Entrepôt de données
(Datawarehouse)

M. BEN SALAH

Introduction

Concepts de datawarehouse et architecture

Modèle dimensionnel

Introduction

Concepts de datawarehouse et architecture

Modèle dimensionnel

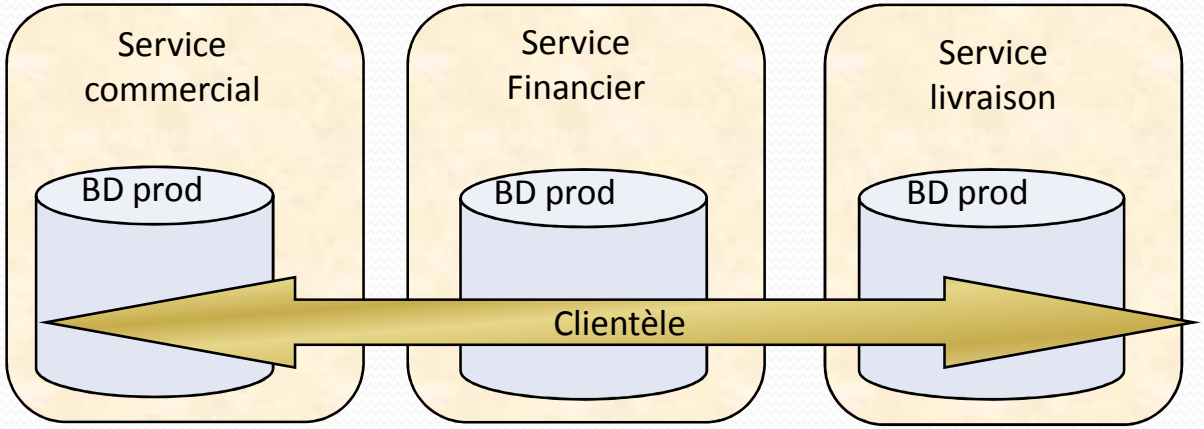
Acquisition des données et alimentation

OLAP

Introduction

- Besoin :
  - Les décideurs ont besoin d'informations pour pouvoir prendre des décisions stratégiques
  - Retrouver une information historique et transversale à l'entreprise

OLTP: On-Line Transactional Processing



- Analyser, décider

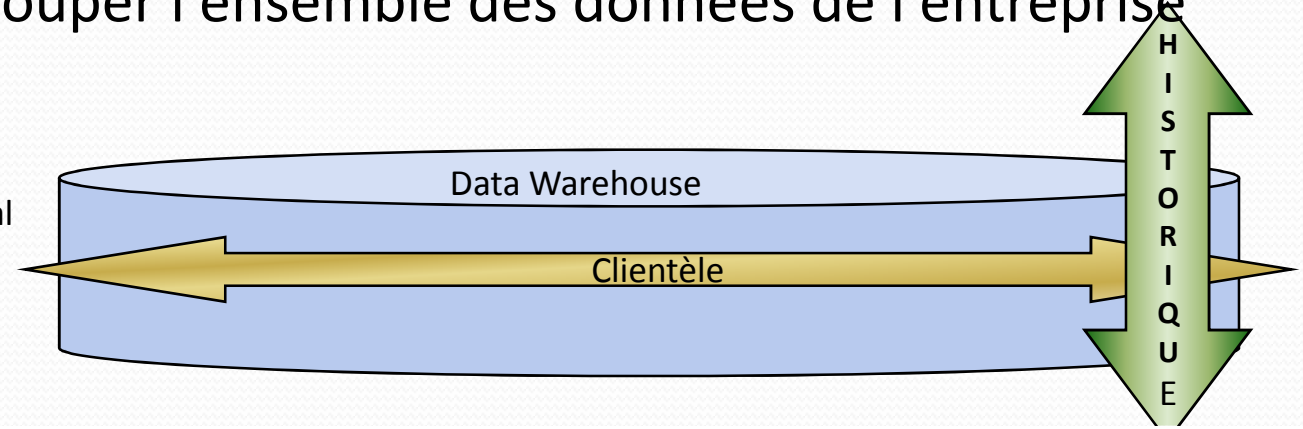
- Pour la décision : requêtes complexes

Introduction

Problématique	Conséquence
- Les informations s'obtiennent à partir des données de divers sources (BDs de productions, …) - les requête d'aide à la décision sont complexe et couteuses	- Hétérogénéité - On peut pas utiliser les BDs de production pour l'analyse

- Comment résoudre une partie de ces problèmes ?
  Fédérer/Regrouper l'ensemble des données de l'entreprise

OLAP: On-Line Analitical Processing



Introduction

- Comment prendre des décisions sur la base d'informations issues de systèmes hétérogènes n'aillant pas de moyens pour communiquer facilement entre eux ?

