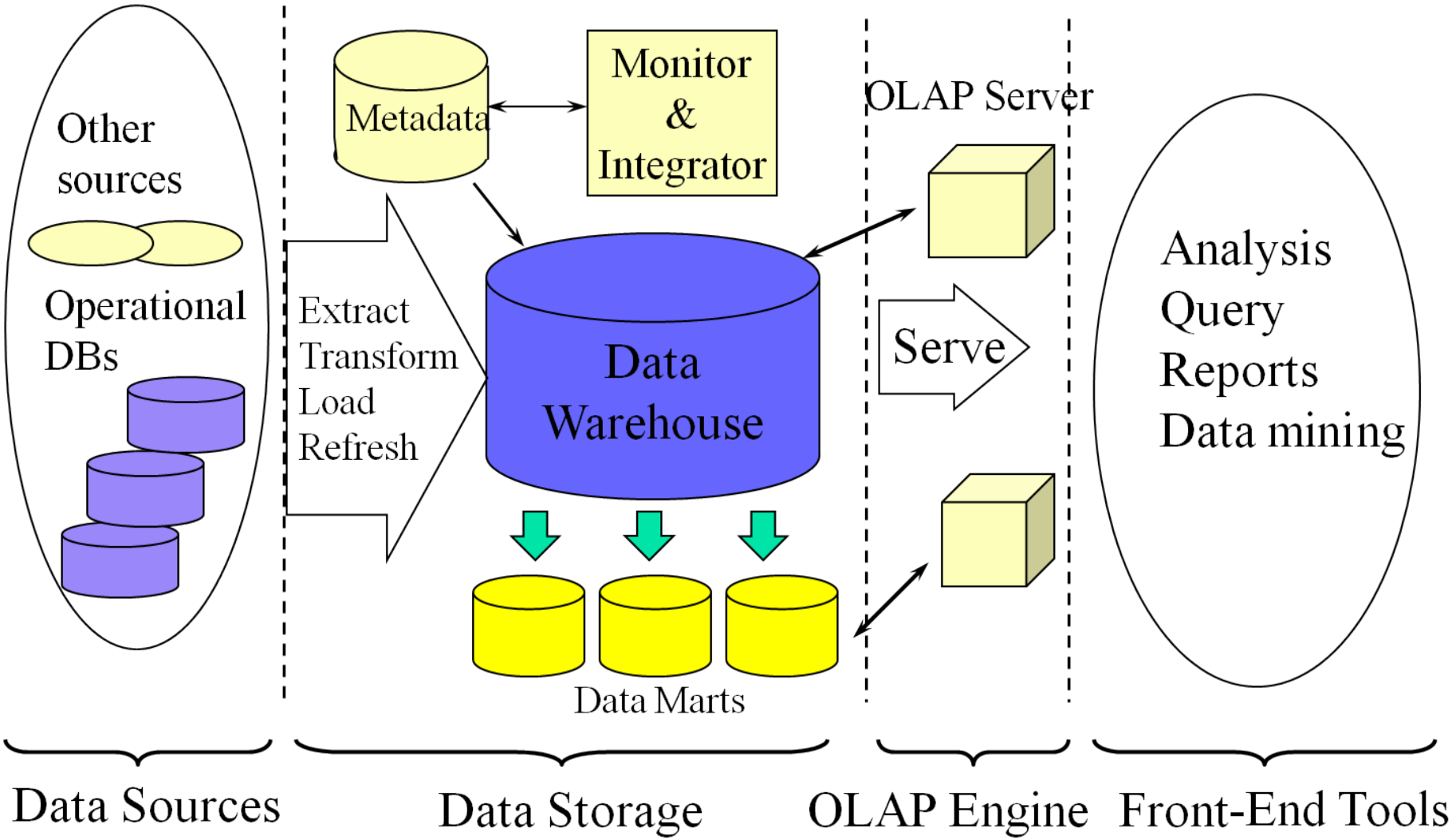
Définition

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 par l'entreprise
- **Ordonner :** Structurer l'information dans le but de la prise de décision (structure différente des BDs opérationnelles)
- **Journaliser :** Stocker l'historique des données

