

Business Intelligence (BI)

L'informatique Décisionnelle

1- Les entrepôts de données

Sonia GUEHIS
Sonia.guehis@parisnanterre.fr

Contexte

Prise de décisions stratégiques et politiques au sein des entreprises par des décideurs souvent *non-informaticiens* :

- Quelles sont les ventes du produit X pendant le trimestre A de l'année B dans la région C ?
- Comment se comporte le produit X par rapport au produit Y?
- Quel type de client peut acheter le produit X?
- Est-ce qu'une baisse de prix de 10% par rapport à la concurrence ferait redémarrer les ventes du produit X

Contexte

Les outils traditionnels de gestion et d'exploitation des données sont du type **transactionnel** ou **OLTP (On-Line Transaction Processing)**

- L'exploitation de données centrée sur la saisie, le stockage, la mise à jour, la sécurité et l'intégrité des données.
- Le système transactionnel est développé pour gérer les transactions quotidiennes
- Ces bases de données supportent habituellement des applications particulières telles que les inventaires de magasins, les réservations d'hôtel, etc
- Le contenu est constitué de données présentes en temps réel, pas d'archives
- Les données sont très détaillées (détails de chacune des transactions)
- La mise à jour s'effectue par de nouvelles transactions
- Très souvent plusieurs de ces systèmes existent indépendamment les uns des autres.

Business Intelligence

Le terme décisionnel « Business Intelligence » couvre l'ensemble des technologies permettant en bout de chaîne d'apporter une aide à la décision.

- Capable d'agréger les données internes ou externes et de les transformer en information servant à une prise de décision rapide.
- Capable de répondre à certains types de questions décisionnels.
- Les questions doivent pouvoir être formulées dans le langage de l'utilisateur en fonction de son secteur d'activité: marketing, service clients, économique...

Applications transactionnelles Vs Applications décisionnelles

- Les applications transactionnelles sont constituées de traitements factuels de type OLTP (On Line Transaction Processing)
- Les applications d'aide à la décision sont constituées de traitements ensembliste de type OLAP: On Line Analytical Processing
- Les deux activités (OLTP & OLAP) ne peuvent co-exister sur des données dans le même système d'information: leurs objectifs de performance sont exactement opposés:
 - Les requêtes complexes et lourdes dégradent les performances des systèmes transactionnels
 - Les données temporelles sont réparties entre données actuelles et données archivées, rendant la vue historique des données très difficile ou impossible
 - Le support efficace d'une activité OLAP nécessite la constitution d'un système d'information propre: Le **Datawarehouse ou entrepôt de données**

Entrepôt de données

L'entrepôt de données doit :

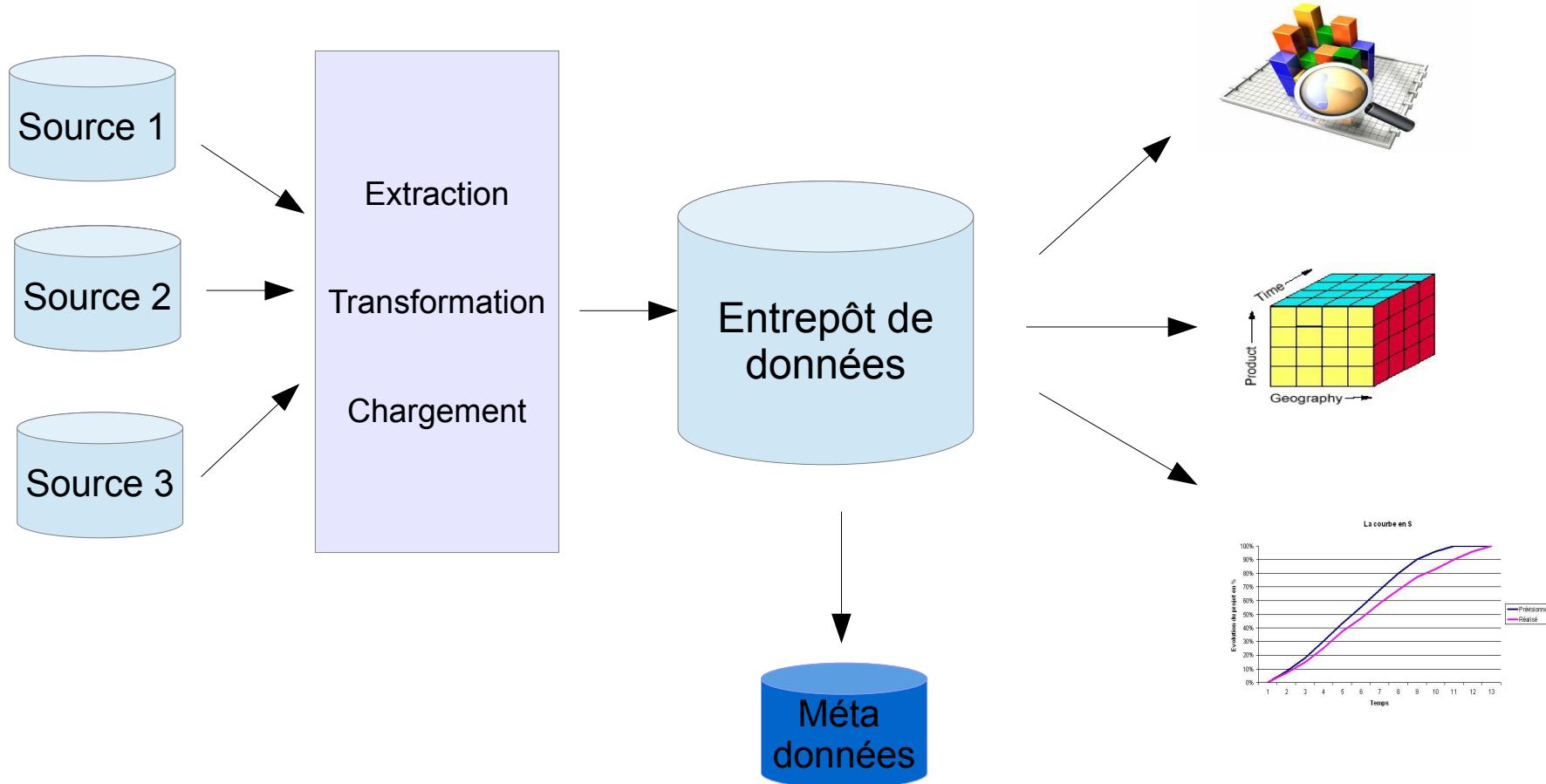
- * **Rendre accessibles les informations de l'entreprise.** Le contenu de l'entrepôt doit être compréhensible et l'utilisateur doit pouvoir naviguer facilement et avec rapidité.
- * **Rendre cohérente l'information d'une entreprise** Les informations provenant d'une branche de l'entreprise peuvent être mises en corrélation avec celles d'une autre branche.
- * **Constituer une source d'information souple et adaptable** L'entrepôt de données est conçu dans la perspective de modifications perpétuelles.
- * **Représenter un bastion sécurisé qui protège le capital information.** L'entrepôt de données ne contrôle pas seulement l'accès aux données, mais il offre à ses gestionnaires une bonne visibilité des utilisations, bonnes et mauvaises, des données.
- * **Constituer la base décisionnelle de l'entreprise** L'entrepôt de données recèle en son sein les informations propres à faciliter la prise de décisions.

Entrepôt de données

Définition (W. H. Inmon) :

« Le Data Warehouse est une collection de données
orientées sujet,
intégrées,
non volatiles et
historisées,
organisées pour le support d'un processus
d'aide à la décision. »

Flux de Données



Flux de données

➤ Flux entrant (ETL):

- Extraction des données de gestion :
 - multi-sources
 - hétérogènes
- Transformation avant insertion dans l'entrepôt:
 - filtrer
 - homogénéiser
 - nettoyer
 - Agréger

➤ Flux sortant :

- Mise à disposition des données
- Tableaux de bord
- Cubes OLAP
- Rapports d'analyse ...
- Les métadonnées

Nettoyage des données

- Ajout de règles d'intégrité pour :
 - identifier et corriger les valeurs invalides
 - identifier (et remplacer) les valeurs manquantes
 - traiter les violations de contraintes d'intégrité
- Ajouter des heuristiques pour :
 - trouver si deux entités représentent la même information,
 - pour ajouter des attributs à certaines entités

Les métadonnées

- Les métadonnées apportent des précisions et informations utiles concernant:
 - Quelles sont les données «entreposées», leur format, leur signification
 - La validité, précision, historique de données
 - les processus de récupération/extraction dans les bases sources
 - la date du dernier chargement de l'entrepôt
 - l'historique des données sources et de celles de l'entrepôt
 - les liens entre les données opérationnelles et celles de l'entrepôt

Importance des métadonnées

- Pour l'utilisateur :
 - évite l'interprétation et facilite l'usage
 - donne des informations sur
 - Le contexte métier de l'utilisateur et l'usage des indicateurs
 - La source des informations et les règles de gestion métier
- Pour l'administrateur :
 - évite les erreurs et facilite la gestion
 - procure des informations sur
 - La description des serveurs, BD, règles de gestion
 - Les volumétries des informations et dates de rafraîchissement
 - Les traces d'activité, temps de chargement, erreurs ...