== > Le datawarehouse répond en partie à cette problématique. En effet, cette base de données regroupe l'ensemble des informations de l'entreprise de façon cohérente dans le but de faciliter l'analyse et la prise de décision.

Introduction

- la plus par du temps le système décisionnel se compose :

  - d'un entrepôt de données ou d'une structure permettant d'accéder à l'ensemble de l'information.

  - d'outils d'analyse qui assurent la présentation des documents à l'aide d'interfaces graphiques.

OLTP vs OLAP

- BD-OLTP : représentation de données sous forme plat
- ED : Permet et facilite la représentation de données sous forme multidimensionnelle : cube

Produit	Pays	Quantité
Stylo	France	2000
Stylo	Allemagne	3000
Stylo	Suisse	1000
Stylo	Maroc	2500
Crayon	France	10000
Crayon	Allemagne	12000
Crayon	Suisse	7000
Crayon	Maroc	10000
Gomme	France	25000
Gomme	Allemagne	25000
Gomme	Suisse	15000
Gomme	Maroc	15000

BDD

DW

	France	Allemagne	Suisse	Maroc	Total
Stylo	2000	3000	1000	2500	**8500**
Crayon	10000	12000	7000	10000	**39000**
Gomme	25000	25000	15000	15000	**80000**
Produit	**37000**	**40000**	**23000**	**27500**	**127500**

OLTP vs OLAP

OLTP	OLAP
Orienté transaction	Orienté analyse
Orienté application	Orienté sujet
Données courantes	Données historisées
Données détaillées	Données agrégées
Données dynamiques	Données statiques
Utilisateurs nombreux, administrateurs/opérationnels	Peu de nombre d'utilisateurs dirigeant
Requête simple (insertion, update, …) Temps d'exécution: court	Requêtes complexe Temps d'exécution: long

Introduction

Concepts de DW et architecture

Modèle dimentionnel

Qui ce que un entrepôt de données?

- L'entrepôt de données est l'élément central de l'informatique décisionnelle. C'est un moyen pour modéliser et stocker l'information pour des fins d'analyse
- C'est une base de données conçus pour l'aide à la prise de décision
- ED offre une vision transversale des données de l'entreprise : intégration des données de différentes BDs et d'autres sources
- Optimisé pour répondre à des questions complexes pour décideurs et analystes.
- Les données sont organisées pour avoir un accès rapide et sous forme synthétique à l'information stratégique dont on a besoin pour la prise de décision

Qui ce que un entrepôt de données?

- ED est une base architecturée pour des requêtes et des analyse plutôt que pour le traitement transactionnel des données.

- Mode de travail : OLAP (On-Line Analytical Processing).

Le but d'un entrepôt de données est de supporter le traitement analytique en-ligne (OLAP).

Qui ce que un entrepôt de données?

- Définition : Le Datawarehouse (DW) ou entrepôt de données (ED) est une collection de données orientée sujet (thématique), intégrées, non volatiles et historisées (donc datées), organisées pour des analyses et la prise de décision.

Bill Inmon.

.

- Caractéristiques :
  - orientation sujets («métiers»)
  - données intégrées
  - données non volatiles
  - données datées

Data Warehouse

Caractéristiques

- **Orientation sujet :**

- Regroupe les informations par thèmes ou par déférents métiers ou sujets (vente, …) == > Structuration autour des principaux sujets de l'entreprise

== > disposer de l'ensemble des informations utiles sur un sujet souvent transversal aux structures fonctionnelles et organisationnelles de l'entreprise

== > Les données collectées (extraite de divers sources) doivent être orientées sujet et donc triées par thème : on n'intègre pas toutes les données de l'entreprise, mais agrège celles qui concerne un sujet précis, qui est son objet d'analyse.

