Intelligence d’affaires - BI

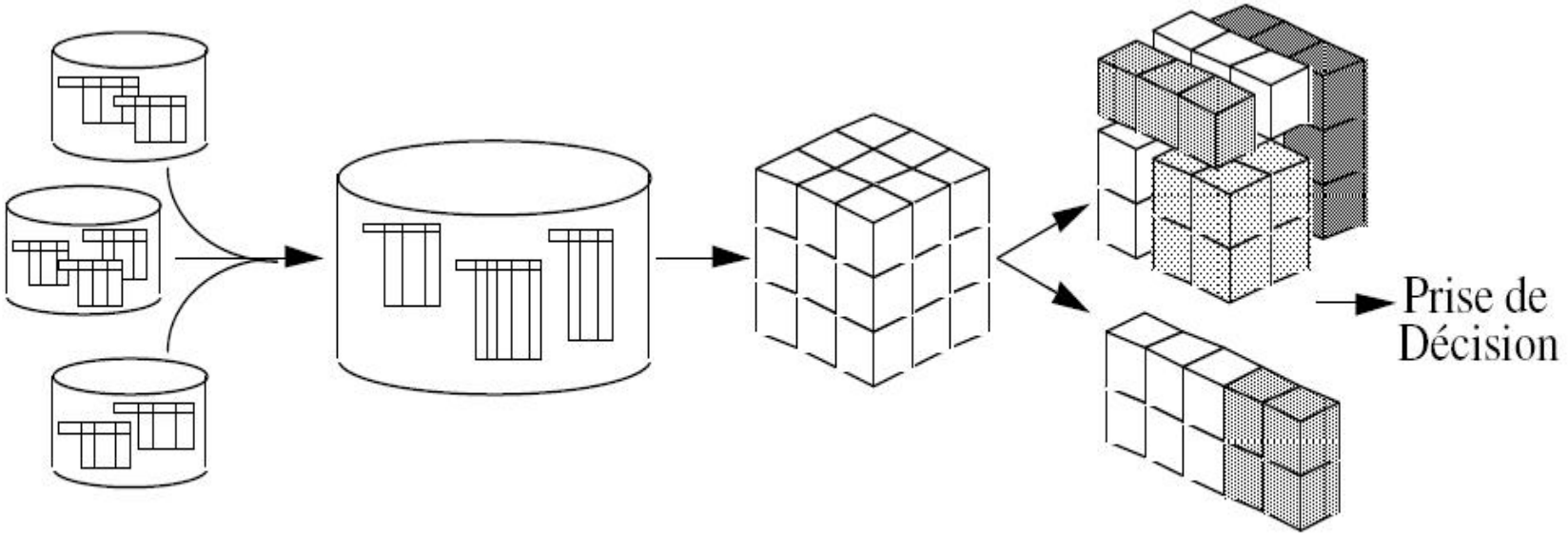
Entrepôt de données
Version Simplifiée

Plan du cours

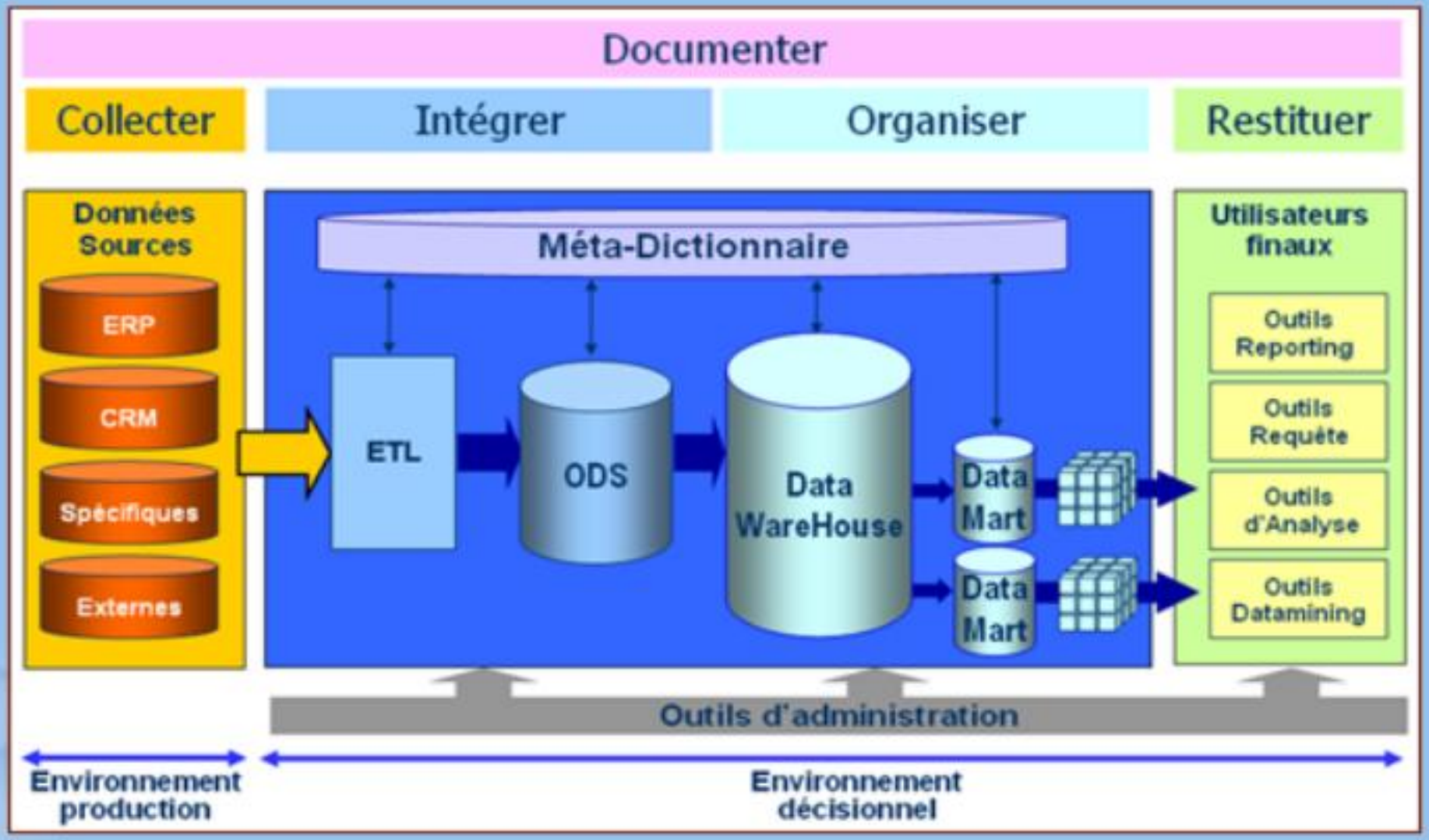
- Définition et concept d'un entrepôt de données

- Architecture générale d'un entrepôt de données
- Modélisation d'un entrepôt de données
- Les types de modèles
- Alimentation et mise à jour d'un entrepôt de données
- Les bases de données dimensionnelles
- Conclusion

De l’entrepôt à la décision



De l'entrepôt à la décision



— Processus détaillé de traitement des données d'un système décisionnel —

Entrepôt de données

Définition et concept

Définition

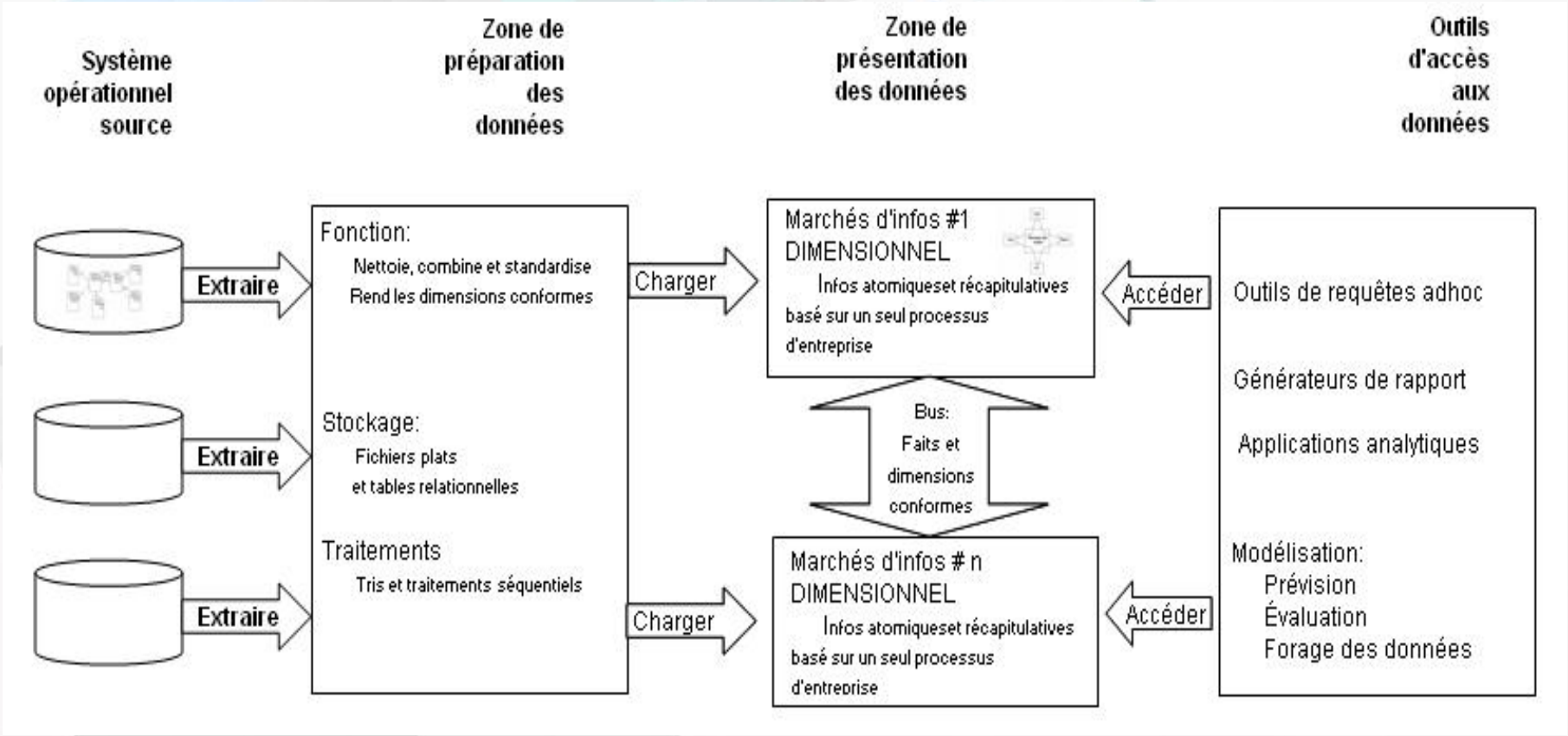
- Ensemble de données historisées variant dans le temps, organisé par sujets, consolidé dans une base de données unique, géré dans un environnement de stockage particulier, aidant à la prise de décision dans l'entreprise.
- Définition de Bill Inmon (1996):
  « Une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision. »

  Principe: mettre en place une base de données utilisée à des fins d'analyse.