Administration

- Conduite du chargement des données à partir de sources multiples
- Tests d'intégrité et de qualité des données
- Construction et gestion des méta-données
- Performances de l'entrepôt : réponse aux requêtes et utilisation des ressources
- Audit d'usage de l'entrepôt pour retour d'information vers usagers
- RéPLICATION, calculs, répartition des données
- Gestion efficace des données stockées
- Purge des données
- Archivage et back-up
- Reprise en cas de panne
- Sécurité

Magasins de données = Data Mart

Définition :

Sous-ensemble d'un entrepôt de données destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise : ex service marketing, service gestion de personnel...

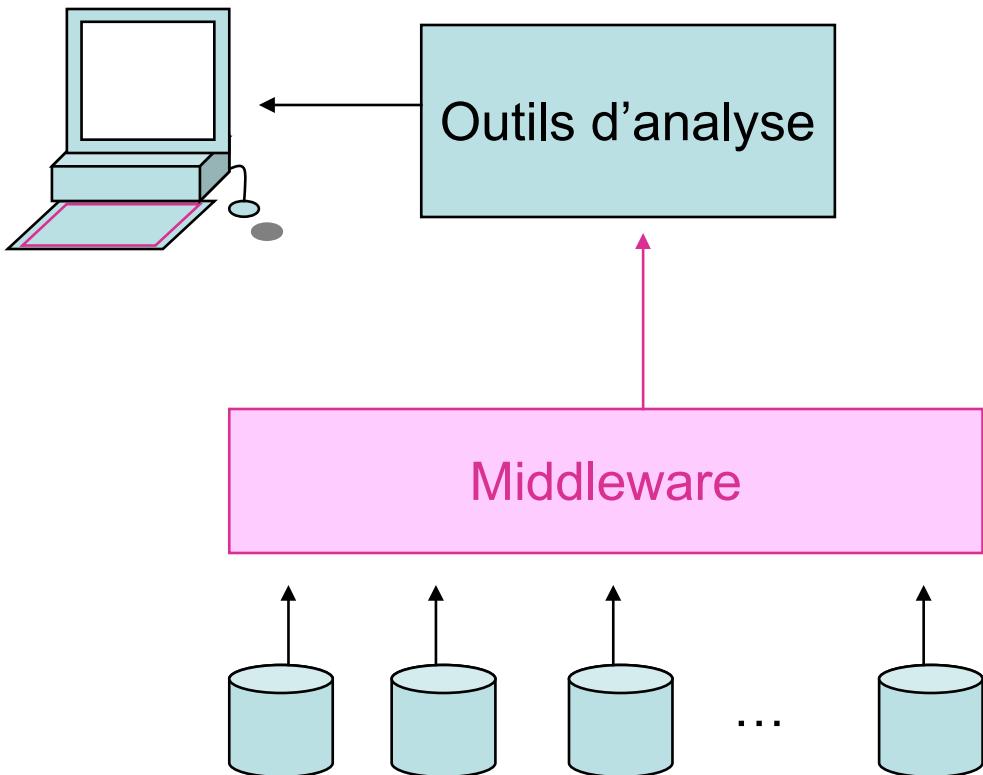
- + Pas de données opérationnelles détaillées (en général)
- + Moins de données que l'entrepôt
 - => Plus facile à comprendre, à manipuler
- + Plus performant
- Possibilité de se retrouver avec plusieurs data marts non corrélés et sans liens.
- Résultats limités étant données l'absence de vue global sur l'ensemble de l'activité de l'entreprise.

Pourquoi des magasins de données

- Utilisateurs plus ciblés => plus facile à définir
- Accès aux données particulières
- Vue collective des données à un service sous une forme adaptée à ses besoins et aux traitements à faire (OLAP, data-mining...)
- Réduction du volume des données => Améliorer les temps de réponse
- Coût d'implantation plus faible que l'entrepôt
- Temps de mise en œuvre plus court

Architectures d'entrepôts de données

1. L'entrepôt virtuel (ou le non-entrepôt)

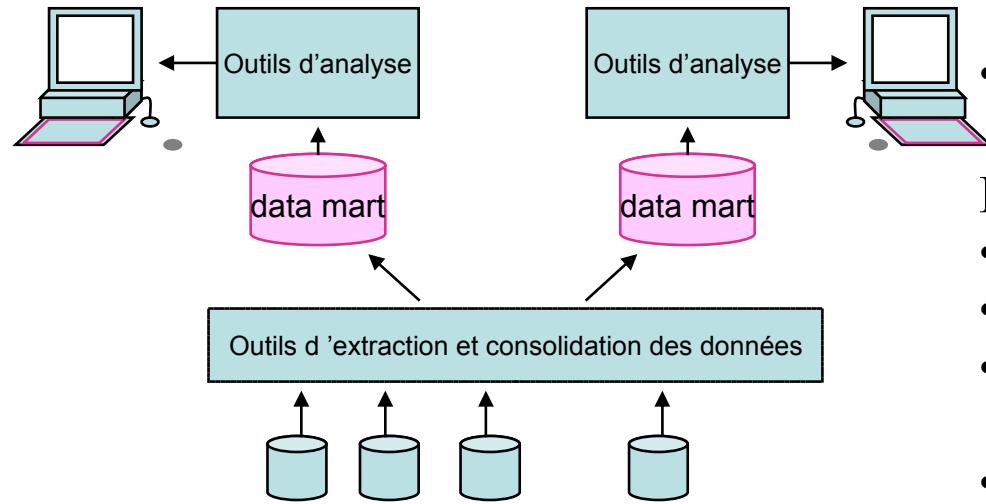


Inconvénients:

- pas de réelle intégration des données
- différentes vues non-réconciliées
- pas d'historique
- les requêtes peuvent facilement bloquer

Architectures d'entrepôts de données

2. Les magasins de données adhoc : Data Marts indépendants



Avantages:

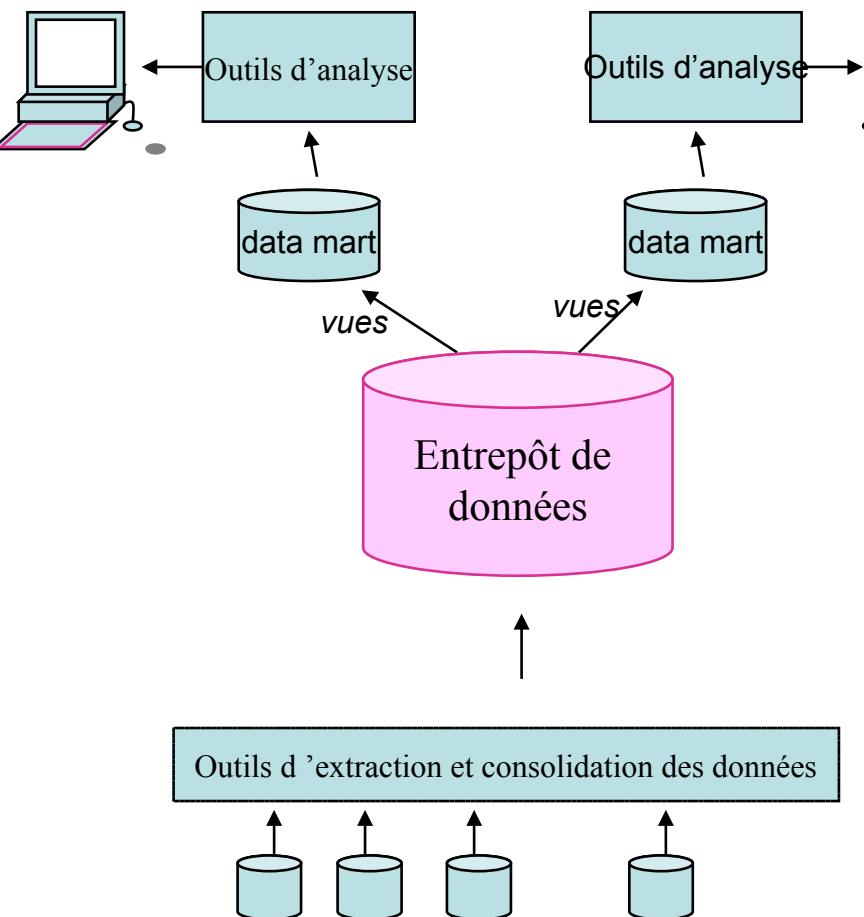
- plus d'interférences avec les BD opérationnelles,
- données consolidées et temporalisées

Inconvénients:

- pas de vue d'ensemble des données
- risque de divergences d'interprétations
- pertes dues aux recouvrements (importants) de données entre data marts.
- Incohérence potentielle des flux et des modèles
- Utilisation de technologies hétérogènes
- Investissement démultiplié
- Administration répartie

Architectures d'entrepôts de données

3. L'entrepôt centralisé



Avantages

- exploitation d'informations unifiée sur toute l'organisation
- magasins de données spécialisés pour des besoins particuliers:
 - données cohérentes
 - agrégations pré-calculées,
 - indexes ajustés,
 - SGBD's spécialisés avec options décisionnelles
- environnement dédié
 - indépendance par rapport aux données de gestion
- ...