Caractéristiques

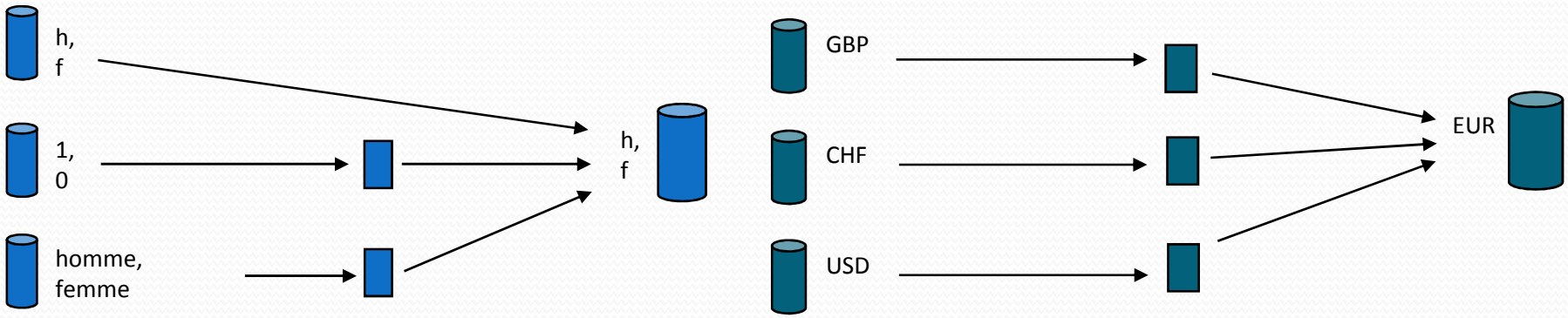
- **intégrées :**

Normalisation des données, définition d'un référentiel unique :

- Remise en forme avant l'intégration dans le datawarehouse.

- Uniformisation et unification pour rendre les données cohérentes

Puisque les données sont extraite de différentes sources elles sont souvent incohérentes (arborent des formats différents). Les sources de données intégrées doivent subir une mise en cohérence pour présenter une vue unifiée des données aux utilisateurs == > les données doivent êtres nettoyer au préalable, mises en forme, normalisées et unifiées afin d'voir une description et un codage unique

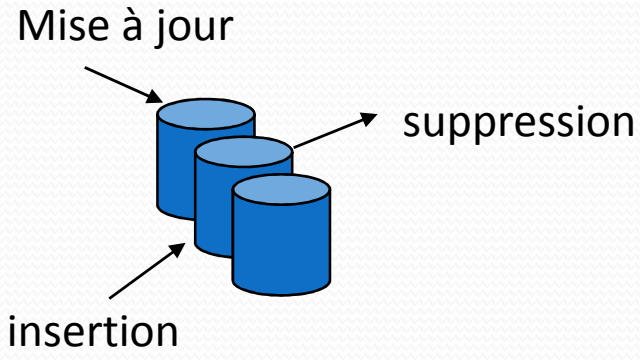


Caractéristiques

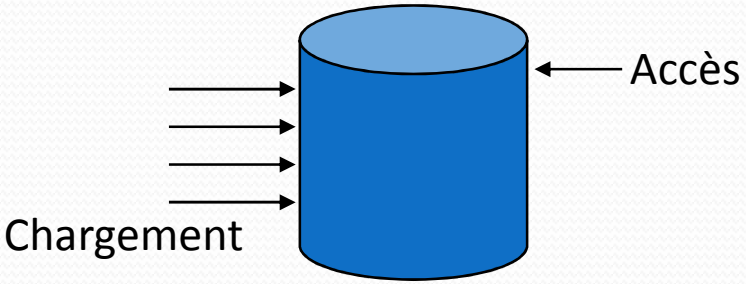
- **Données non-volatilité :**
  - Pas de changement au fil du temps (lecture seul)
  - Traçabilité des informations
  - Copie des données de production, uniquement des inserts

Dans un système de production; la donnée est mise à jour à chaque nouvelle transaction. Dans un Data Warehouse, la donnée ne doit jamais être supprimée ou modifiée. Seule les actions d'ajout et de lecture de donnée de ED doivent être autorisées.