- Trois fonctions essentielles :
  - collecte de données de bases existantes et chargement
  - gestion des données dans l'entrepôt
  - analyse de données pour la prise de décision

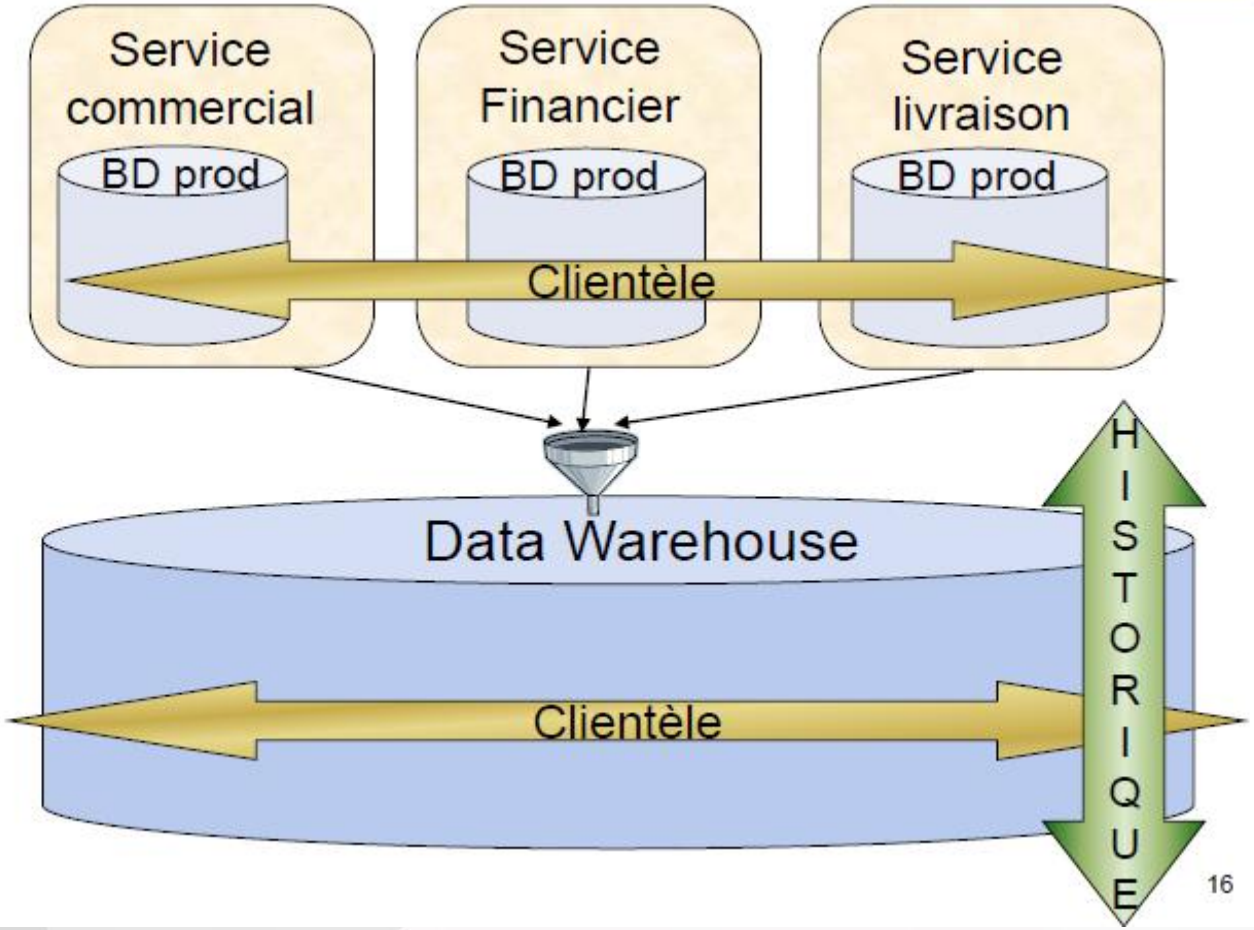
Concept

- **Trois composantes d'un entrepôt de données**



Concept

OLTP: On-Line Transactional Processing



OLAP: On-Line Analitical Processing

Concept

OLTP	ED
Orienté transaction	Orienté analyse
Orienté application	Orienté sujet
Données courantes	Données historisées
Données détaillées	Données agrégées
Données évolutives	Données stattiques
Utilisateurs nombreux, administrateurs/opérationnels	Utilisateurs peu nombreux, administrateur (manager)
Temps d’exécution court	Temps d’exécution long

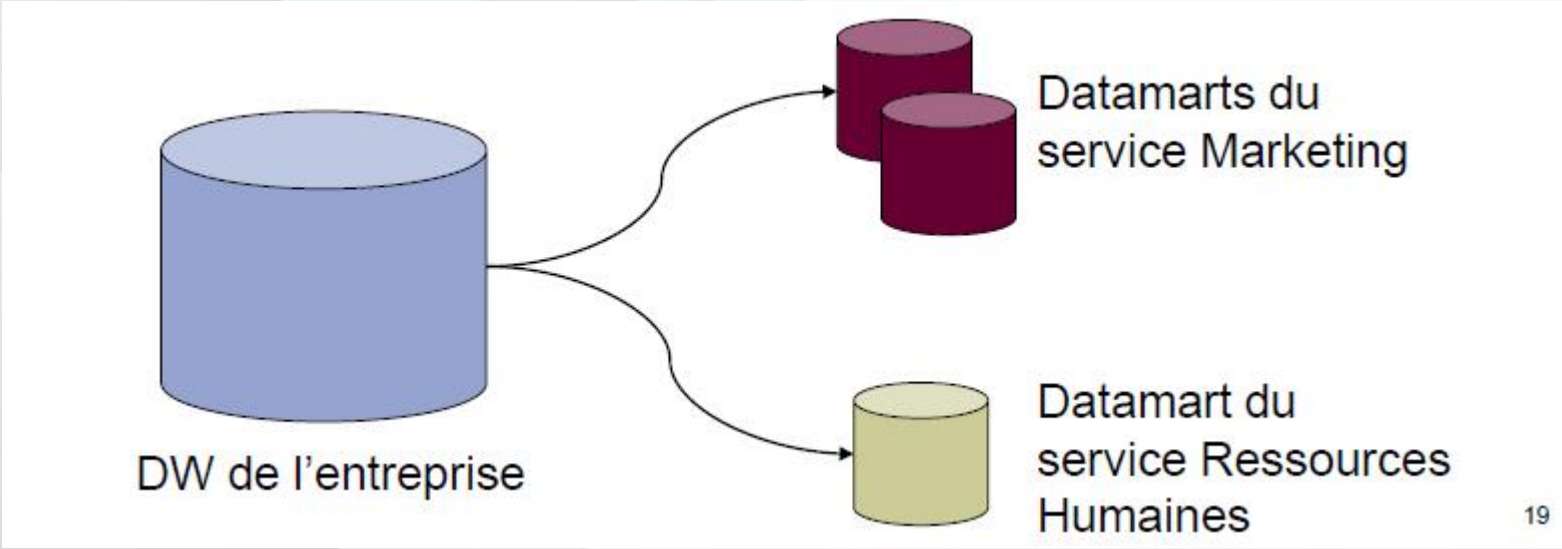
Concept

- Pourquoi une modélisation différente

Caractéristique	BD	ED
Opération	Opérations courantes, transactionnelles	Système d'Aide à la Décision Analyses stratégiques
Modèle	Entité-Association	Schéma en étoile
Redondance des données	À éviter	Permise
Données	Actuelles, brutes	Historiques, agrégées
# d'usagers	Plusieurs	Quelques uns
Mise à jour	Immédiate	Différée
Champs calculés	Aucun	Nombreux
Représentation mentale	Tabulaire	Hypercube
Requête	Simples, quelques enreg.	Complexe, beaucoup d'enreg.
Opérations	Lecture/Écriture	Lecture
Taille	Go (Gigaoctets)	To(TeraOctets)

Concept

- **DATAMART**
  - Sous-ensemble d'un entrepôt de données
  - Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
  - Point de vue spécifique selon des critères métiers



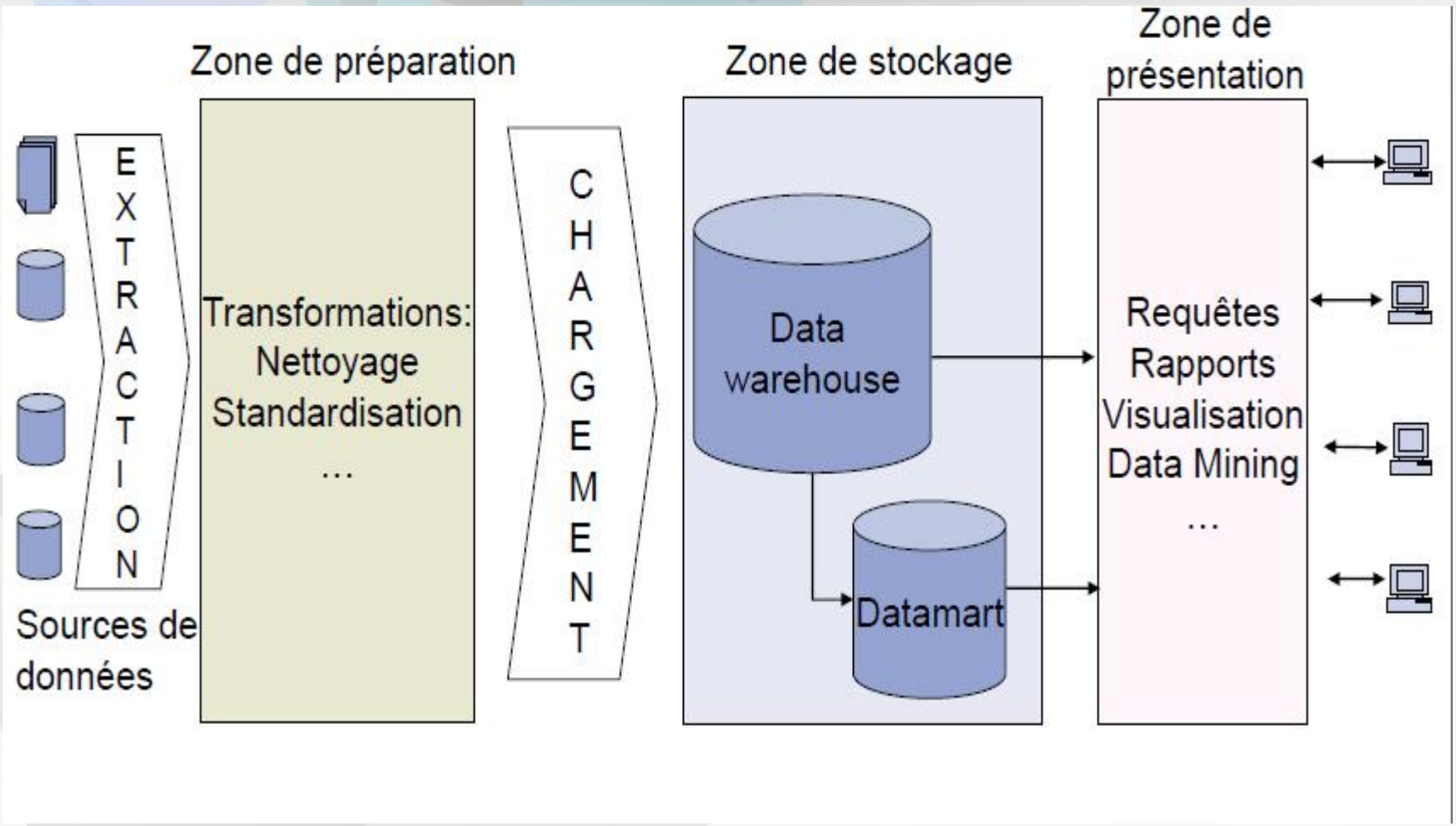
Concept

- **Intérêt des datamart(s)**
  - Nouvel environnement structuré et formaté en fonction des besoins d'un métier ou d'un usage particulier
  - Moins de données que DW
    - Plus facile èa comprendre, èa manipuler
    - Amélioration des temps de réponse
  - Utilisateurs plus ciblés: DM plus facile à définir

Entrepôt de données

Architecture générale

Architecture générale



Architecture générale

- Les flux de données
  - Flux entrant
    - Extraction: multi-source, hétérogène
    - Transformation: filtrer, trier, homogénéiser, nettoyer
    - Chargement: insertion des données dans l'entrepôt
  - Flux sortant
    - Mise à disposition des données pour les utilisateurs finaux

Architecture générale

- **Les différentes zones**
  - Zone de préparation (Staging area)
    - Zone temporaire de stockage des données extraites
    - Réalisation des transformations avant l'insertion dans le DW:
    - Nettoyage
    - Normalisation…
    - Données souvent détruites après chargement dans le DW
  - Zone de stockage (DW, DM)
    - On y transfère les données nettoyées
    - Stockage permanent des données
  - Zone de présentation
    - Donne accès aux données contenues dans le DW
    - Peut contenir des outils d'analyse programmés:
      - Rapports
      - Requêtes…