Construction d'un ED: points critiques

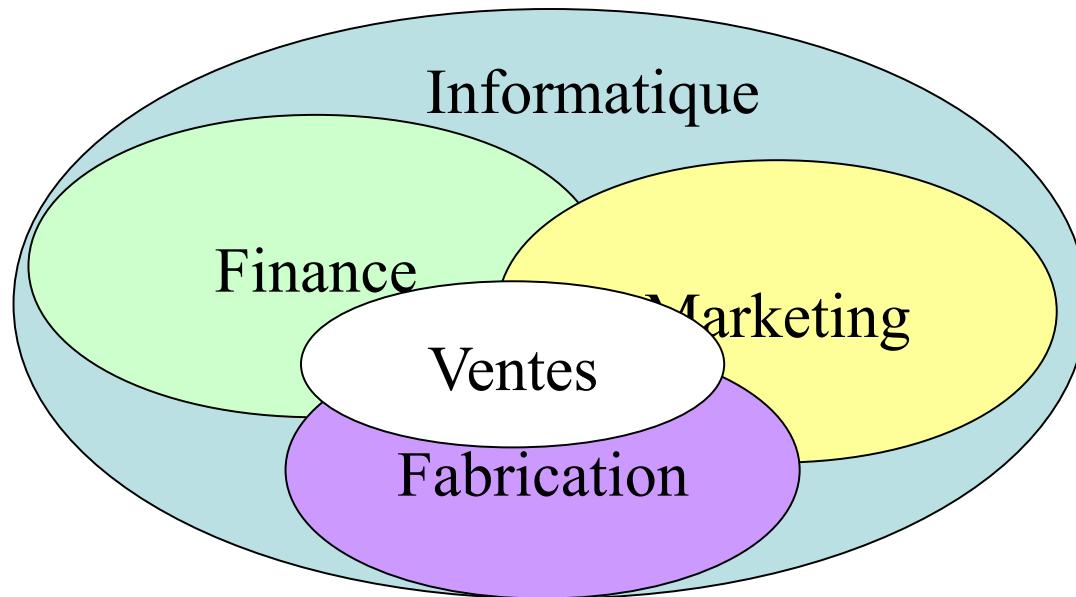
- Assurer une véritable conduite de projet
 - Identifier les besoins des utilisateurs
 - Fédérer l'ensemble des besoins
 - Evaluer les ressources, ...
- Bien connaître les métiers utilisateurs
 - Impliquer les utilisateurs :
 - Participer à la construction itérative du modèle d'information
 - Participer à la définition et l'évolution des métadonnées
 - Former à comprendre la logique de l'entrepôt
 - Chef de projet orienté utilisateur
- Procéder par étapes
 - Datamart
- Gérer l'évolution de l'entrepôt
 - Alimentation et administration

Modélisation Multidimensionnelle

- Modèle en étoile
- Modèle en flocon

Introduction : Modélisation

Objectif : trouver un langage commun



Introduction : Modélisation E/A

Avantages :

- Normalisation :
 - Eliminer les redondances
 - Préserver la cohérence des données
- Réduction d'espace de stockage
- Optimisation des transactions

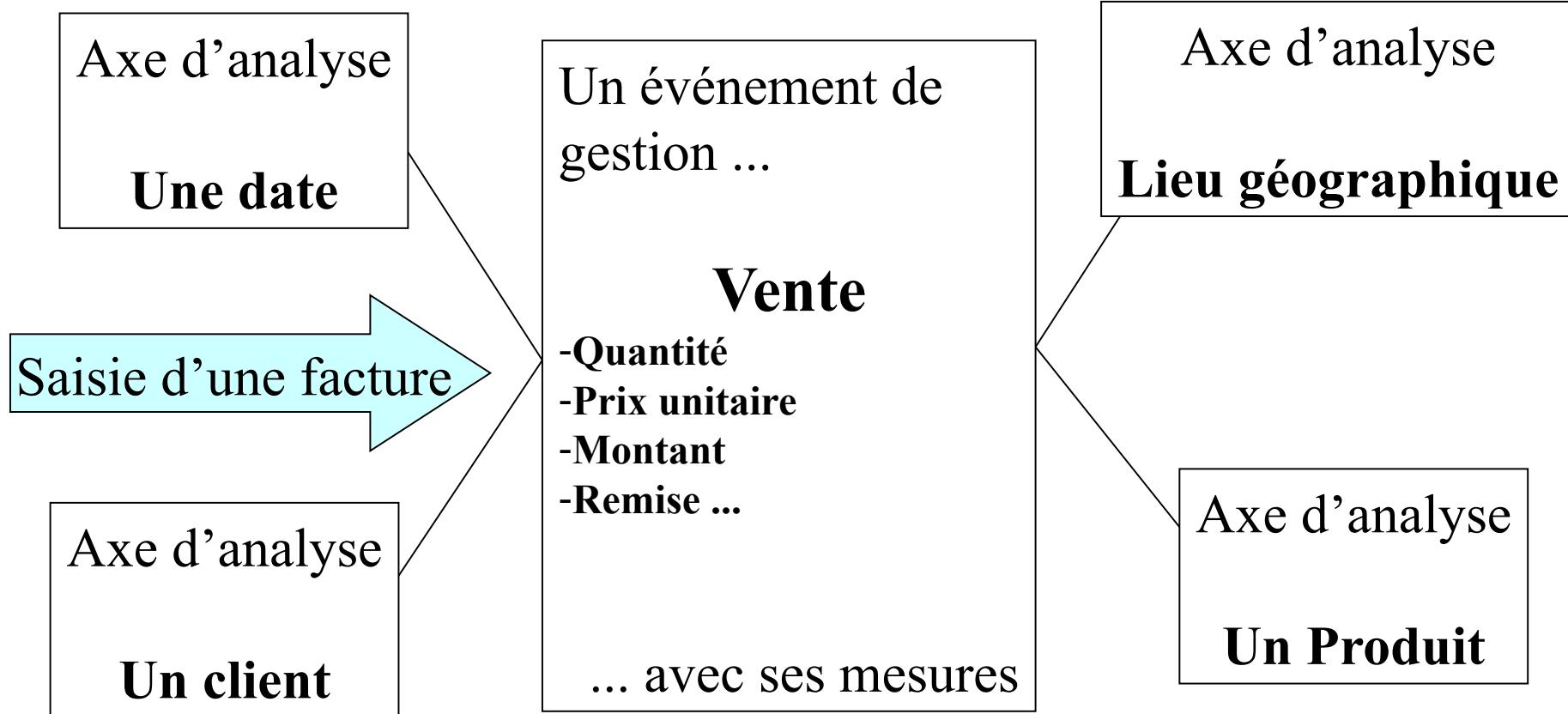
Limites pour les entrepôts de données :

- Inadapté pour l'analyse
- Schéma incompréhensible pour l'utilisateur final
- Conséquences désastreuses sur le performances

Introduction

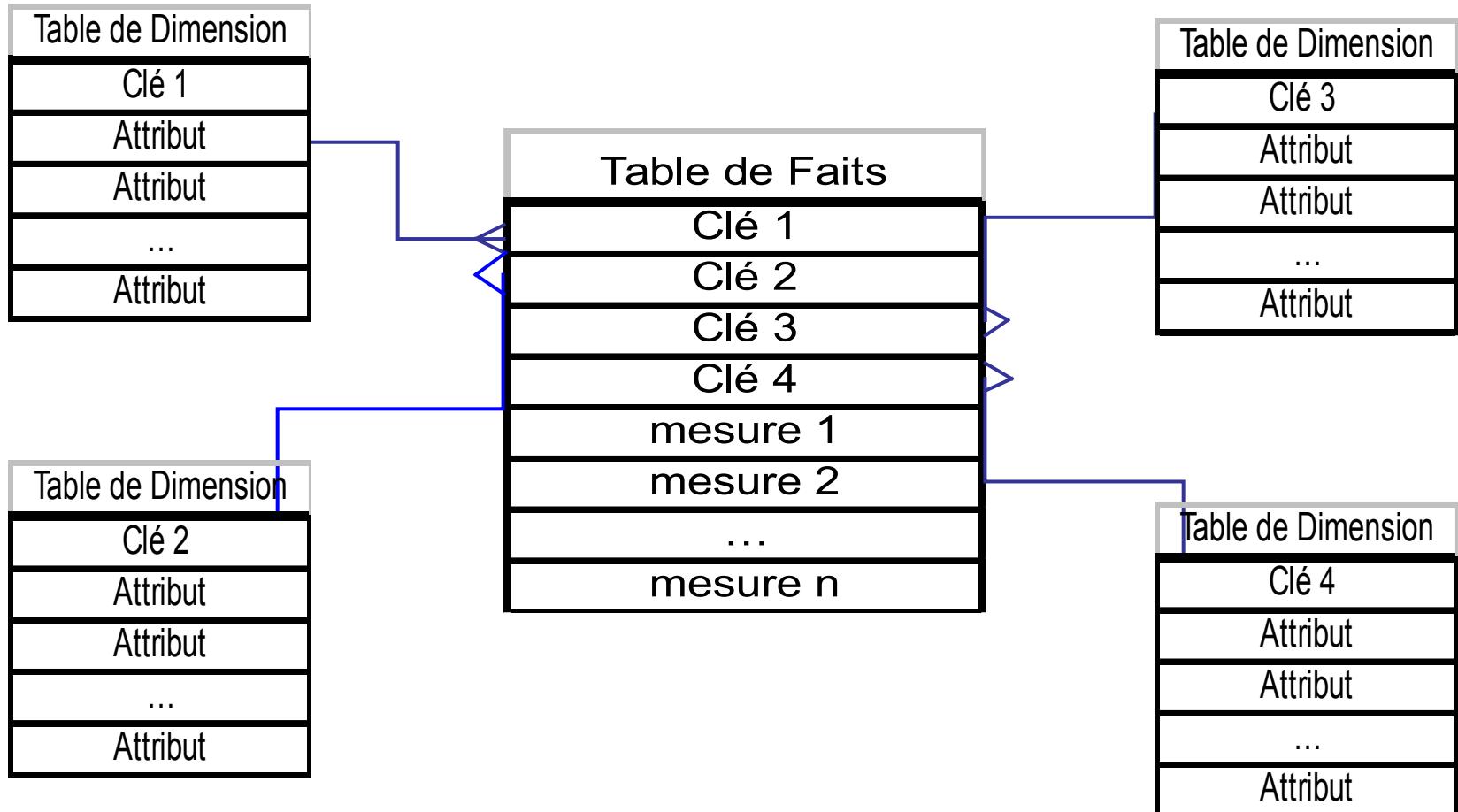
- But : Modèle dédié à l'analyse autour des concepts métiers
 - Lisibilité et simplicité d'usage
 - Performance des traitements
 - Capacité à évoluer
 - Transparence technique et fonctionnelle
- Solution : Modèle multidimensionnel / Modèle en étoile