Datawarehouse



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

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

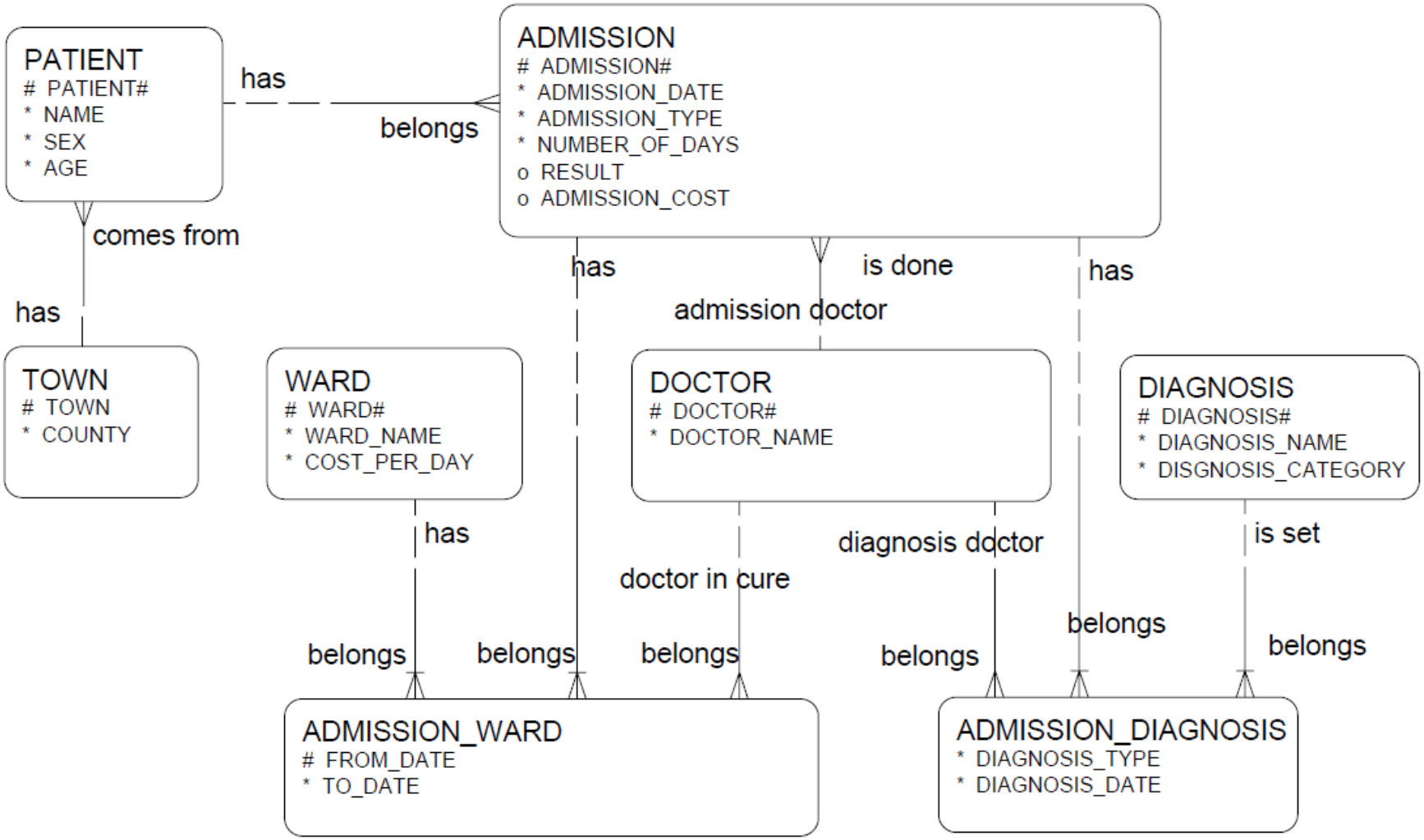
Good DW architecture

"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

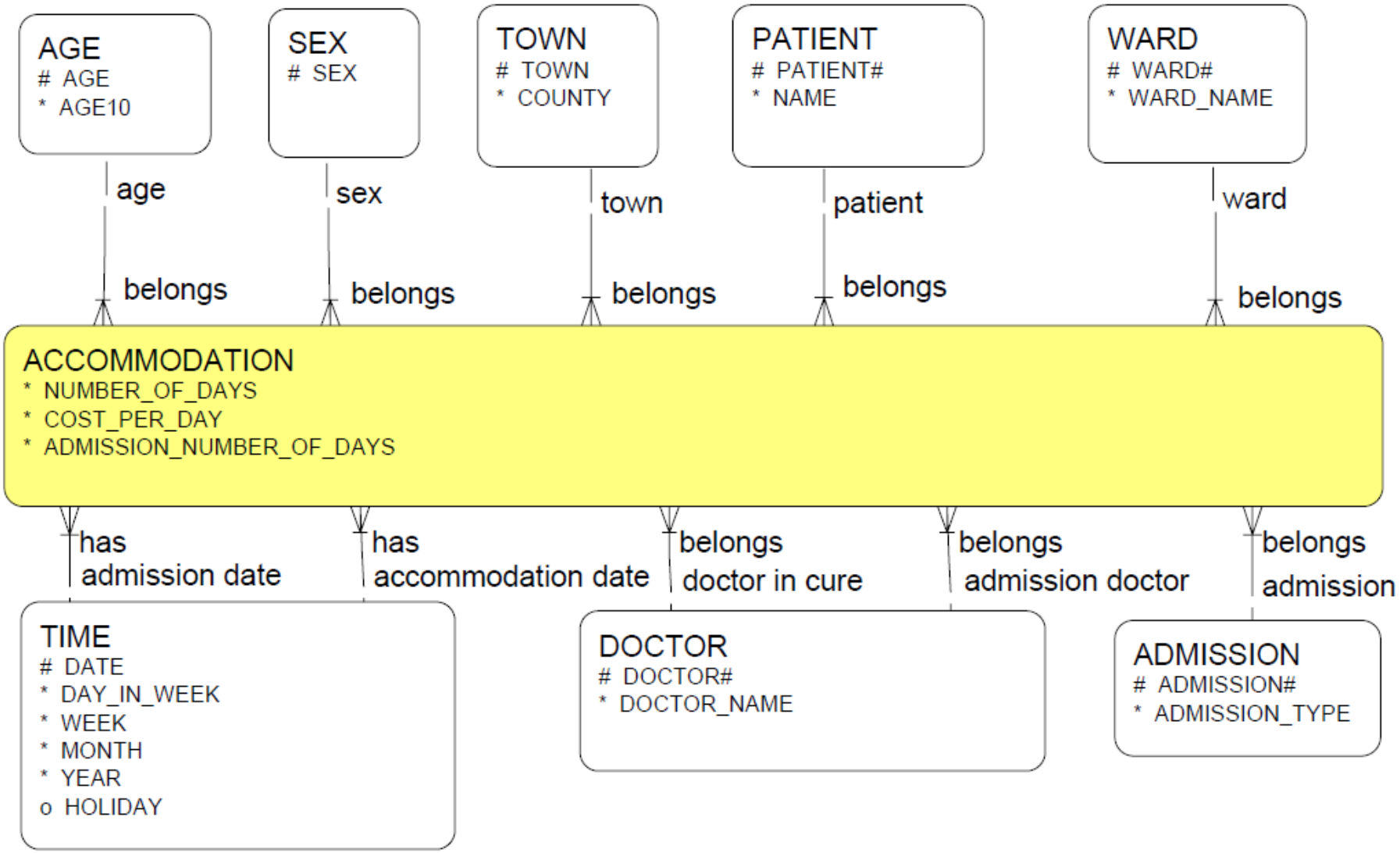
Modèle dimensionnel