Entrepôt de données

Modélisation

Modélisation

- **Modélisation multidimensionnelle**
  - Souvent appelée OLAP (cold 1993) se présente comme une alternative au modèle relationnel.
  - Elle correspond au mieux aux besoins du décideur tout en intégrant la modélisation par sujet.
  - Méthode de conception logique qui vise à présenter les données non plus sous une forme de tables mais de cube centré sur une activité.
  - Un cube de dimension (n >3) est aussi dit hyper cube**.**

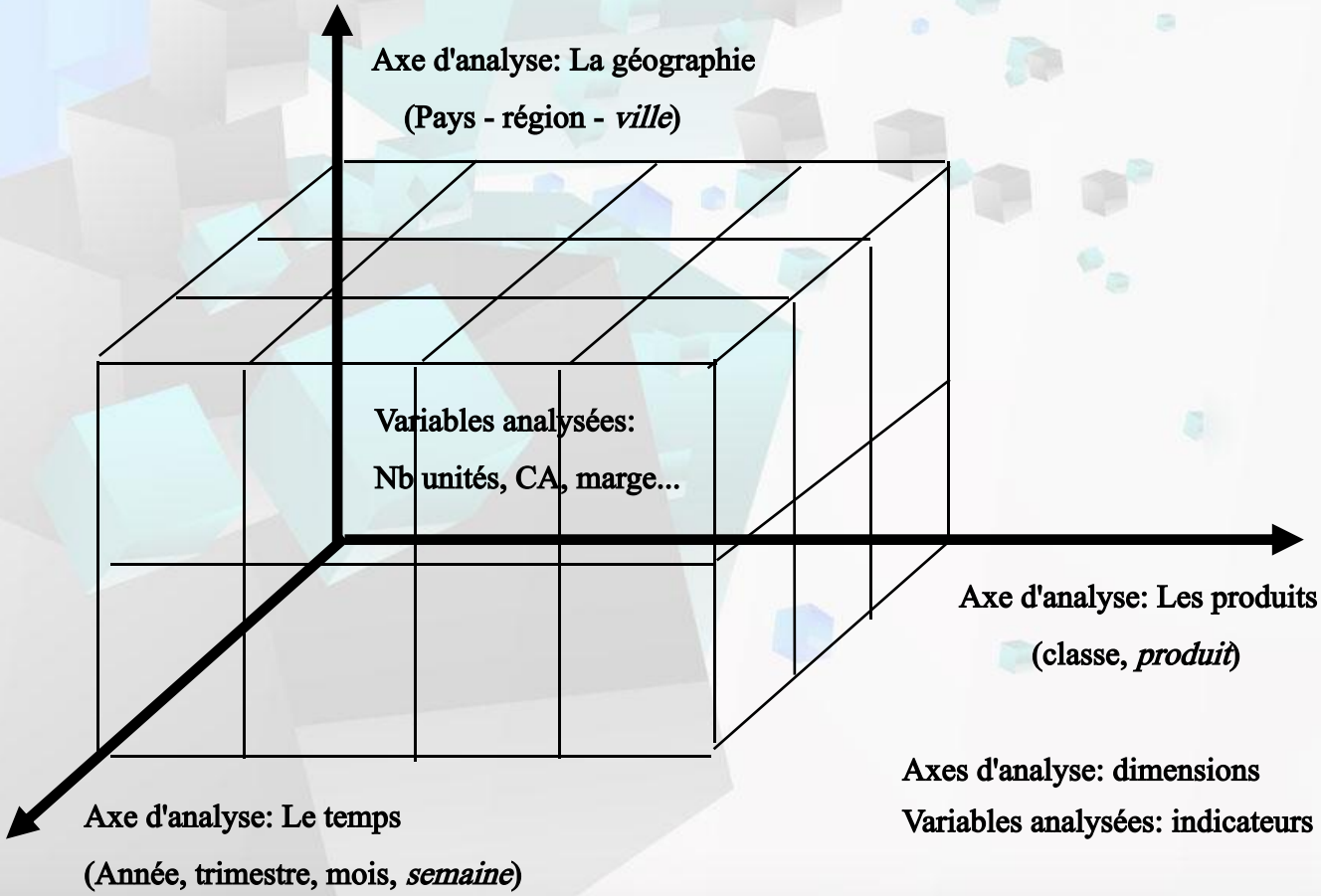
- **Modélisation multidimensionnelle**
  - Consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions.
  - Les données sont organisée de manière à mettre en évidence le sujet analysé et les différentes perspectives de l'analyse.
  - La modélisation multidimensionnelle a donné naissance aux concepts de fait et de dimension (Kimball 1996).

Modélisation

- Nouvelle méthode de conception autour des concepts métiers
  - Ne pas normaliser au maximum
- Introduction de nouveaux types de table:
  - Table de faits
  - Table de dimensions
- Introduction de nouveaux modèles:
  - Modèle en étoile
  - Modèle en flocon

Modélisation

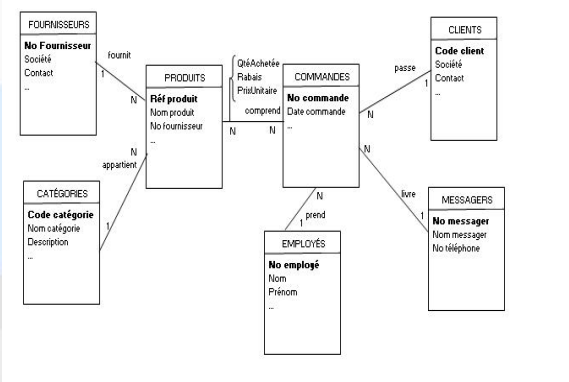
- Le cube et la dimension

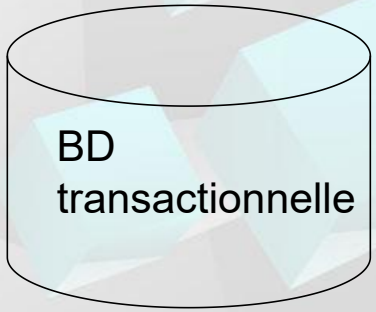


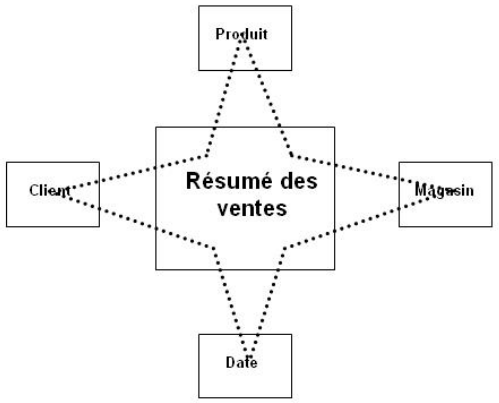
Axes d'analyse: dimensions

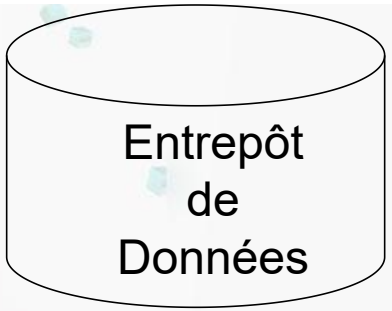
Variables analysées: indicateurs

Modélisation





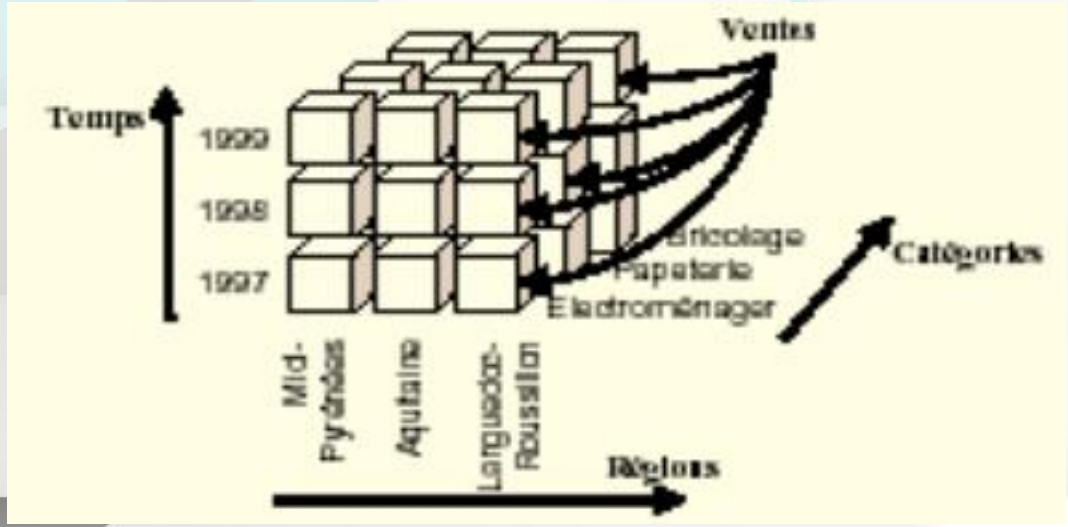




Modélisation par dimensions
Produire un ED dont la structure est facile à comprendre et à manipuler
Créer une bd dont l'interrogation est efficace.

Modélisation

- Considérons plusieurs tables, relatives aux ventes de chaque année entre 1997 et 1999. On peut alors observer les données dans un espace à 3 dimensions :
  - la dimension **catégories produit**
  - la dimension **régions**
  - la dimension **temps**
- Chaque intersection de ces dimensions représente une cellule comportant le montant des ventes :

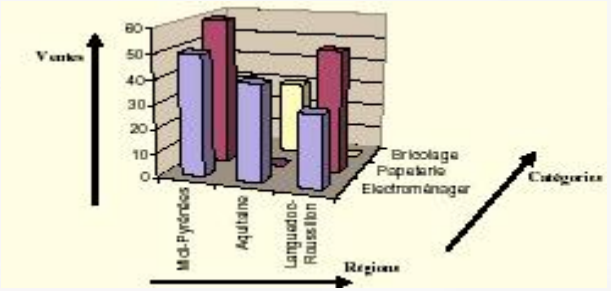


Modélisation

- Soit les données suivantes 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

- On peut distinguer différentes perspectives:
  - Une dimension relative à la catégorie des produits
  - Une dimension relative à la région

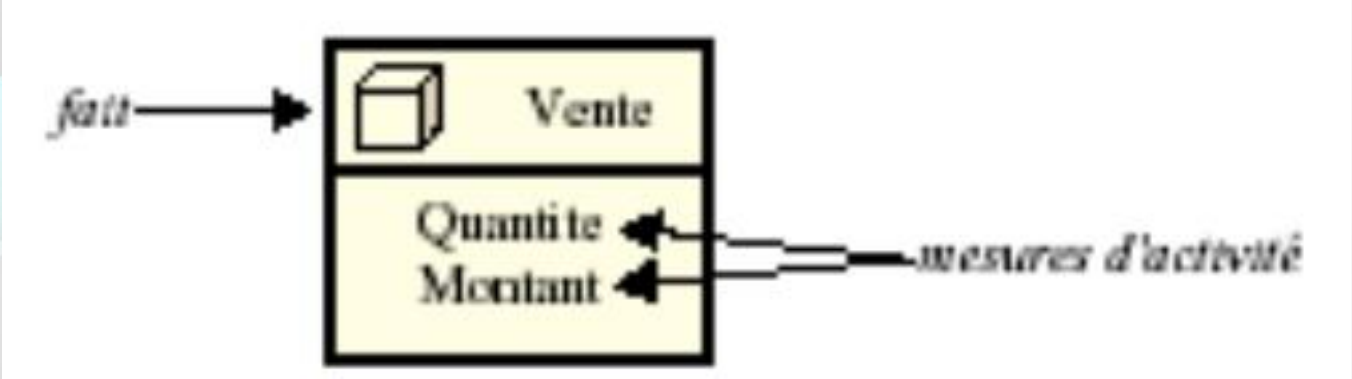


Modélisation

- **Le concept de fait :**
  - Ce que l'on souhaite mesurer
    - Quantités vendues, montant des ventes,…
  - 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.