= > conserver la traçabilité des informations et des décisions prises



Base de production



Entrepôts de données

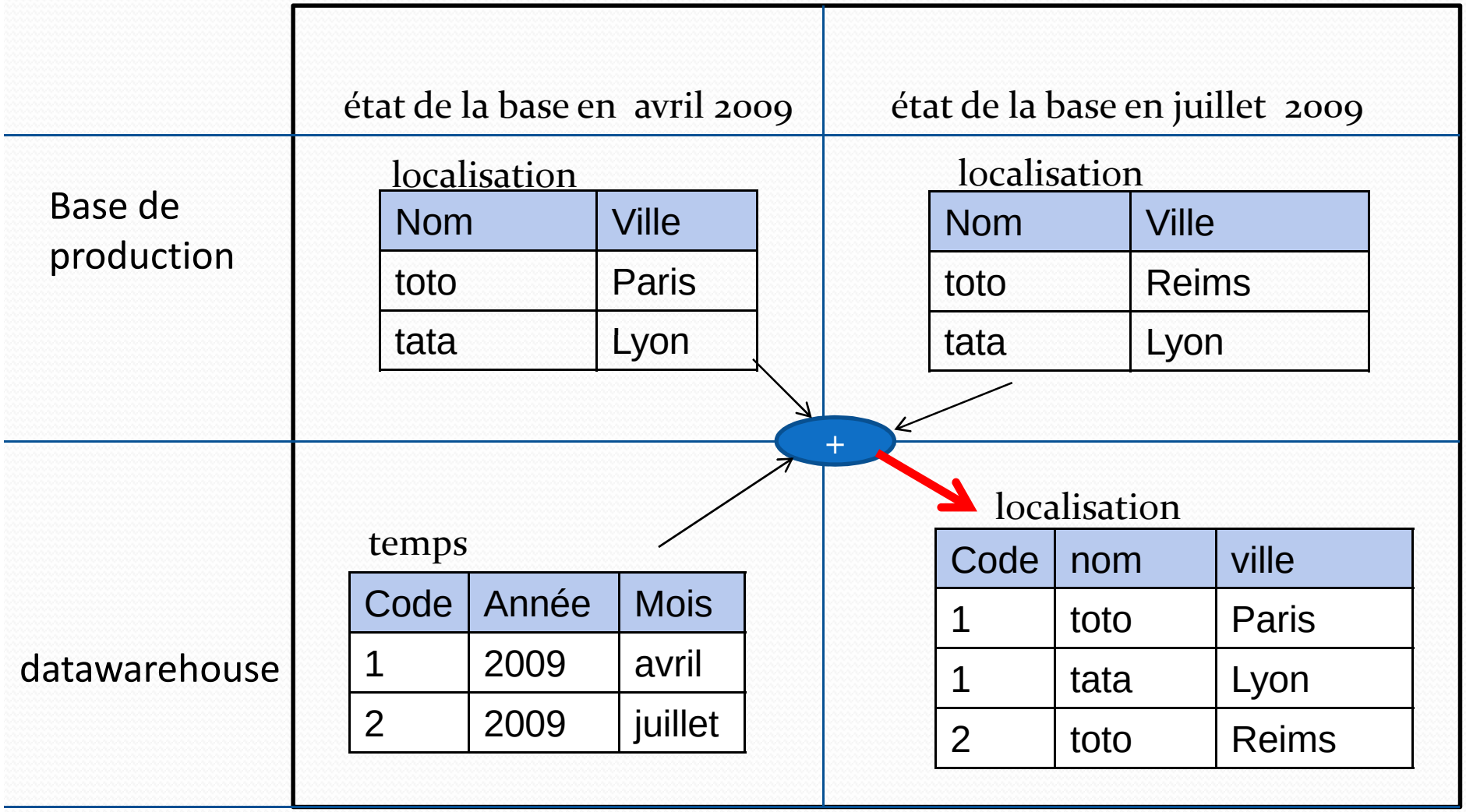
Caractéristiques

- **Données historisées :** Référentiel de temps associé aux données.

Besoin de suivre l'évolution, dans le temps, des différent indicateurs à analyser. Les données de ED doivent être historisées, donc datées afin d'être capable d'identifier une valeur particulière dans le temps.