La modélisation multidimensionnelle



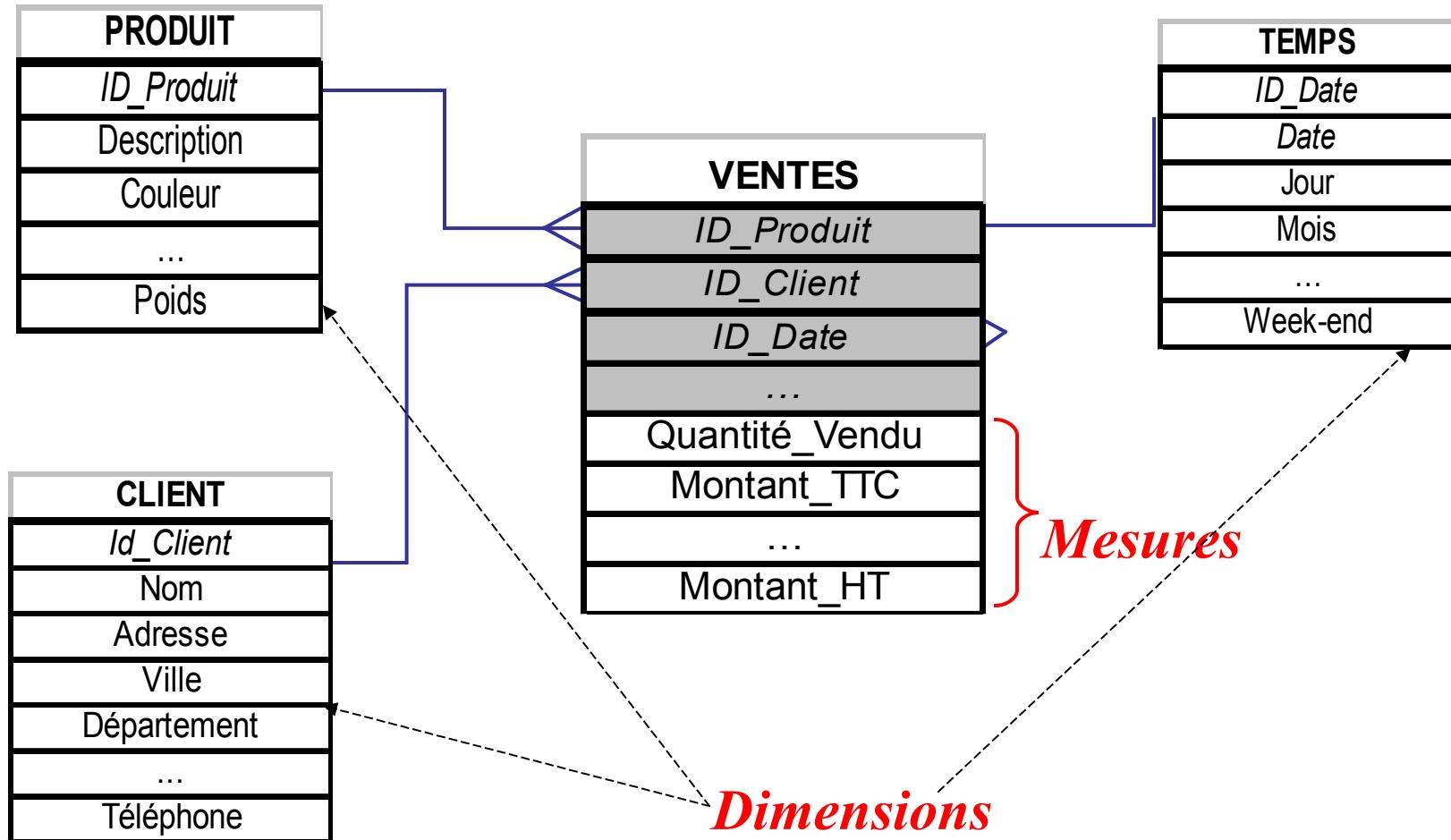
Modèle en étoile

- Une table centrale : la table de faits
- Des tables périphériques (dimensions) : les axes d'analyse



Modèle en étoile

Exemple : Analyse des ventes par client, par période, par produit



Caractéristiques des tables de dimension

- Un attribut pour clé primaire (générée)
- Autres attributs contiennent des données descriptives
 - Textuelles, non-mesurables
 - Ensemble limité des valeurs
 - Exemple : dimension produit, nb produits = 3000, dimensions temps sur 10 ans = 3650 lignes ...
 - Granularité
- Fortement dénormalisée
 - Exemple : Dimension Client
- Relativement statique
 - Exemple : Dimension Temps

CLIENT
<i>Id_Client</i>
Nom
Adresse
CodePostal
Ville
Département
...
Téléphone

Dimension Temps

- Commune à tout ED :
 - Analyse → observation des faits dans le temps
- Reliée à toutes les tables de faits

TEMPS	
Id_Date (CP)	
Date	
Jour_de_Semaine	
NumJour_Dans_Mois	
NumJour_Dans_Année	
NumSemaine_Année	
Mois	
Trimestre	
Année	
Jour_Férié	
Période_fiscale	
WeekEnd	
Saison	
...	

Caractéristiques des tables de faits

- Contient :
 - des faits / mesures
 - Grain de mesures de l'activité : valeur numérique
 - Exemple : chiffre d'affaire, montant de vente, ...
 - Zéro, un ou plusieurs faits
 - Trois types de faits : additif, semi-additif, non additif
 - un ensemble de clés étrangères des tables de dimension
 - Exemple : Id_période, Id_produit, Id_Magasin, ...
 - Une clé par dimension
- Une granularité
 - Détermine la taille de la table
- Mince et longue

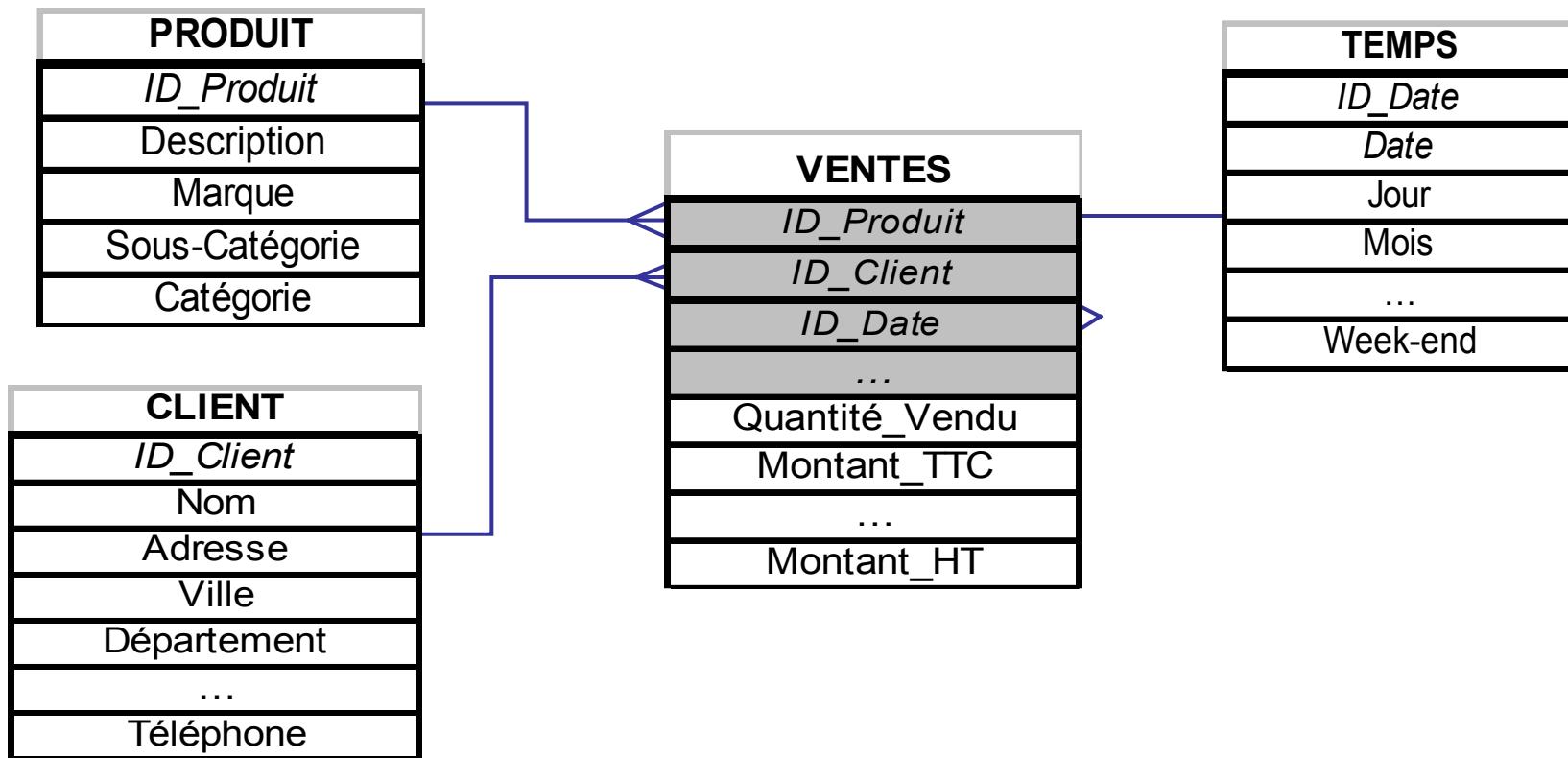
Exemple: 200 magasins, 500 produits par année pendant 10 ans
Nombre de lignes = $200 \times 500 \times 10 \times 365 = 365\,000\,000$