Modélisation

- ***Exemple :*** le fait de « Vente » peut être constitué des mesures d'activités suivantes :
  - quantité de produits vendus et montant des ventes



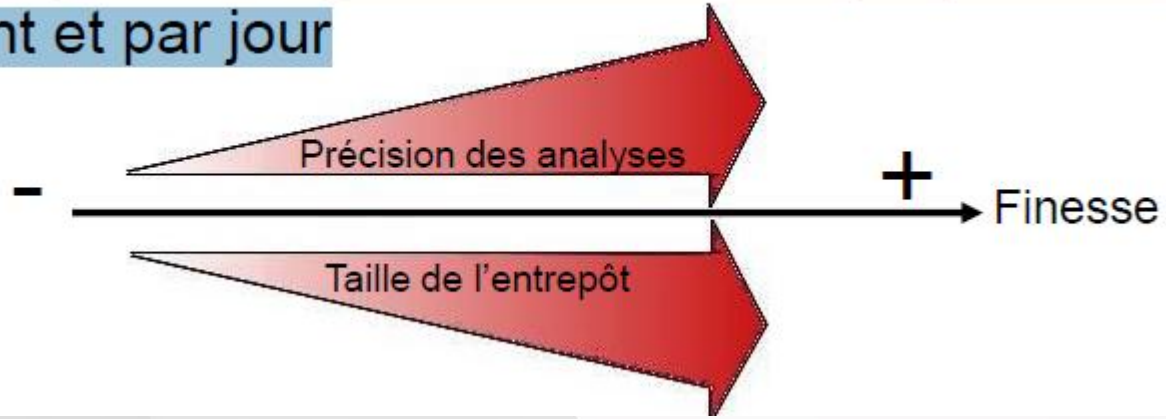
Modélisation

- **Table des faits**
  - Table principale du modèle dimensionnel
  - Contient les données observables (les fait) sur le sujet étudié selon les divers axes d'analyse (les dimensions)

	Table de faits des ventes
Clés étrangères vers les dimensions	Clé date (CE)
	Clé produit (CE)
	Clé magasin (CE)
Faits	Quantité vendue
	Coût
	Montant des ventes

Modélisation

- La granularité de la table de faits
  - La granularité définit le niveau de détails de la table de faits:
- Exemple: une ligne de commande par produit, par client et par jour

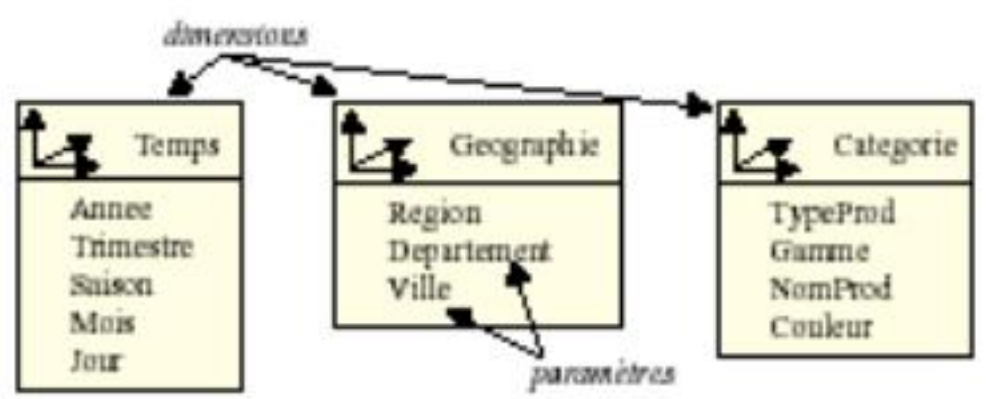


Modélisation

- **Le concept de dimension**
  - Le sujet analysé, c'est à dire le fait, est analysé suivant différentes perspectives correspondant à une catégorie utilisée pour caractériser les mesures d'activité analysées : on parle de dimensions.
- Dimension = axe d'analyse
  - Client, produit, période de temps…
  - Contient souvent un grand nombre de colonnes
  - L'ensemble des informations descriptives des faits
  - Contient en général beaucoup moins d'enregistrements qu'une table de faits

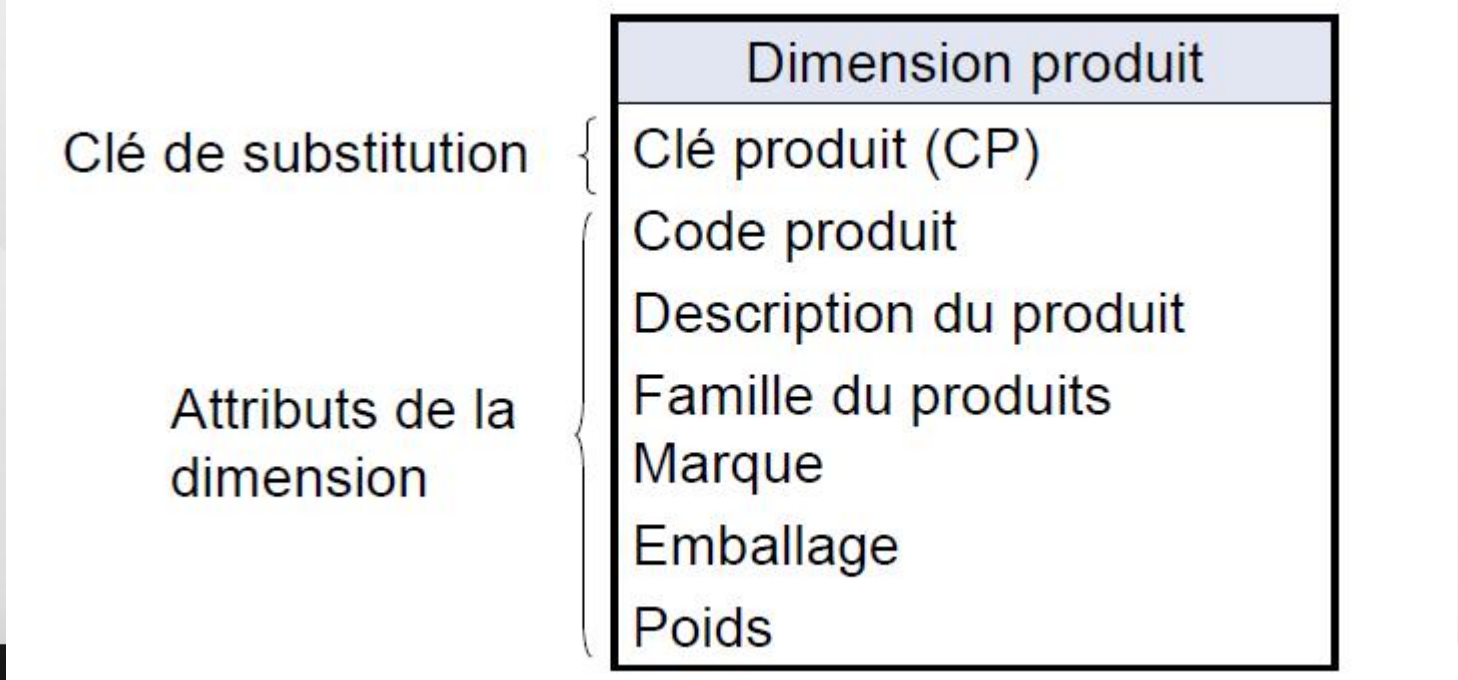
Modélisation

- ***Exemple :*** Dans l'exemple précédent, le fait « vente » peut être analysé suivant différentes perspectives correspondant à trois dimensions : la dimension ***Temps***, la dimension Géographie et la dimension catégories



Modélisation

- **Table de dimension**
  - Axe d'analyse selon lequel vont être étudiées les données observables (faits)
  - Contient le détail sur les faits.



Modélisation

- **La dimension temps**
  - Commune à l'ensemble du DW
  - Reliée à toute table de faits

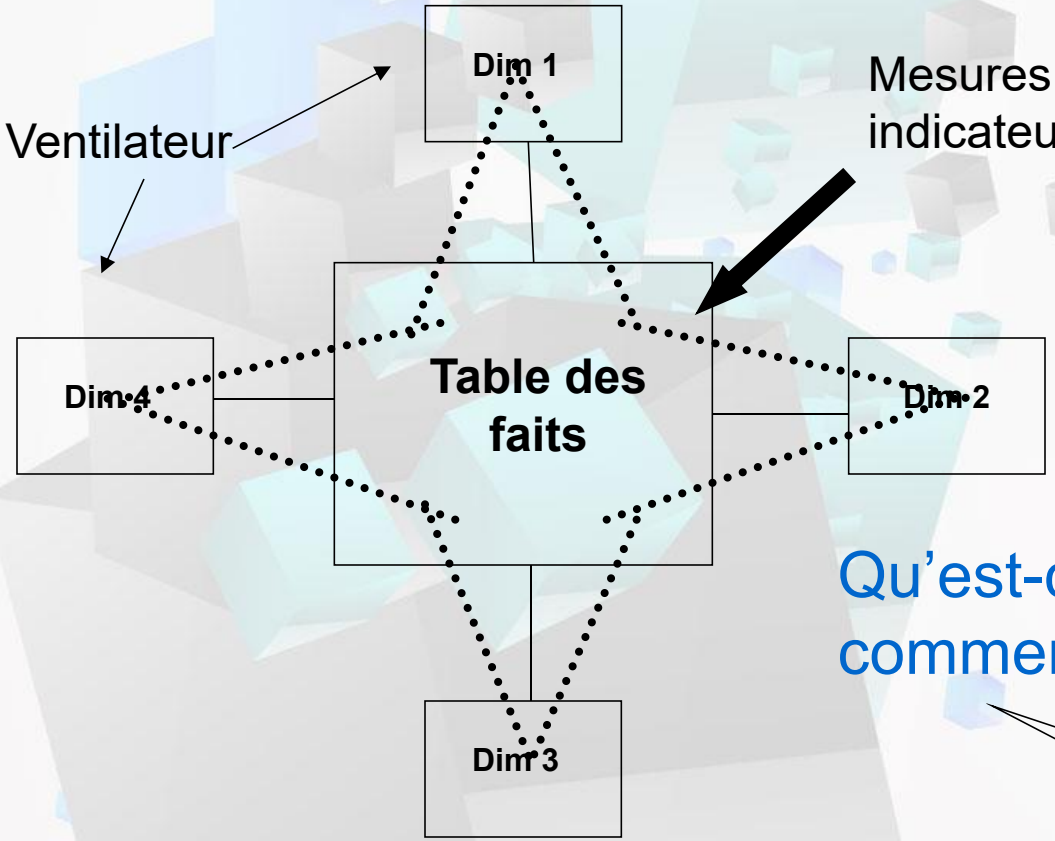
Dimension Temps
Clé temps (CP)
Jour
Mois
Trimestre
Semestre
Année
Num_jour_dans_année
Num_semaine_ds_année

Modélisation

- Granularité d'une table de dimension
  - Une dimension contient des membres organisés en hiérarchie :
    - Chacun des membres appartient à un niveau hiérarchique (ou niveau de granularité) particulier
    - Granularité d'une dimension : nombre de niveaux hiérarchiques
    - Temps :
      - année – semestre – trimestre - mois