Caractéristiques

- **Données historisées**



Objectifs de Datawarehouse

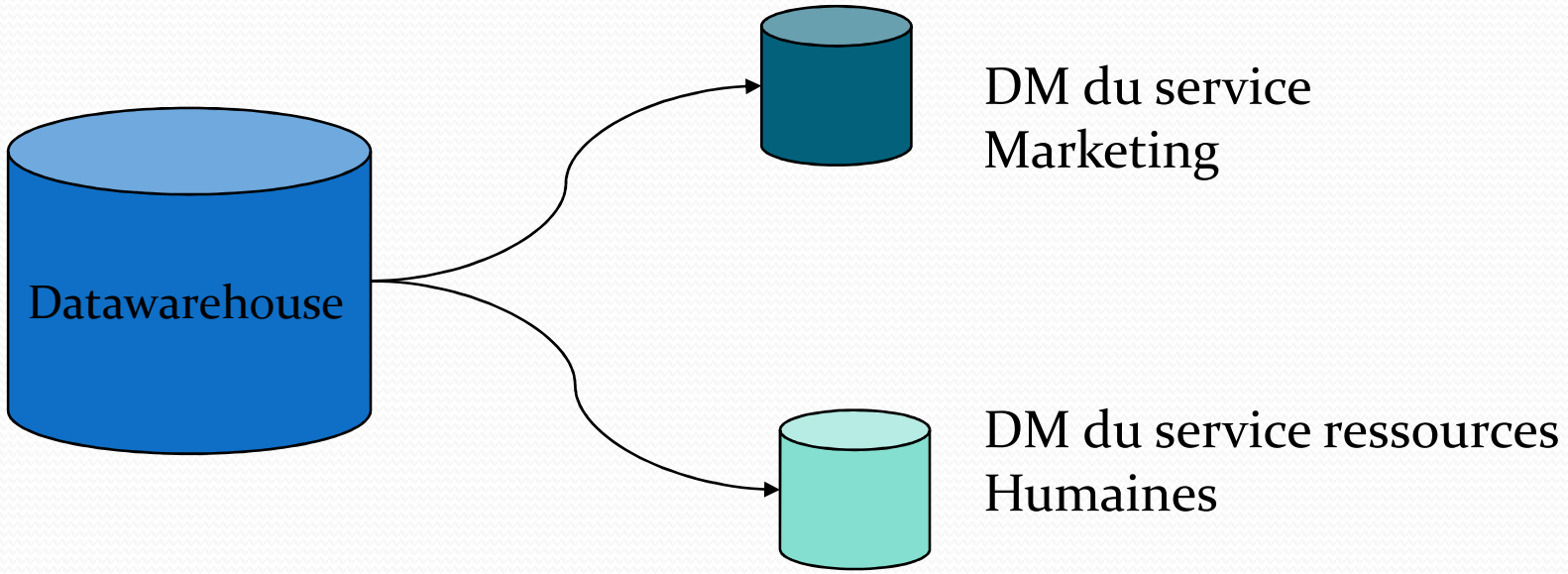
- Les objectifs principaux de Datawarehouse sont :
  - Récupérer les données existants dans divers sources
  - les intégrer, les grouper (agrégés) afin de fournir une vue orientée métier
  - les stocker et historisées
  - permettre l'**interrogation et** l'**analyse** les données facilement et rapidement.

Magasin de données (Datamart)

- datamart (DM) : magasin de données ciblé sur un sujet précis

  == > un sous-ensemble du datawarehouse contenant les données du datawarehouse pour un secteur particulier de l'entreprise.

Exemple : *DM Marketing*, *DM Commercial*, ...



Intérêt des datamart

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

bases de données et entrepôt de données (ED)

- SGBD et ED :
  - ont des objectifs différents et font des traitements différents
  - stockent des données différentes
  - font l'objet de requêtes différentes

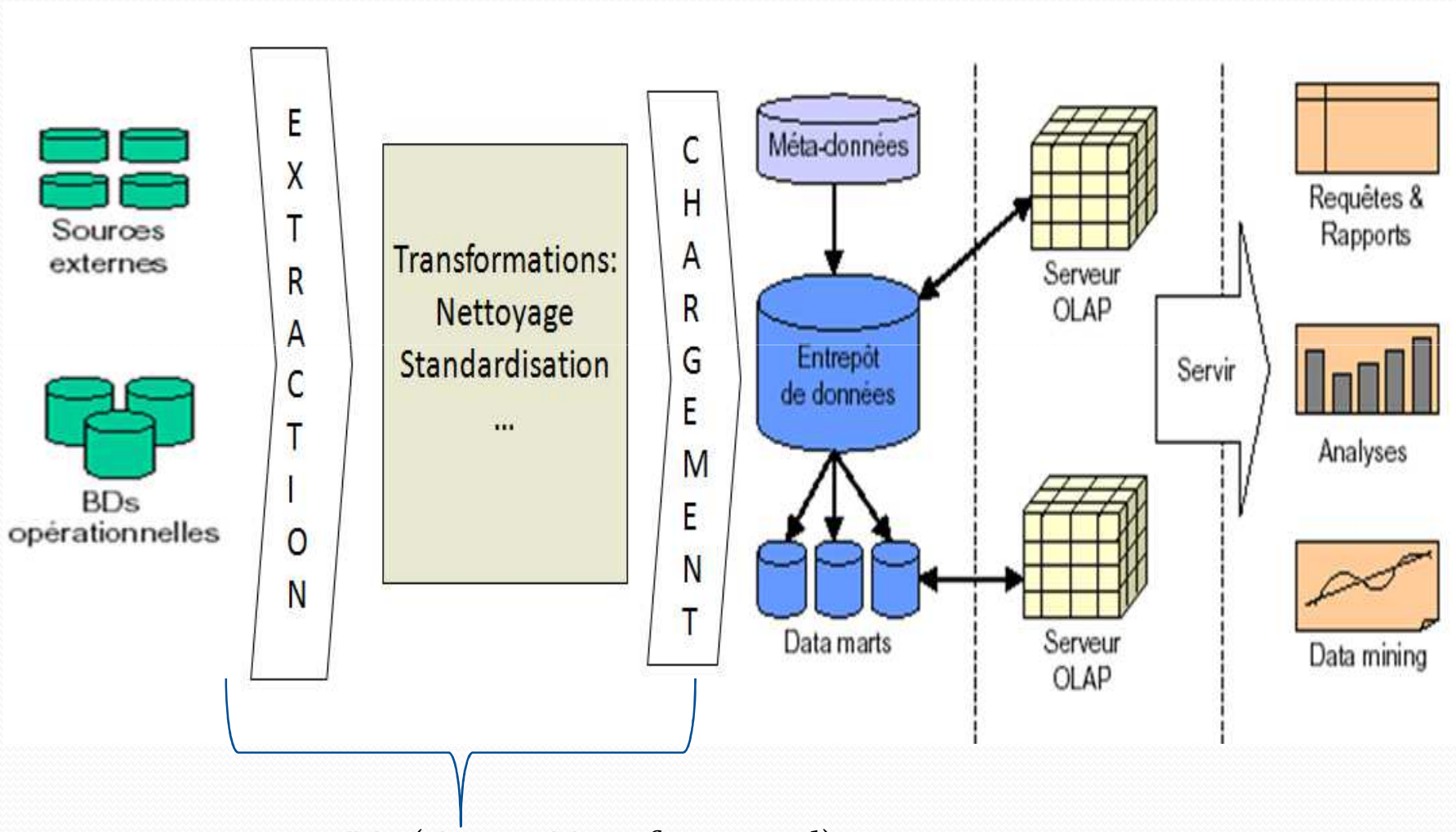
-> SGBD et ED ont besoin des organisation différente des données