Les Faits

- Fait additif
 - Additionnable suivant toutes les dimensions
 - Exemples : quantité vendue, montant, chiffre d'affaire ...
- Fait semi-additif
 - Additionnable suivant certaines dimensions seulement
 - Exemples : solde de compte, niveau de stock, nombre de client
- Fait non-additif
 - Non additionnable quelque soit la dimension
 - Exemple (ratio) : marge brute = $1 - \text{Coût/CA}$

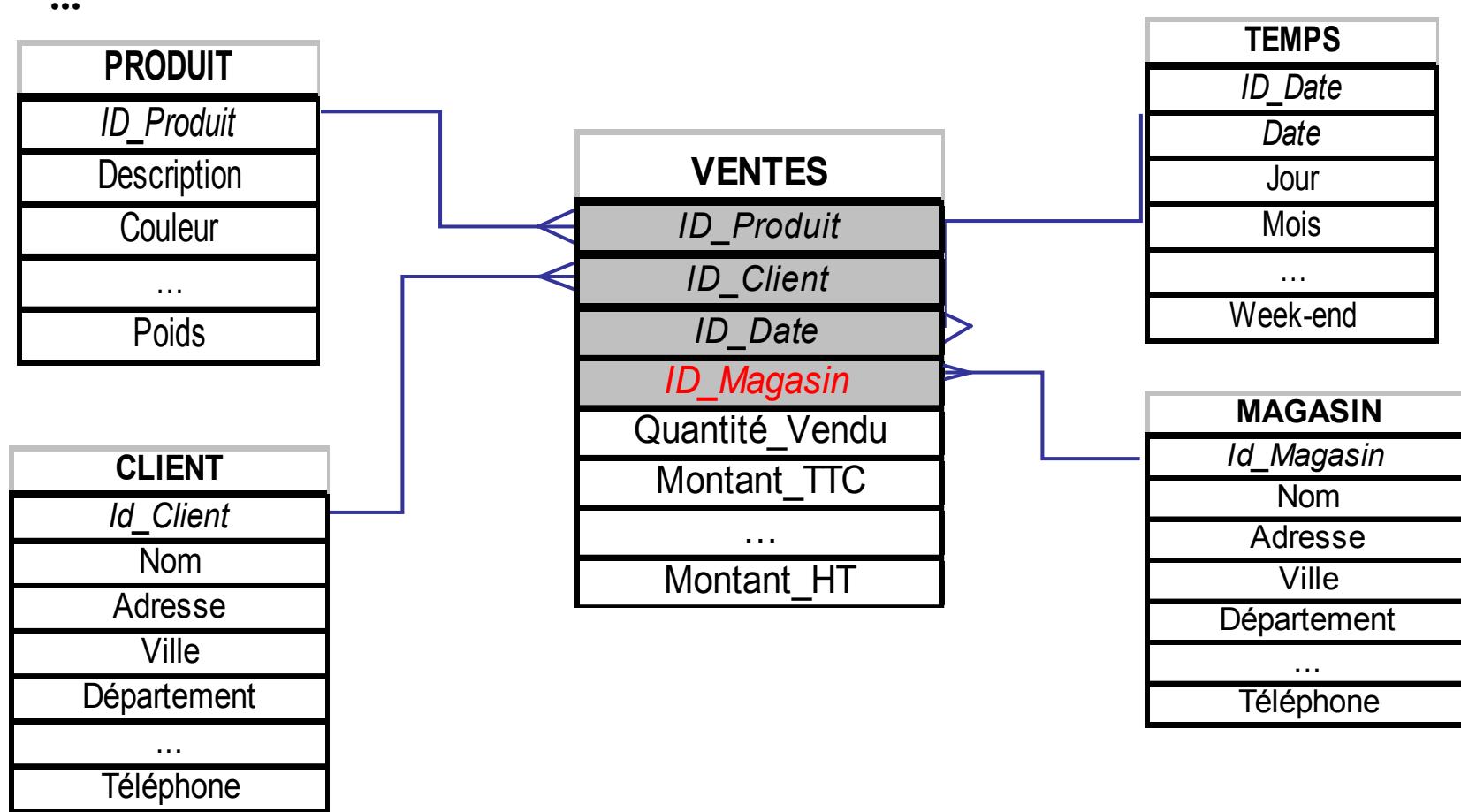
Exemple : Analyse des ventes

- Montant par semaine/mois/année
- Montant par marque/sous-catégorie/catégorie
- Montant par semaine/mois/année et par marque/sous-catégorie/catégorie
- ...



Exemple : Analyse des ventes

- Montant par semaine/mois/année
- Montant par marque/sous-catégorie/catégorie
- Montant par semaine/mois/année et par marque/sous-catégorie/catégorie
- **Montant par semaine/mois/année et par marque/sous-catégorie/catégorie/magasin**
- ...



Avantages du modèle en étoile

- Intuitive, facile à comprendre
- Extensible :
 - Faits nouveaux
 - Dimensions nouvelles
 - Attributs dimensionnels non-prévus
- Temps d'exécution de requêtes plus rapide
- ...

Hiérarchie de dimension

- Analyses ascendantes et descendantes plus facile
- Contenue dans la table de dimension
- Exemples : Produit, Temps, Lieu géographique

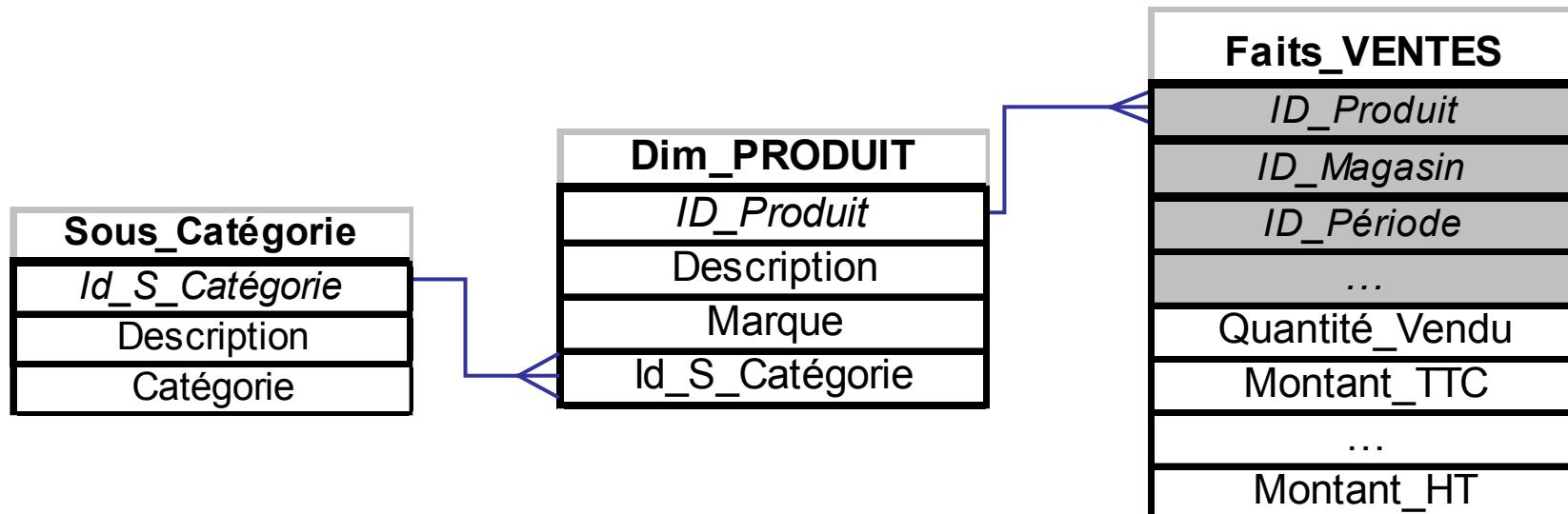
Dim_PRODUCT	
<i>ID_Produit</i>	
Description	
Marque	
Sous-Catégorie	
Catégorie	

Dim_TEMPS	
<i>ID_Date</i>	
Date	
Jour	
Semaine	
Mois	
Trimestre	
Semestre	
Année	
...	