Modélisation

- **Schéma en étoile répondant à un besoin d'affaires**

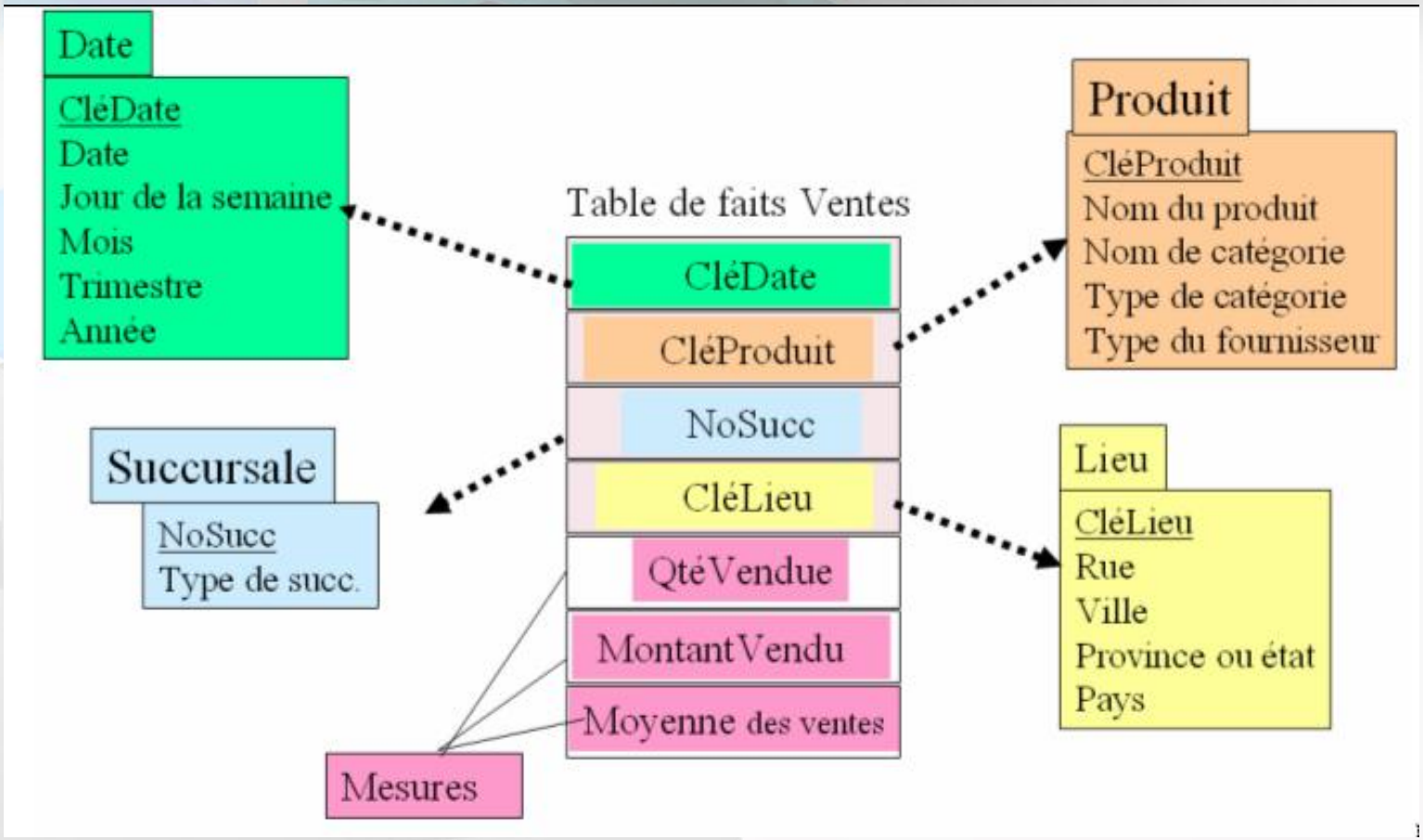


Mesures

Qu'est-ce que je veux voir et comment je veux le voir!

Dimensions

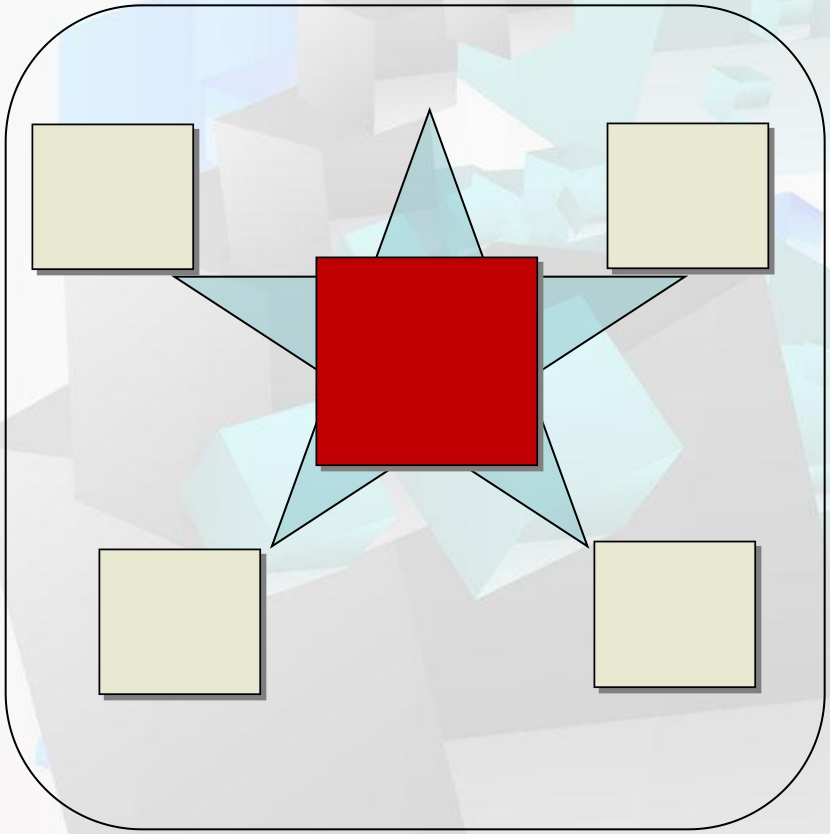
Un exemple



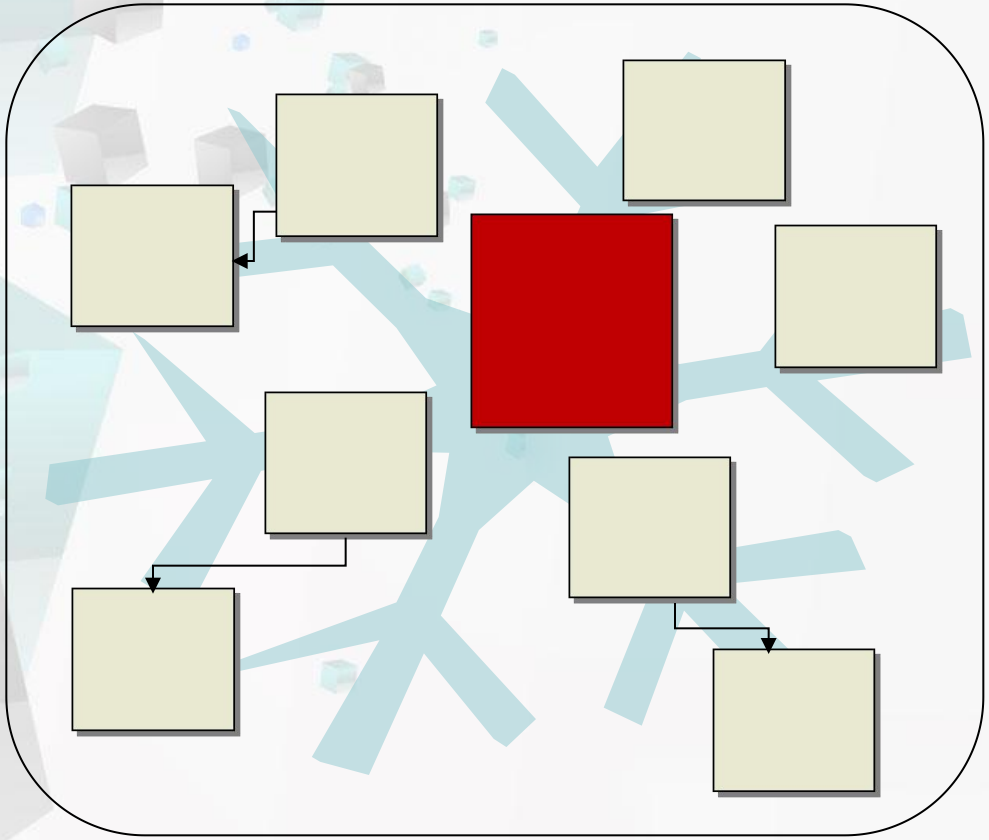
Entrepôt de données

Les types de modèles

Les types de modèle

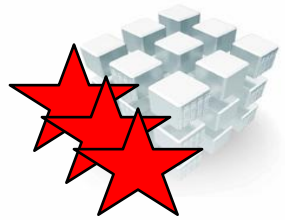


Modèle en étoile



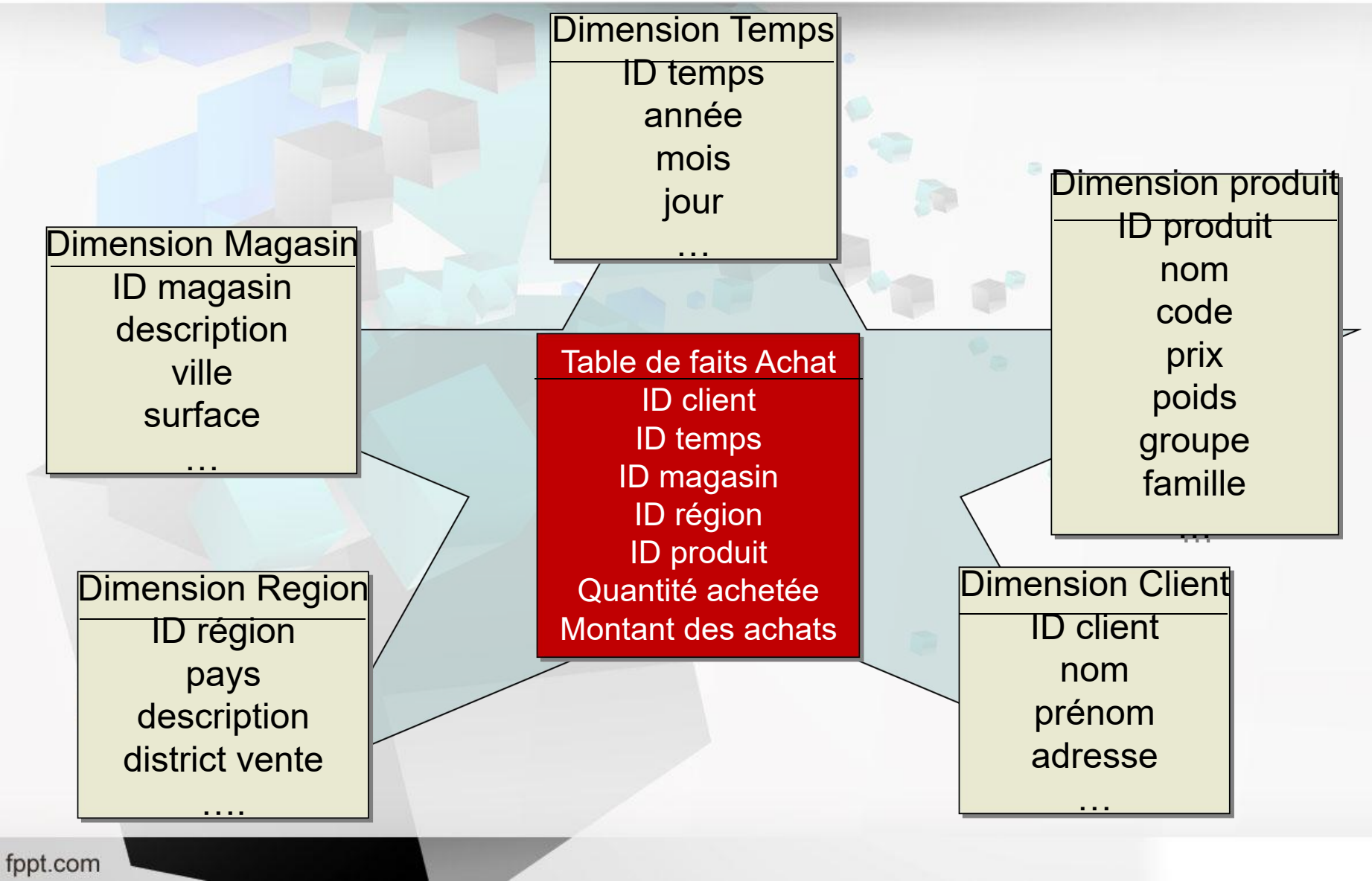
Modèle en flocon

Modèle en étoile

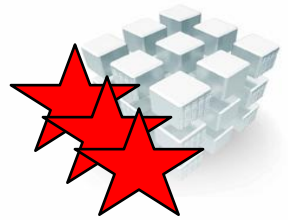


- Une table de fait centrale et des dimensions
- Les dimensions n’ont pas de liaison entre elles
- Avantages
  - Facilité de navigation
  - Nombre de jointures limité
- Inconvénients
  - Redondance dans les dimensions
  - Toutes les dimensions ne concernent pas les mesures