-> pour raison de performances les SGBD et DW sont physiquement séparés.

Architecture des entrepôts de données

- Une architecture d'entrepôt de données possède les caractéristiques suivantes :
  - Extraction (acquisition) des données : les données sources sont extraites des bases de données, des fichiers …
  - Préparation des données et alimentation de l'entrepôt : les données sources sont nettoyées, transformées et intégrées avant d'être stockées dans l'entrepôt
  - Restitution : Interfaces et d'applications (clients) pour accès aux données de l'entrepôt

Architecture des entrepôts de données



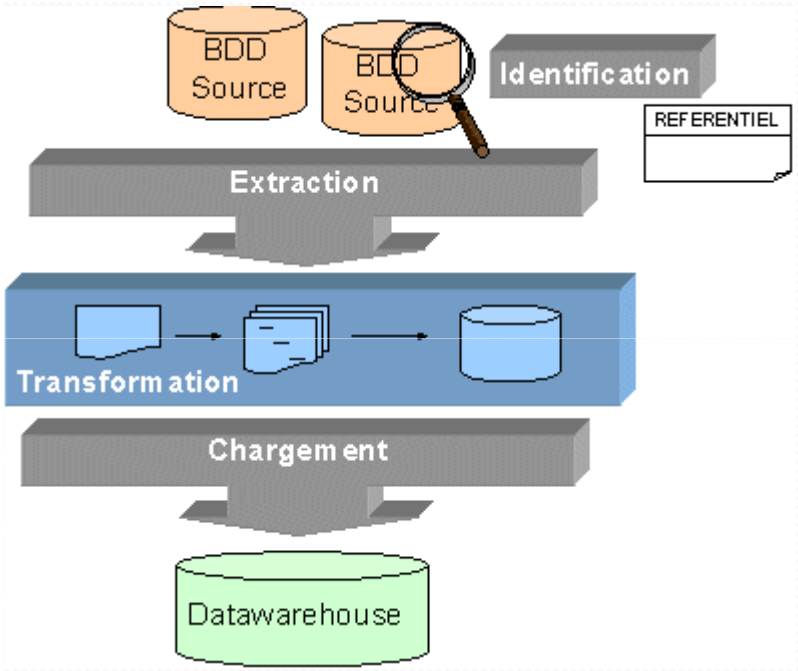
Architecture des entrepôts de données

- ETL : permet l'identification et Extraction des données les transformer puis les charger (stockées) dans l'ED.

  Un environnement BI inclut presque toujours une solution ETL (Extract Transform Load) pour extraire transformer et charger les données sources dans un ED.

  == > ETL = outils pour alimenter un ED

Architecture des entrepôts de données



- Flux de données
  - Flux entrant
    - Extraction: divers sources, hétérogènes
    - Transformation : filtrer, homogénéiser, nettoyer
    - Chargement : insertion des données dans l'entrepôt
  - Flux sortant :
    - Mise à disposition des données pour les utilisateurs finaux

Architecture des entrepôts de données

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

Architecture des entrepôts de données

- Métadonnées : Ensemble d'informations nécessaires à la construction, à la compréhension et à l'exploitation des données du datawarehouse.