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



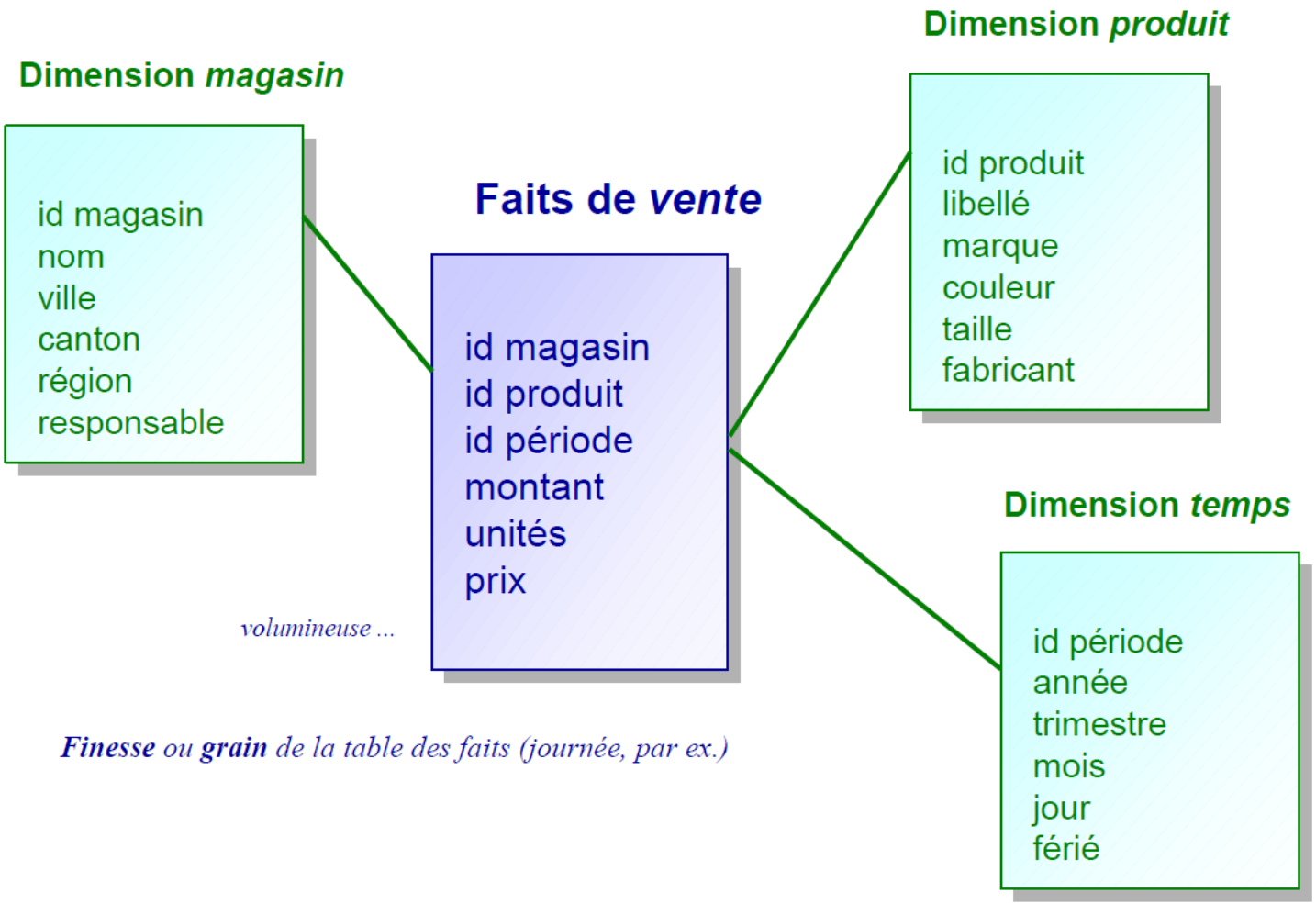
Modèle dimensionnel

On considère le problème d'analyse suivant : notre société vend des produits dans des magasins. Les produits sont décrits par un nom, une marque, un type, une couleur, une taille et un fabricant. Un client (nom, prenom, adresse) peut acheter pour un certain montant plusieurs unités d'un même produit (une vente = 1 ou plusieurs unités). Un magasin possède un responsable et une région.
le service marketing soubhaite être en mesure d'obtenir les réponses aux questions suivantes :