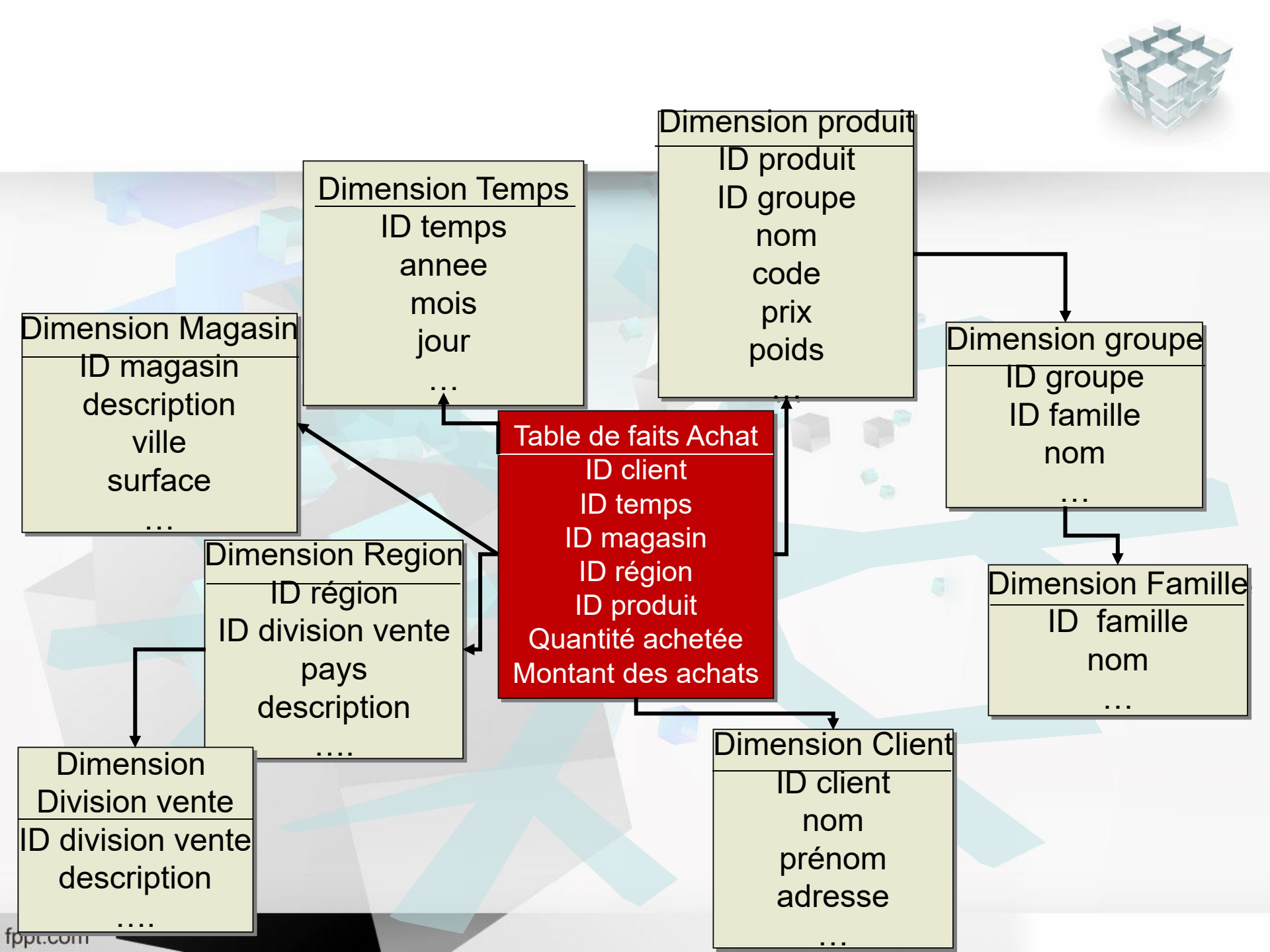
Modèle en étoile



Modèle en flocon

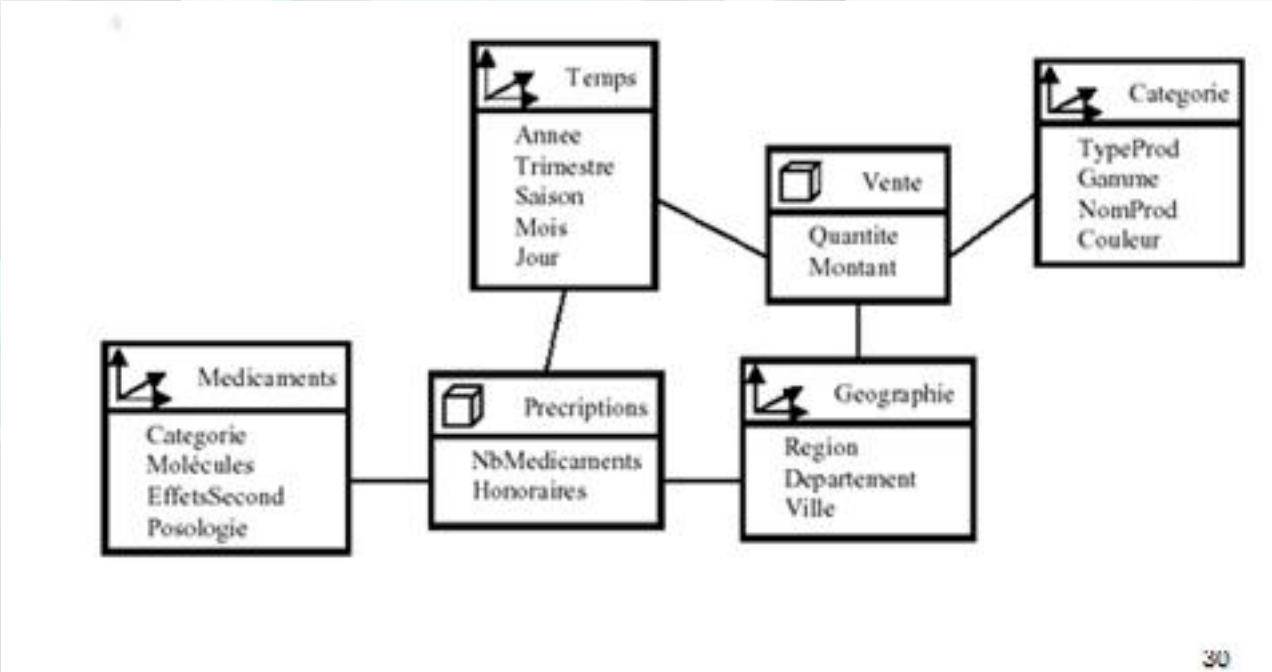


- Une table de fait et des dimensions décomposées en sous hiérarchies.
- On a un seul niveau hiérarchique dans une table de dimension.
- La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine.
- Avantages
  - Normalisation des dimensions
  - Économie d'espace disque
- Inconvénients
  - Modèle plus complexe (jointure)
  - Requêtes moins performantes


Dimension produit
ID produit
ID groupe
nom
code
prix
poids
…
Dimension Temps
ID temps
annee
mois
jour
…
Dimension Magasin
ID magasin
description
ville
surface
…
Dimension groupe
ID groupe
ID famille
nom
…
Table de faits Achat
ID client
ID temps
ID magasin
ID région
ID produit
Quantité achetée
Montant des achats
Dimension Region
ID région
ID division vente
pays
description
….
Dimension Famille
ID famille
nom
…
Dimension Division vente
ID division vente
description
….
Dimension Client
ID client
nom
prénom
adresse
…

Constellation

- On rassemble plusieurs tables de faits qui utilisent les mêmes dimensions



Méthodologie: 9 étapes de Kimball

1. **Choisir le sujet**
2. **Choisir la granularité des faits**
3. **Identifier et adapter les dimensions**
4. **Choisir les faits**
5. Stocker les pré-calculs
6. Établir les tables de dimensions
7. Choisir la durée de la base
8. Suivre les dimensions lentement évolutives
9. Décider des requêtes prioritaires, des modes de requêtes

Entrepôt de données

Alimentation/ mise à jour de l’entrepôt

Alimentation/ mise à jour de l'entrepôt

- Le processus d'alimentation consiste 'a :
  - Rassembler de multiples données sources souvent hétérogènes et les homogénéiser
- Entrepôt mis à jour régulièrement
  - Besoin d'un outil permettant d'automatiser les chargements dans l'entrepôt

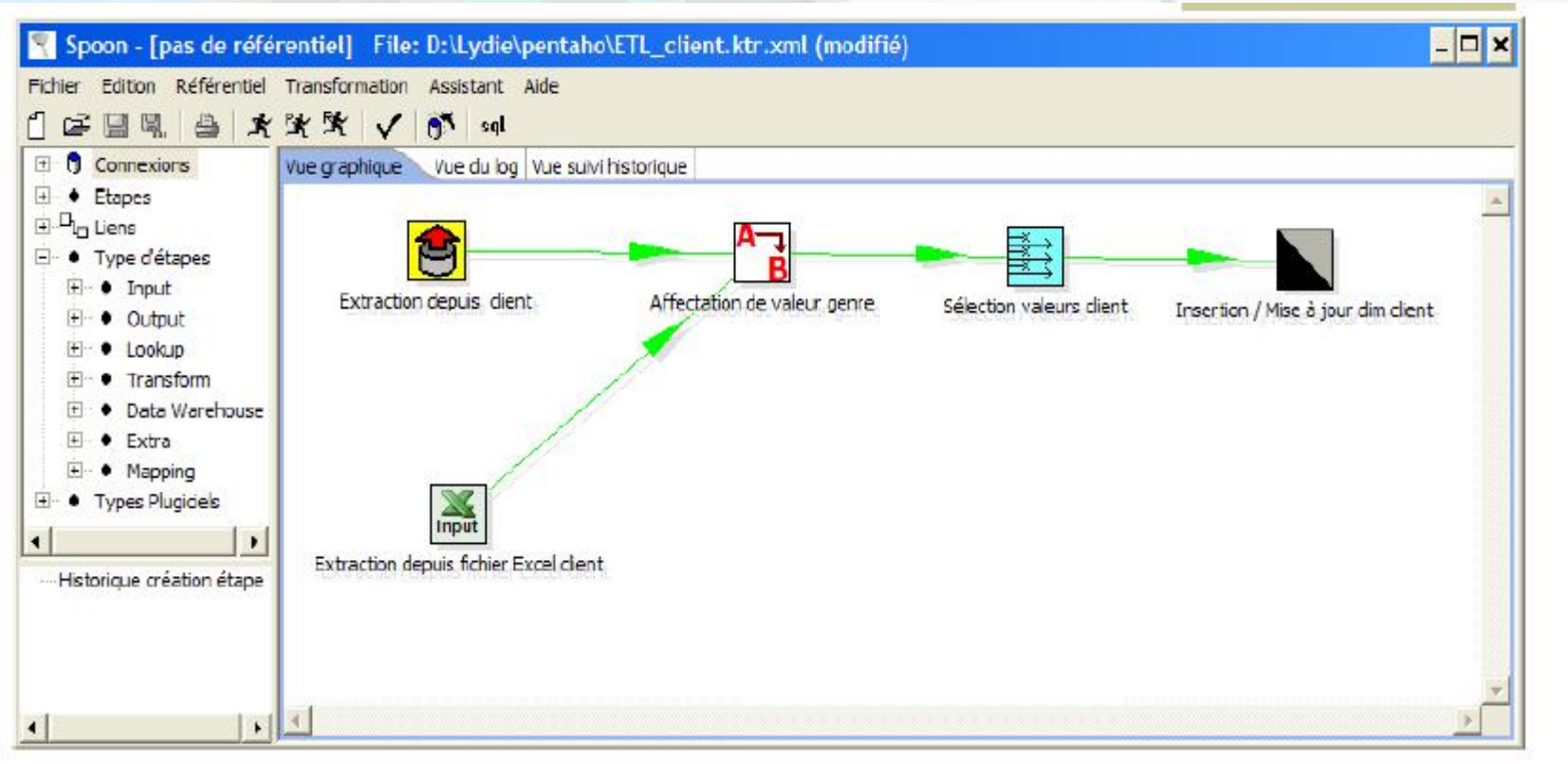
➢ Utilisation d'outils ETL (Extract, Transform, Load)