- Le référentiel de l'entrepôt de données = métadonnées + outils d'administration

Introduction

Concepts de DW et architecture

Modèle dimentionnel

Mise en place d'un datawarehouse

Différente phases :

1- Conception

définir la finalité de l'entreprise : quelle activité analysée ?

Définition du modèle de données (modèle en étoile/flocon)

2- Acquisition des données et alimentation

Extraction, transformation et rechargement des données

3- Définir les aspects techniques de la réalisation

4- restitution des données :

requêteurs SQL, analyse multidimensionnelle …

Conception

- La conception est extrêmement complexe
- le but finale : la création d'un entrepôt de données qui comble les exigences de l'entreprise
- En générale les concepteur évitent de se livrer à une conception à l'échelle de l'entreprise. On procéder par étape dont l'objectif de mettre en place une solution simple et plus réaliste sujet par sujet.

La description du composant de base de données d'un entrepôt de données s'effectue à l'aide d'une technique, dénommée **modélisation dimensionnelle**

Recueil des besoins

- Identifier l'objectif principal:
  Quelles sont les attentes de l'ED ?

- Identifier les décisions:
  Quelles sont les décisions à prendre? (Quoi?)
  Quels sont les critères qui influencent la prise de décision ? (Comment?)
  Dans quel(s) but(s) les décisions sont elles prises? (Pourquoi?)

- Identifier les difficultés actuelles :
  Quelles sont les difficultés actuellement rencontrées dans la prise de décision ?
  - précision des données(détails), actualisation, vérification
  - synthèse des données (regroupements)
  - évolution (temps)
  - autres…

Conception

- Travail du concepteur:
  - Identifier le besoin
  - connaitre les exigences (avec les priorités) des utilisateurs
  - cibler les données pertinentes : identifier les données à prendre en compte avec leurs sources
  - Modéliser l'entrepôt
  - Concevoir les outils de visualisation
  - Contrôler la validité des données présentées.
  - Enrichir constamment l'ntrepôt.

La description du composant de base de données d'un entrepôt de données s'effectue à l'aide d'une technique, dénommée **modélisation dimensionnelle**

Recueil des besoins

- Avoir une idée sur la fréquence d'actualisation des informations
  Quels sont les besoins concernant la fréquence de mise à jour des informations proposées par le DataWarehouse?

- Comment présenter les informations ?
  - Quelles sont les préférences dans la présentation des informations
  - tableaux, graphiques ?
  - Type de graphiques (barres, nuages de points, camemberts, ...)?
  - Existe-t-il une présentation actuelle ou habituelle à conserver?

Modélisation

- Pour la modélisation conceptuelle de BD pour un systèmes de traitement de transactions en ligne (OLTP) on utilise le modèle entité-association (E-A) basé sur la normalisation

1 - La normalisation 3FN entraine la création de nombreuse tables.

== > Une requête décisionnelle requiert de nombreuses jointures et d'agrégats.

2- La normalisation en 3FN se justifie par la difficulté de gérer la redondance dans un système OLTP lors des mises à jours des données. Cette justification disparaît dans un système décisionnel.

Modélisation

- Pour la conception des EDs on utilise un nouveau modèle :

  la modélisation dimensionnelle par fois appelée modélisation OLAP (Codd1993)

  - Ce modèle répond mieux aux besoin décisionnel
  - Méthodologie de conception logique et de structuration données dédiée au reporting et à l'analyse. Elle a pour but de présenter les données dans une forme standard, intuitive, permettant des accès à grandes performances.
  - Méthodologie de conception autour des concepts métier

    => ne pas normaliser au maximum

Concepts du modèle dimensionnel

- Caractéristiques :
  - Appel au concepts de modélisation E-A avec des restriction importantes : on dé-normalise
  - Dans ce modèle, les activités réalisées par un processus métier sont décrites en terme de faits, mesurés par des indicateurs, et en terme de dimensions.
  - Un fait représente un événement (la vente d'un produit dans une région à une certaine date).
  - mesure : critère d'évaluation du processus décisionnel (ex : chiffre d'affaires (CA), quantité en stock)
  - dimensions : Axe d'analyse associé à un indicateur, représentant un sujet d'intérêt (ex : temps, produit, localisation)

Concepts du modèle dimensionnel

- Caractéristique :
  - Introduction de nouveaux types de tables
    - table de dimensions :

      Axes d'analyse

      ex. : dimension produit, région …

    - table des faits :

      Sur quoi va porter l'analyse

      ex. : fait vente

      - Un fait contient des mesures (ex. CA)
  - Introduction de nouveaux modèles :