- Je veux un diagramme des ventes globales de l'entreprise dans le temps
- Je veux un diagramme des ventes par magasin
- Je veux un diagramme des ventes par produit, et par magain

Question : Dessinez le schéma du DW sous-jacent. Donnez un exemple de 'valeurs' pour chacune des tables créées.

Modèle dimensionnel



Aussi connu sous le nom de modèle en étoile

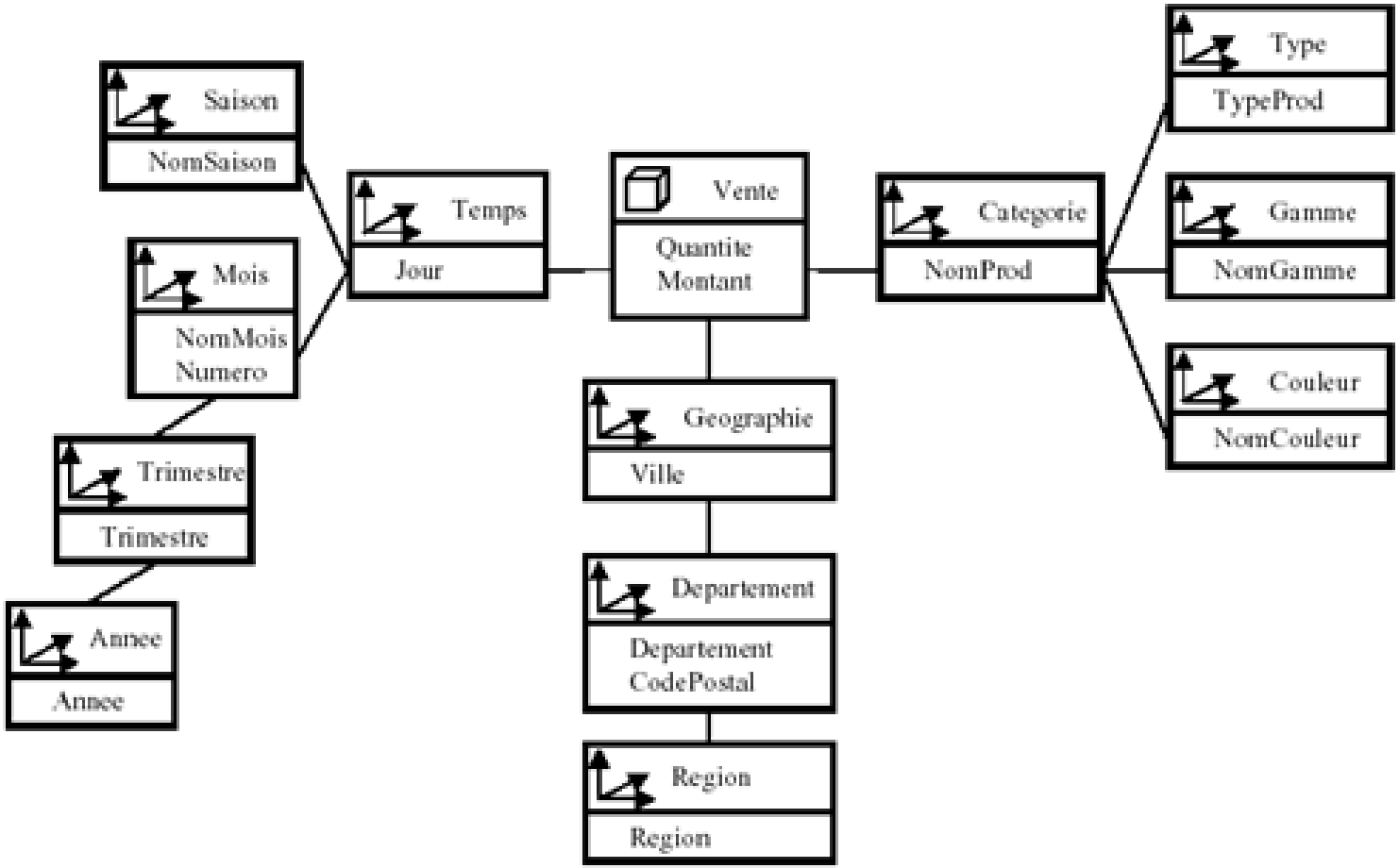
Modèle dimensionnel

On considère maintenant les requêtes suivantes

- Je veux un diagramme des ventes globales de l'entreprise par jour
- Je veux un diagramme des ventes par mois
- Je veux un diagramme des ventes par saison (été, hiver, automne, printemps)
- Je veux un diagramme des ventes par couleur de produit
- Je veux un diagramme des ventes par type de produit
- Je veux un diagramme des ventes par magasin
- Je veux un diagramme des ventes par ville
- Je veux un diagramme des ventes par région

Question : Dessinez le schéma du DW sous-jacent. Donnez un exemple de 'valeurs' pour chacune des tables créées.