Dim_LIEU GEO	
<i>ID_Lieu</i>	
Adresse	
CodePostal	
Ville	
Département	
Région	
Pays	

Modèle en flocon

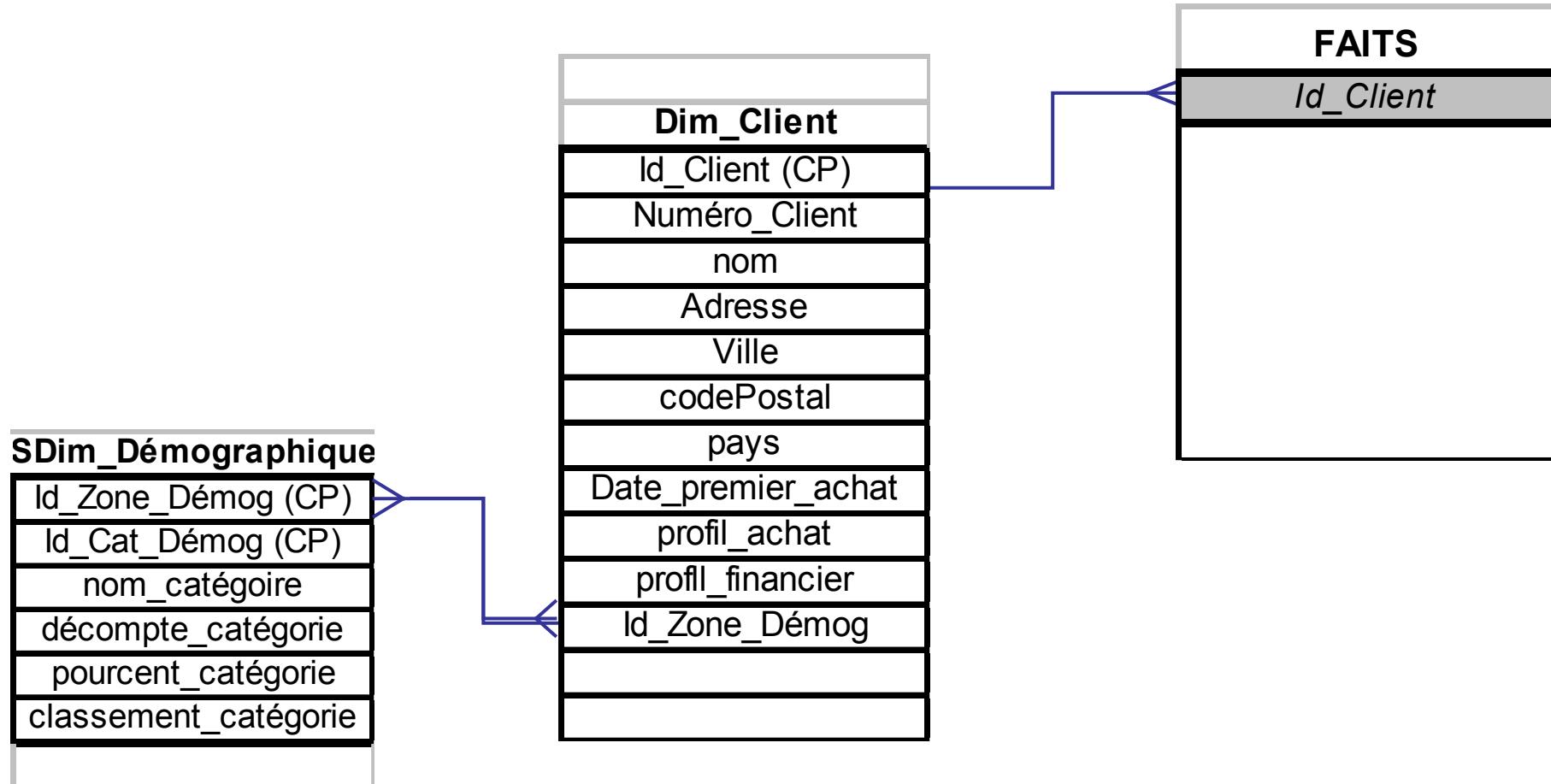
- Décomposition de tables dimensions par normalisation
 - Plusieurs niveaux de tables de dimension
 - Dimension de faible cardinalité



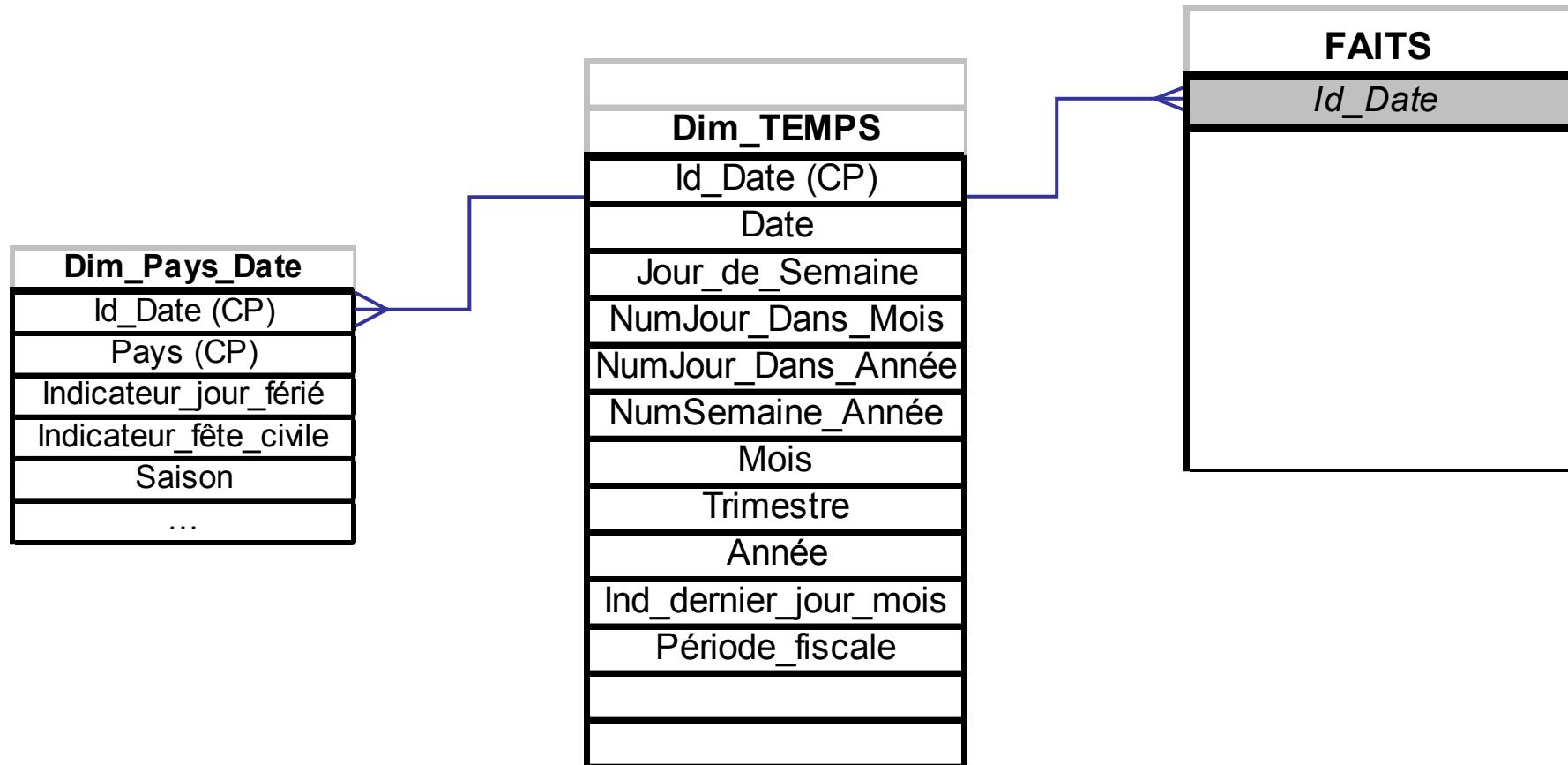
Modèle en flocon

- Avantages :
 - Gain d'espace
 - Exemple :
 - » 250 000 lignes, nom de catégorie codé de taille 15 octets ,
 - » Taille de clé = 2 octets
 - » Gain d'espace : $250\ 000 * 13\ o = 3,25\ Mo$
 - Insignifiant si taille de table de faits est grande
 - » Pour une table de 10 Go, gain = 0,03 %
 - Inconvénients :
 - Modèle plus complexe
 - Requêtes moins performantes