Alimentation/ mise à jour de l'entrepôt

- Définition d'un ETL
  - Offre un environnement de développement
  - Offre des outils de gestion des opérations et de maintenance
  - Permet de découvrir, analyser et extraire les données à partir de sources hétérogènes
  - Permet de nettoyer et standardiser les données
  - Permet de charger les données dans un entrepôt

Alimentation/ mise à jour de l'entrepôt

- Aperçu d'un ETL



Alimentation/ mise à jour de l'entrepôt

Activités	
Extraction	accès aux différentes sources
Nettoyage	recherche et résolution des inconsistances dans les sources
Transformation	entre différents format, langages, etc.
Chargement	des données dans l'entrepôt
Réplication	Des sources dans l'entrepôt
Analyse	Ex: détection de valeurs non valides ou inattendues
Tranfert des données haut débit	pour les très grands entrepôts
Tests de qualité	Ex: poiur correction et complétude
Analyse des méta données	Aide à la conception

Alimentation/ mise à jour de l'entrepôt

- **Extraction**
  - Extraire des données des systèmes de production
    - Dialoguer avec différentes sources:
    - Base de données,
    - Fichiers,
    - Bases propriétaires
  - Utilise divers connecteurs :
    - ODBC,
    - SQL natif,
    - Fichiers plats

Alimentation/ mise à jour de l'entrepôt

- **Tâches 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 ?*

Alimentation/ mise à jour de l'entrepôt

- **Tâches de sélection des données sources**
  - **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.

Alimentation/ mise à jour de l'entrepôt

- **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.*

Alimentation/ mise à jour de l'entrepôt

- **Types d'inconsistances**
  - Présence de données fausses dés leur saisie :
    - *fautes de frappe*
    - *différents formats dans une même colonne*
    - *texte masquant de l'information (ex: "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
  - Fonctions de normalisation et de conversion
  - Usage de dictionnaires de synonymes ou d'abréviation.

Alimentation/ mise à jour de l'entrepôt

- **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 (ex: "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

Alimentation/ mise à jour de l'entrepôt

- Transformation
- Rendre cohérentes les données des différentes sources
    - Transformer, nettoyer, trier, unifier les données
    - Exemple: unifier le format des dates (MM/JJ/AA JJ/MM/AA)
  - Étape très importante, garantit la cohérence et la fiabilité des données

Alimentation/ mise à jour de l'entrepôt

- 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 :
    - *Équivalence de champs*
    - *Équivalence d'enregistrements* : fusion d'enregistrements

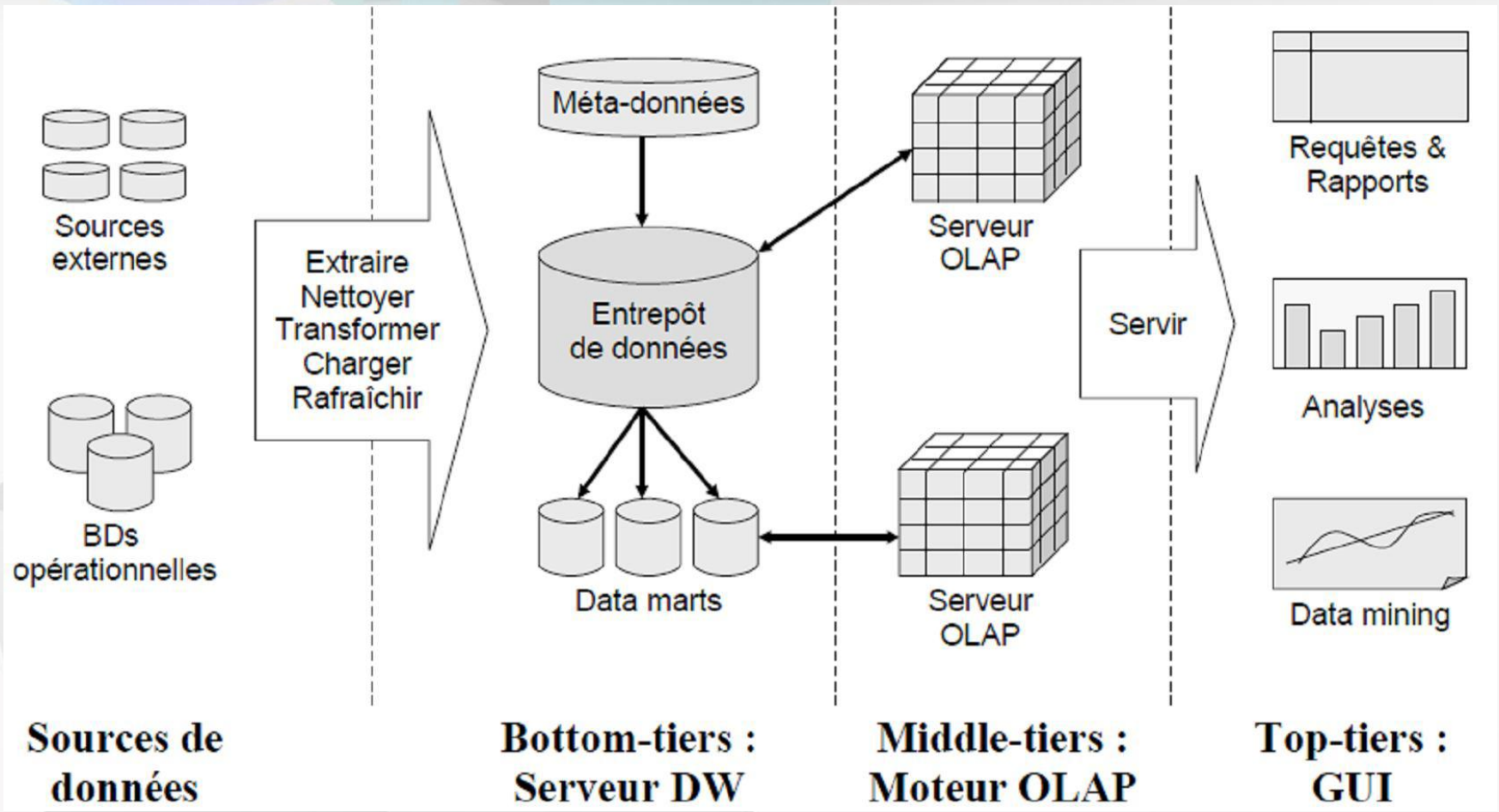
Alimentation/ mise à jour de l'entrepôt

- Chargement
  - Insérer ou modifier les données dans l'entrepôt
    - Utilisation de connecteurs:
    - ODBC,
    - SQL natif,
    - Fichiers plats
  - Opération qui risque d'être assez longue
  - Plutôt mécanique et la moins complexe
  - Il est nécessaire de définir et de mettre en place
    - Des stratégies pour assurer de bonnes conditions à sa réalisation
    - Une politique de rafraîchissement

Entrepôt de données

Les bases de données multidimensionnelles

Les bases de données multidimensionnelle

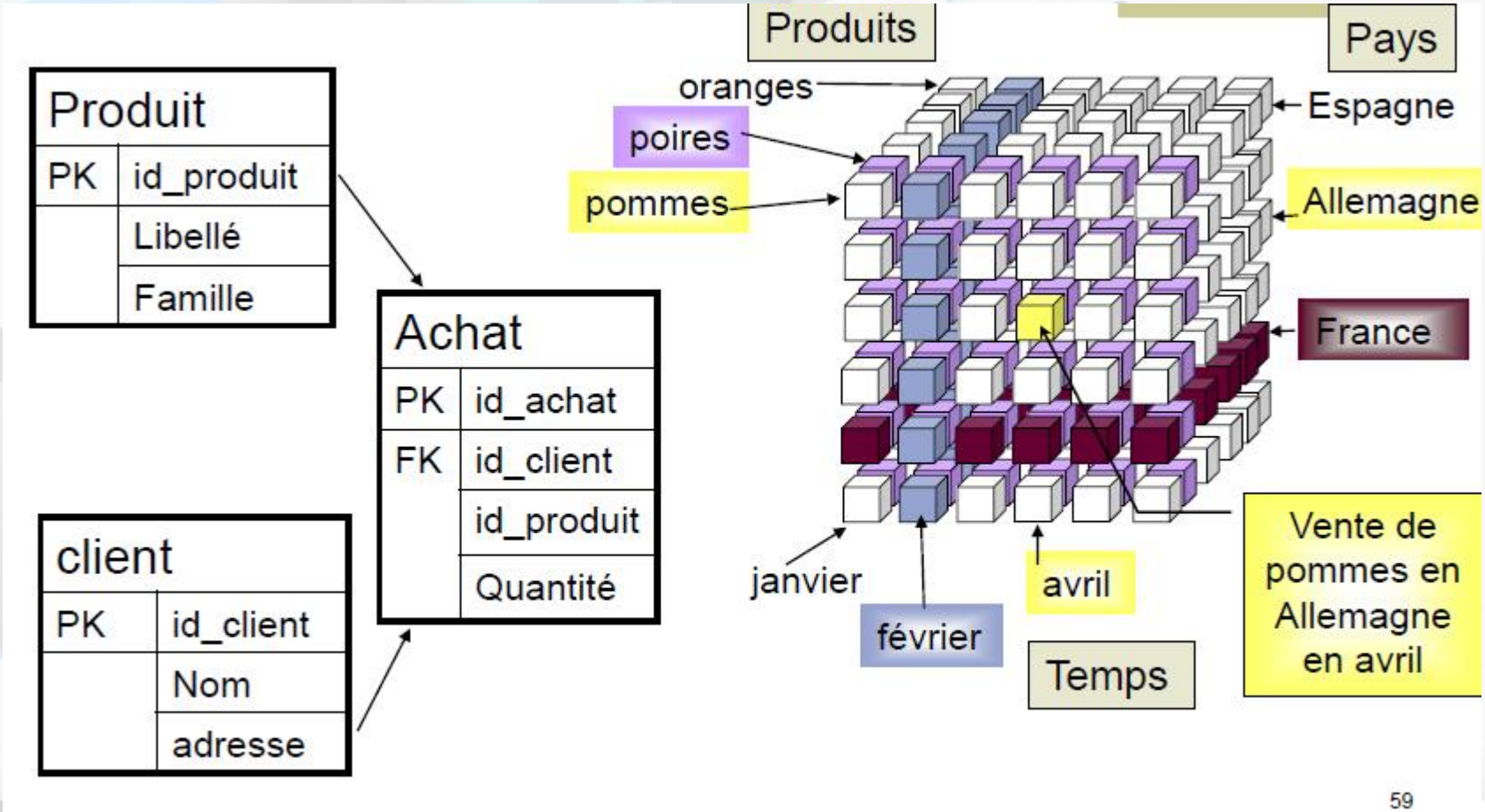


Les bases de données multidimensionnelles

- **Entrepôt de données (DataWare house) vs BD opérationnelle**
  - **OLTP (On-Line Transaction Processing)**
    - Exécution en temps réel des transactions, pour l'enregistrement des opérations quotidiennes : inventaires, commandes, paye, comptabilité
    - Par opposition au traitement en batch
  - **OLAP (On-Line Analytical Processing)**
    - Traitement efficace des requêtes d'analyse pour la prise de décision qui sont par défaut assez complexes (bien qu'a priori, elles peuvent être réalisées p.ar les SGBD classiques)