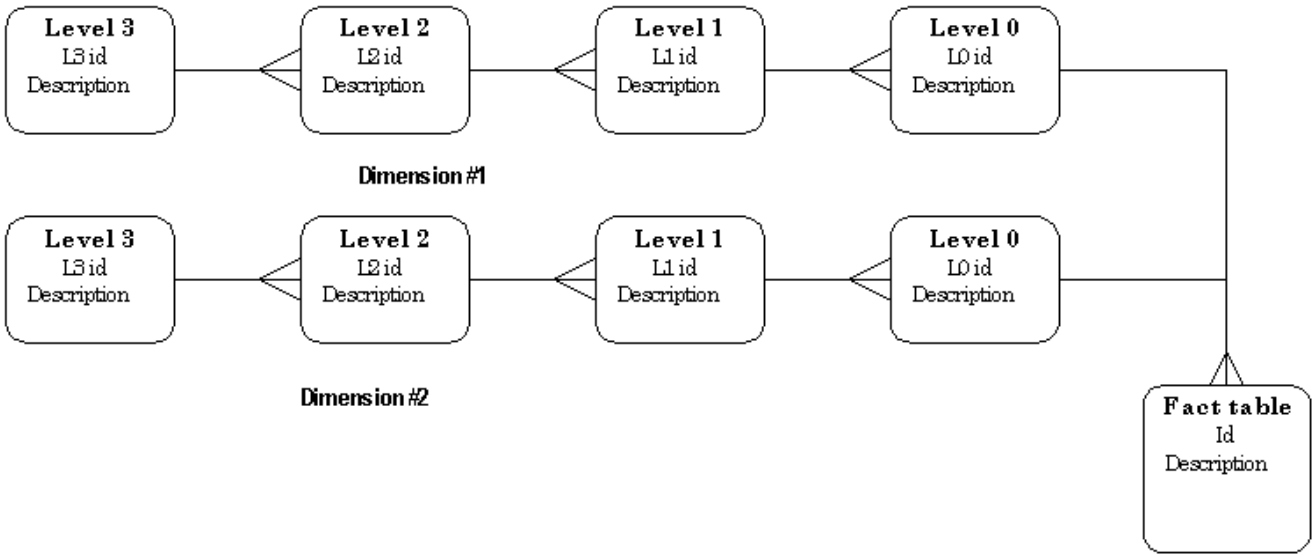
Modèle dimensionnel



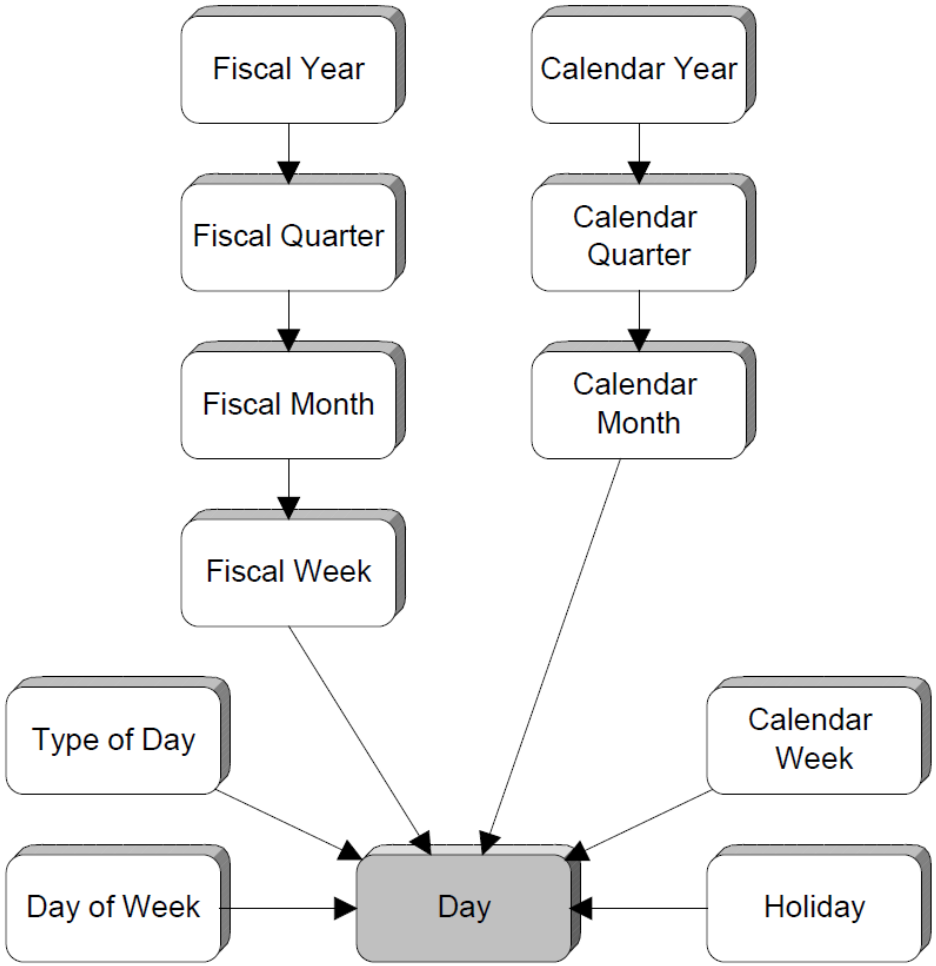
Aussi connu sous le nom de modèle en flocon

Dimensions hiérarchiques

Les dimensions peuvent avoir une organisation hiérarchique :