Modèle en flocon

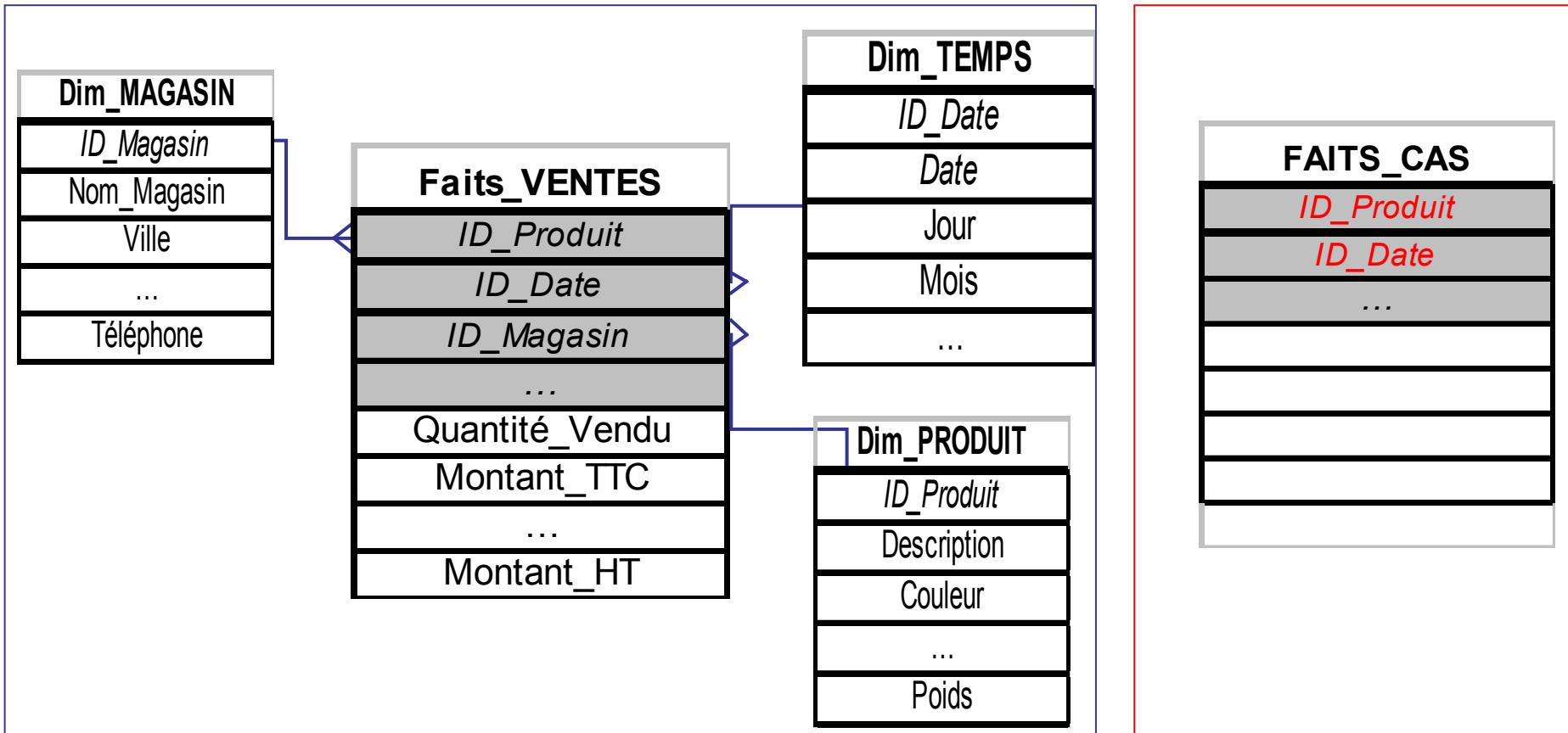


Modèle en flocon

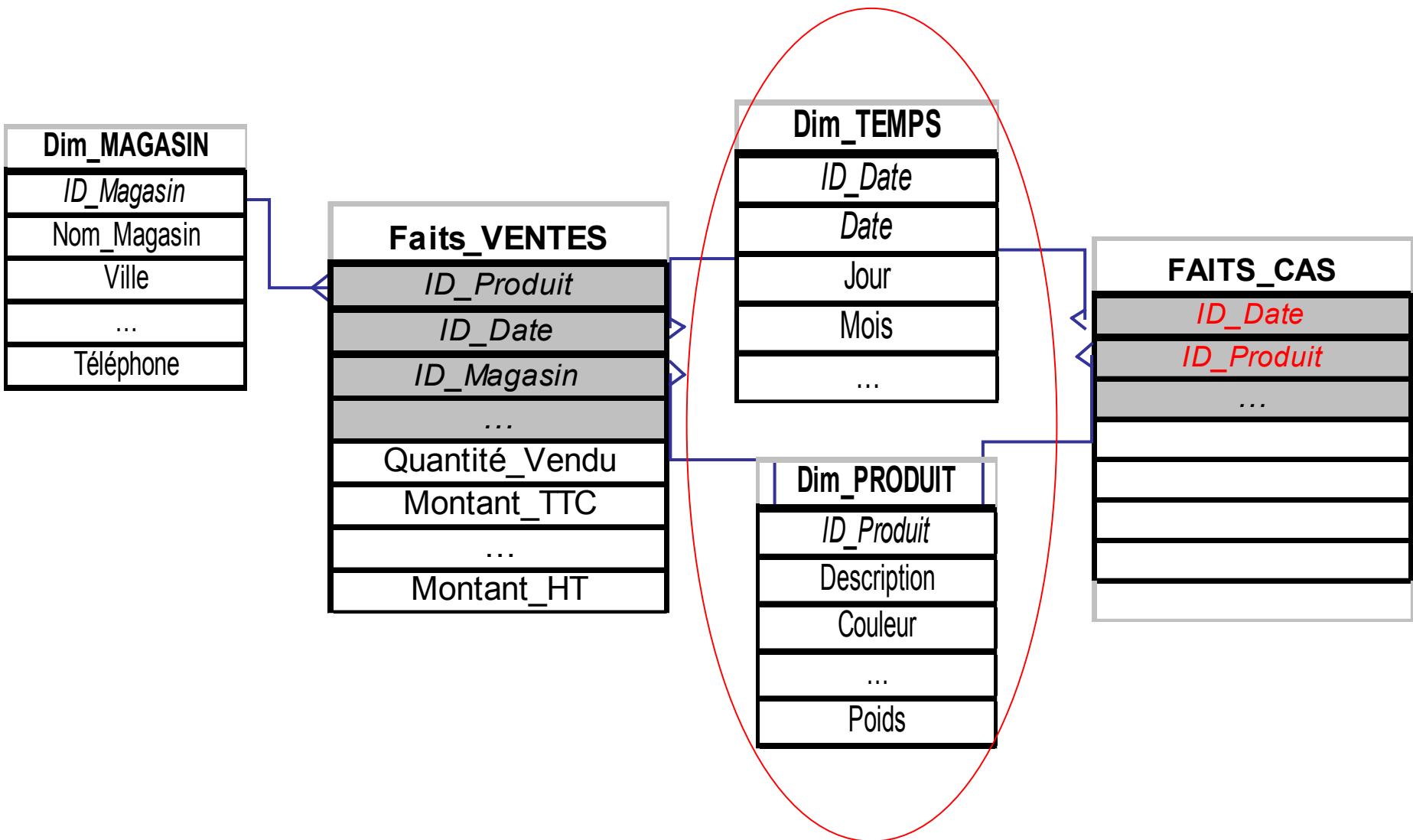


Les dimensions conformes

- Différents datamarts ayant les mêmes dimensions
 - Exemple : les dimensions produits, temps etc.



Les dimensions conformes



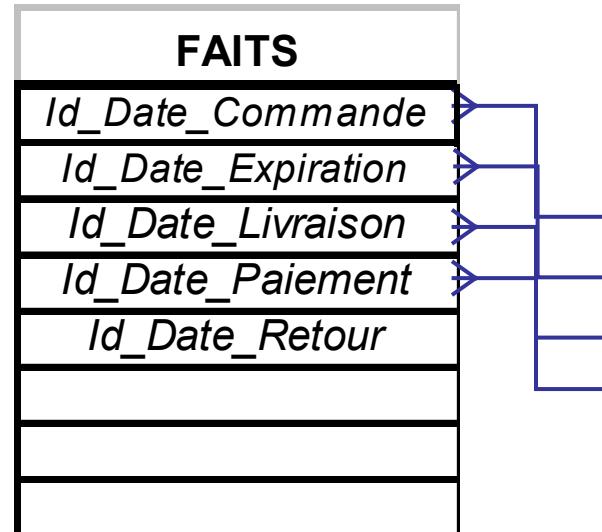
Les dimensions conformes

- Associée à différentes tables de faits
- A une signification commune
- Permet à l'entrepôt de fonctionner comme un ensemble intégré :
 - La cohérence des interfaces utilisateurs
 - La cohérence de l'interprétation des attributs

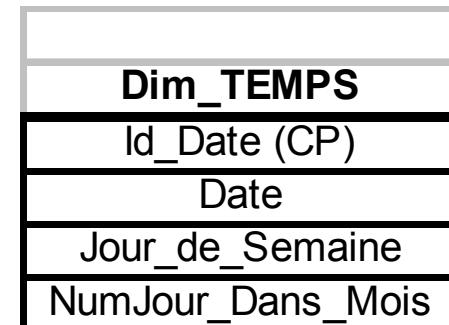
Dimension multi rôles

- Représenter une même dimension plusieurs fois dans la même table de faits

- Date de commande
- Date d'expiration
- Date de livraison
- Date de paiement
- Date de retour ...