Les bases de données multidimensionnelle

- OLTP versus OLAP



Les bases de données multidimensionnelles

- **ROLAP**
  - Relational OLAP
    - Données stockées dans une base de données relationnelles
    - Un moteur OLAP permet de simuler le comportement d'un SGBD multidimensionnel
  - Plus facile et moins cher à mettre en place
  - Moins performant lors des phases de calcul
  - Exemples de moteurs ROLAP:
    - Mondrian

Les bases de données multidimensionnelle

- **MOLAP**
  - Multi dimensional OLAP:
    - Utiliser un système multidimensionnel « pur » qui gère les structures multidimensionnelles natives (les cubes)
    - Accès direct aux données dans le cube
  - Plus difficile à mettre en place
  - Formats souvent propriétaires
  - Conçu exclusivement pour l'analyse multidimensionnelle
  - Exemples de moteurs MOLAP:
    - Microsoft Analysis Services
    - Hyperion

Les bases de données multidimensionnelle

- HOLAP
  - Hybride OLAP:
    - tables de faits et tables de dimensions stockées dans SGBD relationnel (données de base)
    - données agrégées stockées dans des cubes
  - Solution hybride entre MOLAP et ROLAP
  - Bon compromis au niveau coût et performance

Les bases de données multidimensionnelle

- **Le cube**
  - Modélisation multidimensionnelle des données facilitant l'analyse d'une quantité selon différentes dimensions:
    - Temps
    - Localisation géographique
    - …
  - Les calculs sont réalisés lors du chargement ou de la mise à jour du cube

Entrepôt de données

Manipulation des données multidimensionnelles

Manipulation des données

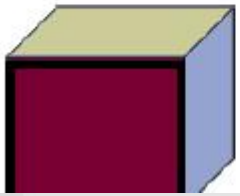
- **MDX (Multidimensional Expressions)**
  - Langage permettant de définir, d'utiliser et de récupérer des données à partir d'objets multidimensionnels
    - Permet d'effectuer les opérations décrites précédemment
  - Equivalent de SQL pour le monde OLAP
  - Origine: Microsoft
  - Opérations ensembliste possibles sur un cube:
    - Rotate
    - Slicing
    - Dicing
    - Scoping
    - Drill-up / down

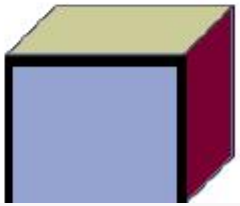
Manipulation des données

- **Opération agissant sur la structure**
  - Rotation (rotate): présenter une autre face du cube

	05	06	07
Œuf	221	263	139
Viande	275	257	116

	05	06	07
Idf	101	120	52
Ain	395	400	203



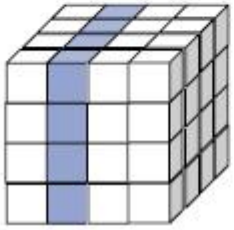


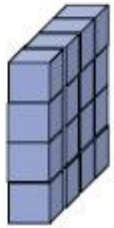
Manipulation des données

- **Opération agissant sur la structure**
  - Tranchage (slicing): consiste à ne travailler que sur une tranche du cube. Une des dimensions est alors réduite à une seule valeur

		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240
Viande	Idf	163	152	145
	Ain	187	174	184

		06
Œuf	Idf	265
	Ain	245
Viande	Idf	152
	Ain	174





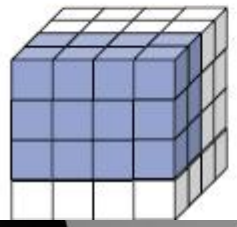
Manipulation des données

- **Opération agissant sur la structure**
  - Extraction d'un bloc de données (dicing): ne travailler que sous un sous-cube

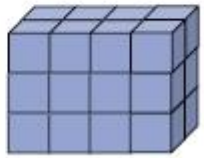
		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240
Viande	Idf	163	152	145
	Ain	187	174	184

→

		05	06	07
Œuf	Idf	220	265	284
	Ain	225	245	240



→

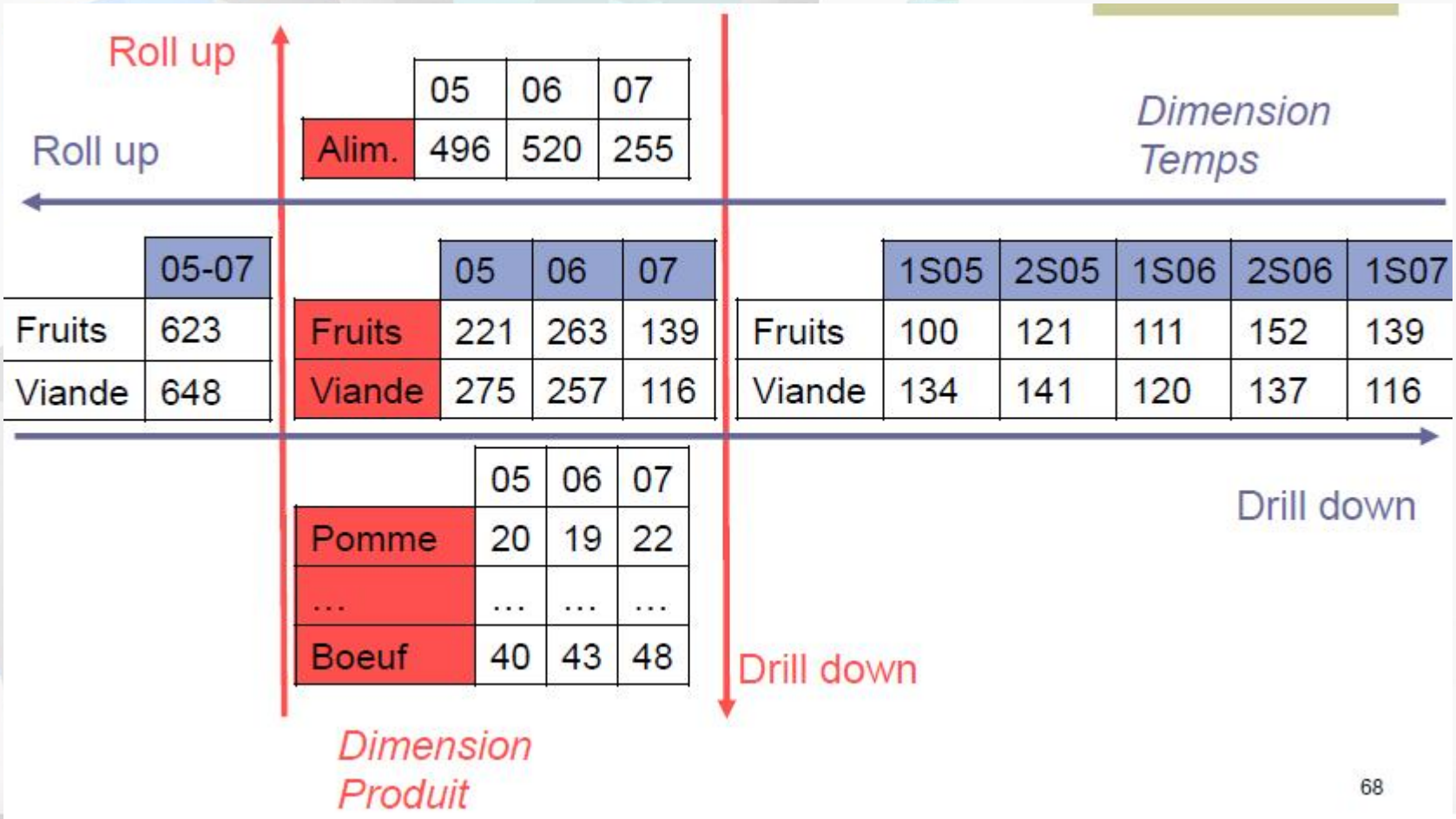


Manipulation des données

- **Opération agissant sur la granularité**
  - Forage vers le haut (roll-up): « dézoomer »
    - Obtenir un niveau de granularité supérieur
    - Utilisation de fonctions d'agrégation
  - Forage vers le bas (drill-down): « zoomer »
    - Obtenir un niveau de granularité inférieur
    - Données plus détaillées

Manipulation des données

- **Drill-up, drill-down**



	05	06	07
Alim.	496	520	255

	05-07
Fruits	623
Viande	648

	05	06	07
Fruits	221	263	139
Viande	275	257	116

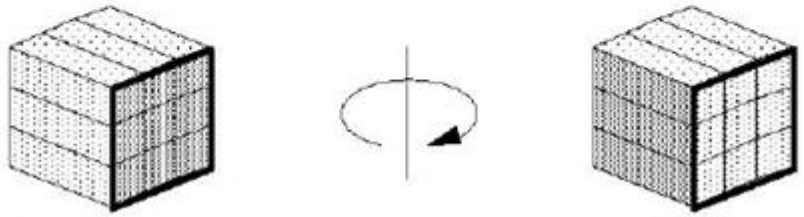
	1S05	2S05	1S06	2S06	1S07
Fruits	100	121	111	152	139
Viande	134	141	120	137	116

	05	06	07
Pomme	20	19	22
…	…	…	…
Boeuf	40	43	48

- Rotate

	95	96	97
Frais	221	263	139
Liquide	275	257	116

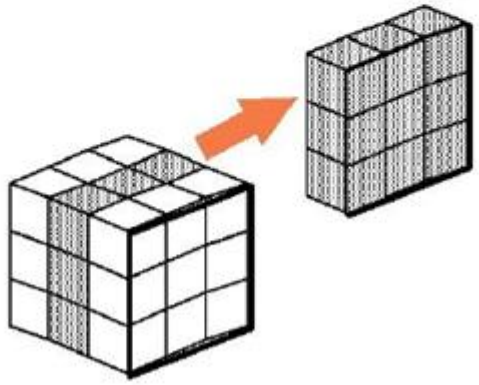
	95	96	97
NordPdC	101	120	52
IdF	395	400	203



- Slicing

		1995	1996	1997
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

		1996
Frais	IdF	265
	Province	245
Liquide	IdF	152
	Province	174



Manipulation des données multidimensionnelles

- **MDX (exemple)**
  - Fournir les effectifs d'une société pendant les années 2004 et 2005 croisés par le type de paiement

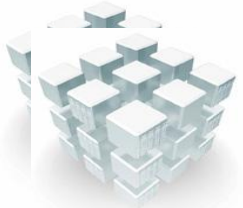
```
SELECT {([Time].[2004]), ([Time].[2005])} ON COLUMNS,
{[Pay].[Pay Type].Members} ON ROWS
FROM RH
WHERE ([Measures].[Count])
```

Cube

Dimensions, axes d'analyse

	2004	2005
Heure	3396	4015
Jour	3678	2056

Manipulation des données



- **Visualisation autour d'un entrepôt de données**
  - Les techniques de visualisation des données doivent faciliter leur analyse et leur interprétation
  - Les techniques de visualisation:
    - convertissent des données complexes en images, graphiques en 2 et 3 dimensions, et en animations
    - qui peuvent être analysées en cherchant des interrelations entre données
  - Elles sont de plus en plus intégrées dans les ED.

Manipulation des données multidimensionnelles

- **Fouille de données (Data Mining) :**
  - **Recherche de connaissance**, sous forme de **modèle de comportement**, **cachés** dans les **données**
  - Domaine jeune à l'intersection de l'Intelligence Artificielle, les Statistiques, les Bases de données
  - **Nombreuses techniques de fouille** : régression linéaire, induction d'arbres de décision, algorithmes génériques, réseaux de neurones, …
  - Les **techniques de fouille sont en pleine évolution** et sont de plus en plus intégrées dans les entrepôts de données.