Dimensions hiérarchiques



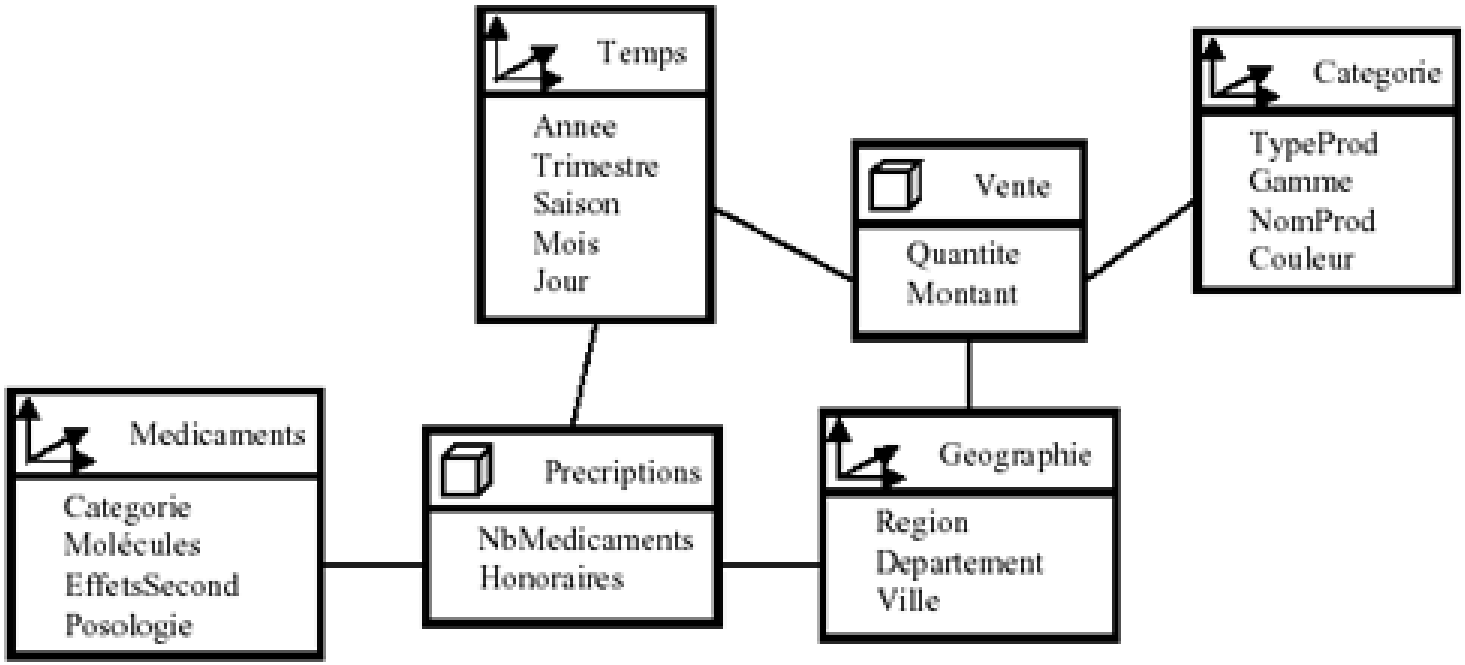
Modèle dimensionnel

On considère maintenant que nos magasins vendent en plus des médicaments. En addition des demandes précédentes, le service du marketing souhaite aussi obtenir :

- Les prescriptions par médicaments (les médicaments ne sont pas des produits, il faut les considérer à part)
- Les prescriptions par médicaments par mois prescriptions par médicaments magasin

Question : Dessinez le schéma du DW sous-jacent. Donnez un exemple de 'valeurs' pour chacune des tables créées.

Modèle dimensionnel



Aussi connu sous le nom de modèle en constellation

Modèle dimensionnel

Enfin, on considère qu'un produit à changer de nom en 2014 (même si cela reste en fait le même produit). On veut continuer à être en mesure de calculer les ventes de ce produit. Quel est le problème ? Proposez une solution. Donnez un exemple de 'valeurs' pour chacune des tables créées.

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 d'adresse, se marier, ...
- Un produit peut changer de nom, de formulation « 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

Supplier Key	Supplier$_{C}ode$	Supplier$_{N}ame$	Supplier State
123	ABC	Acme Supply Co	CA

Supplier Key	Supplier$_{C}ode$	Supplier$_{N}ame$	Supplier State
123	ABC	Acme Supply Co	IL

SCD : Type 2

Supplier Key	Supplier$_{C}ode$	Supplier$_{N}ame$	Supplier State	Version
123	ABC	Acme Supply Co	CA	0
124	ABC	Acme Supply Co	CA	1

Supplier Key End Date	Supplier$_{C}ode$	Supplier$_{N}ame$	Supplier State	Start D
123 21-Dec-2004	ABC	Acme Supply Co	CA	01-Jan-2
124 NULL	ABC	Acme Supply Co	CA	22-Dec-2

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