- Utilisation des vues
 - Unicité des noms d'attributs



Dimension multi rôles

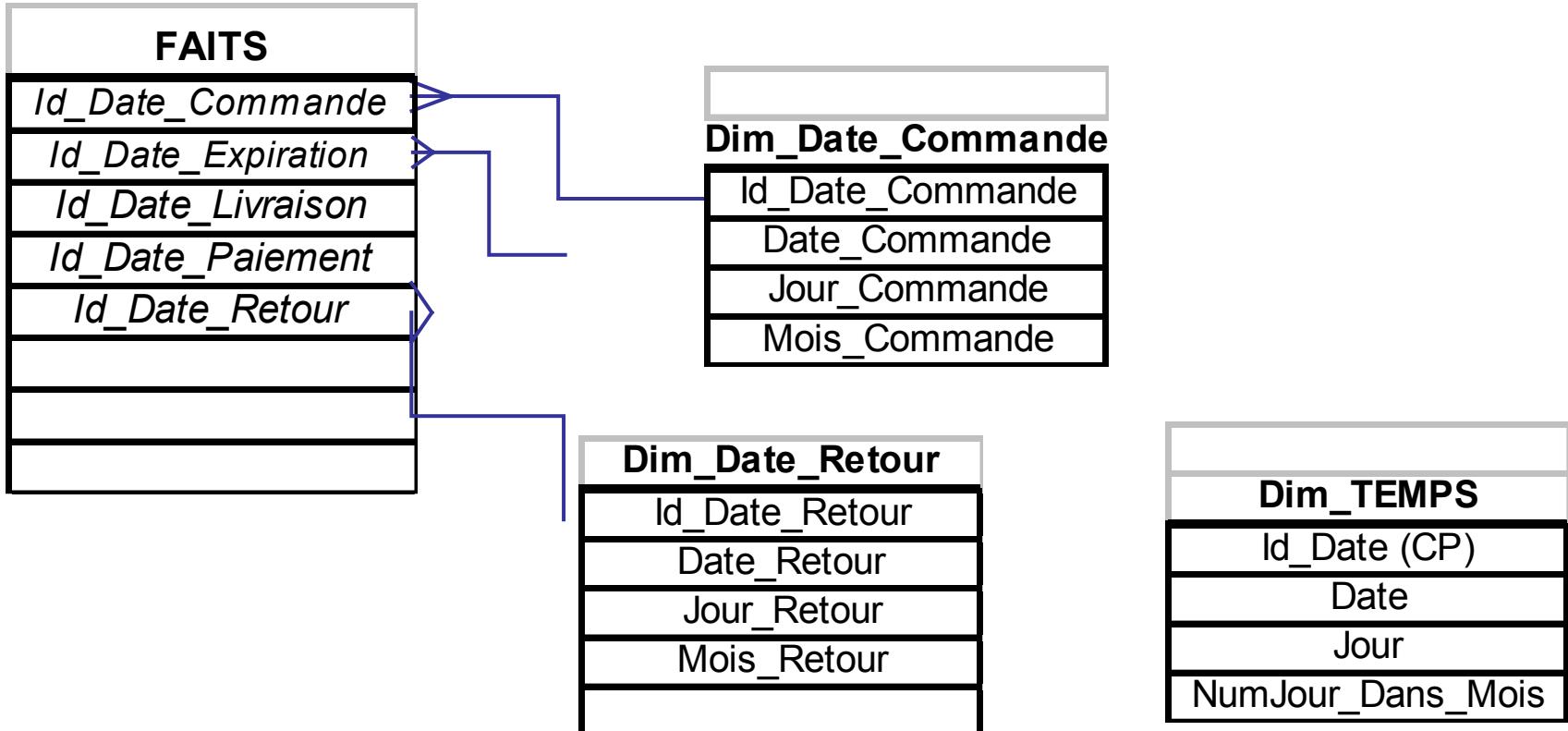
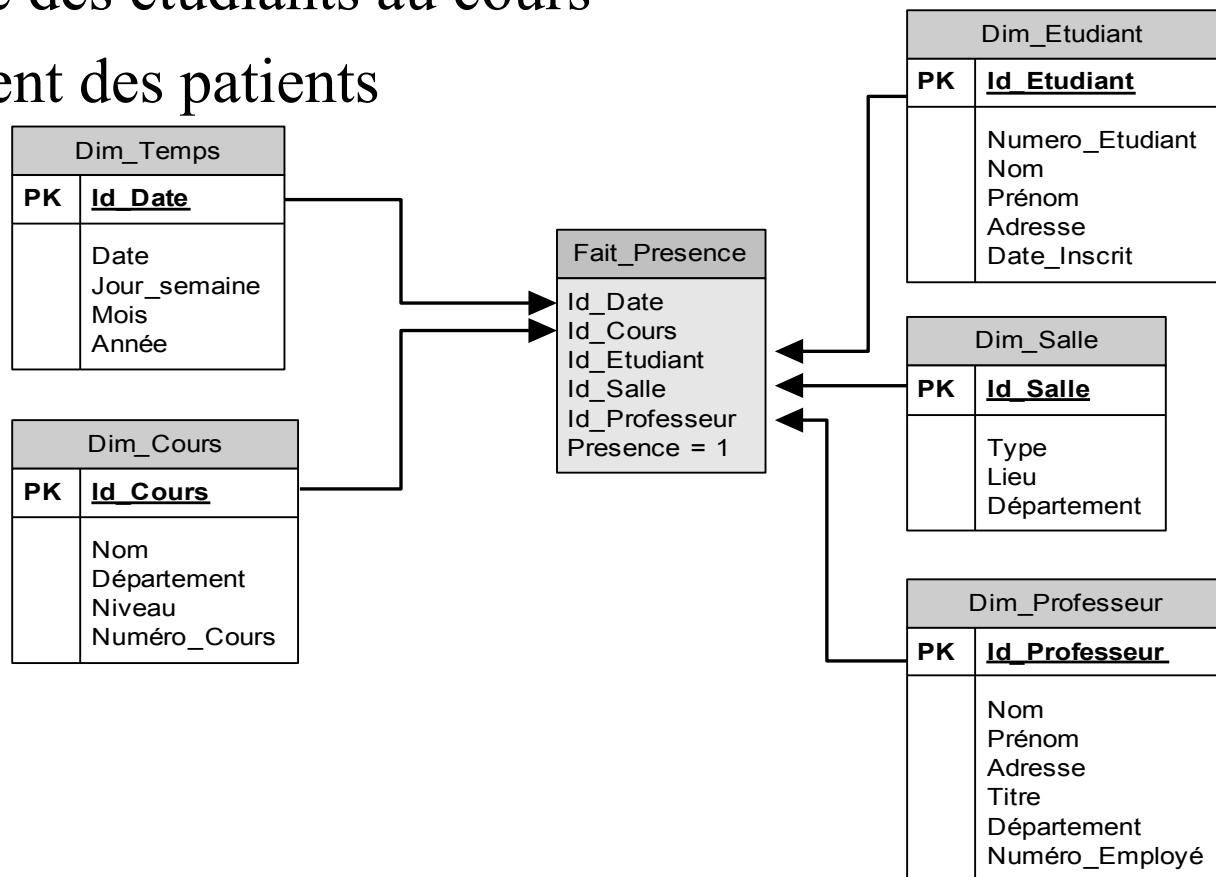


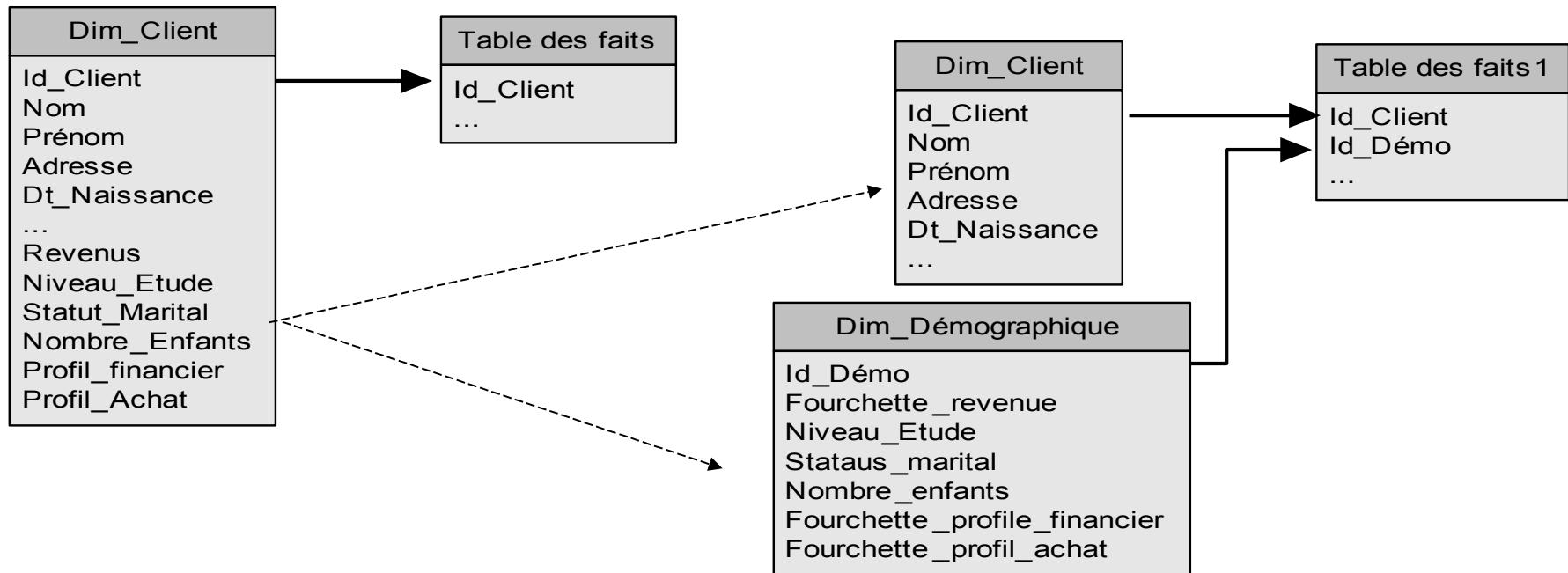
Table de faits sans faits

- Pas de faits
- Suivi des événements
 - Présence des étudiants au cours
 - Traitement des patients



Grande Dimension

- Évolution plus rapide de certains attributs
 - Clients pour les compagnies d'assurance



Méthodologie : 9 étapes de kimbball

- Choisir le sujet
- Choisir la granularité des faits
- Identifier et adapter les dimensions
- Choisir les faits
- Stocker les pré-calculs
- Établir les tables des dimensions
- Choisir la durée de la base
- Suivre les dimensions lentement évolutives
- Décider des requêtes prioritaires, des modes de requête