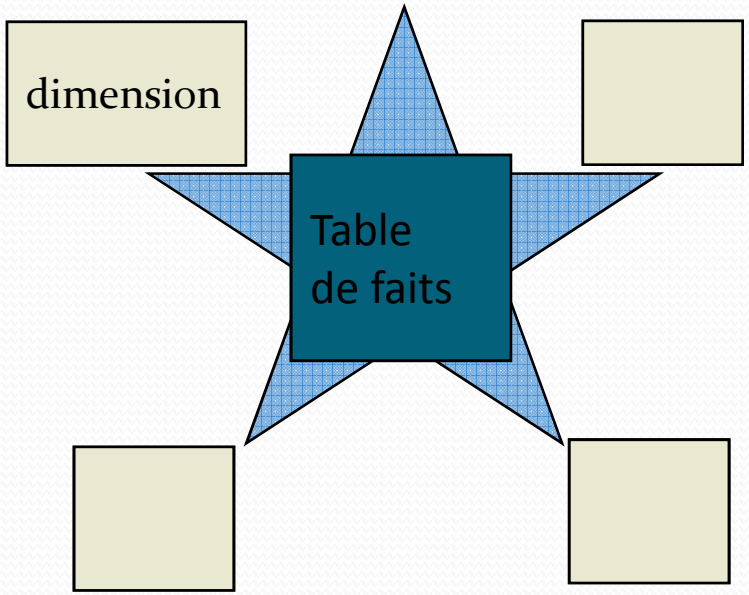
    - modèle en étoile
    - modèle en flocon

Concepts du modèle dimensionnel

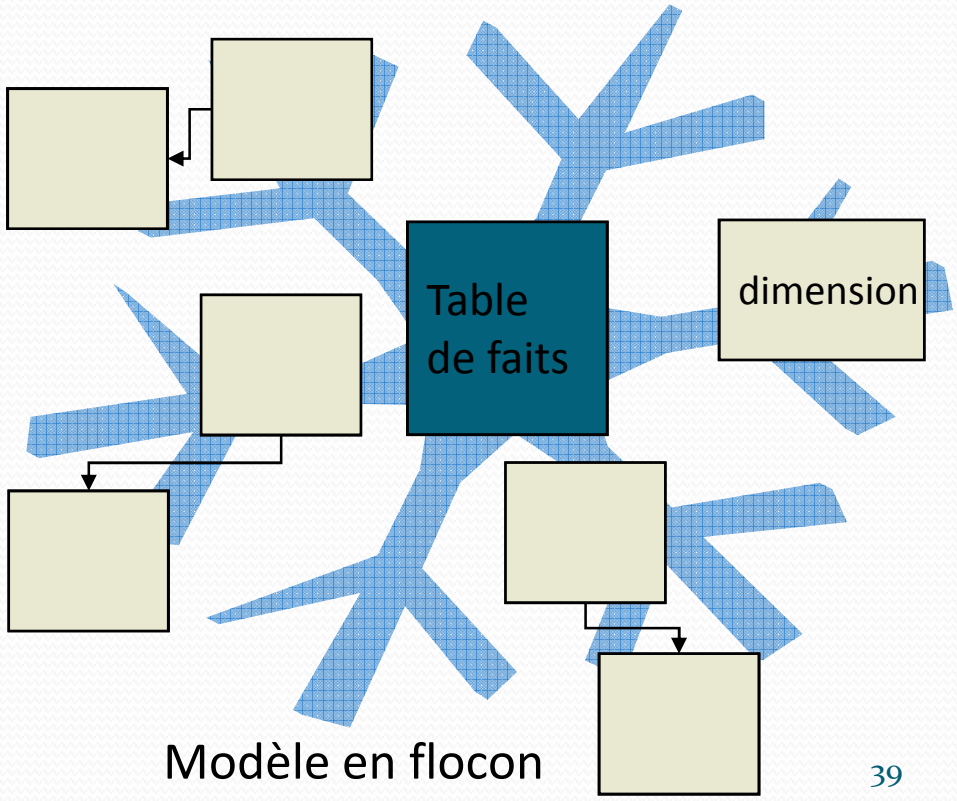
- Caractéristique :

tables de faits + tables de dimensions reliées à une table de faits par une jointure

On obtient un schéma en étoile(dé-normalisation importante) ou flocon



Modèle en étoile



Modèle en flocon

Concepts du modèle dimensionnel

- Intérêt du modèle dimensionnel :

  répondre à des besoins caractéristiques des systèmes décisionnels : la performance et la simplicité des requêtes.

  == > permet d'avoir des requêtes simple

  == > évite les jointures

  == > agrégation des données en une seule opération pour le modèle en étoile

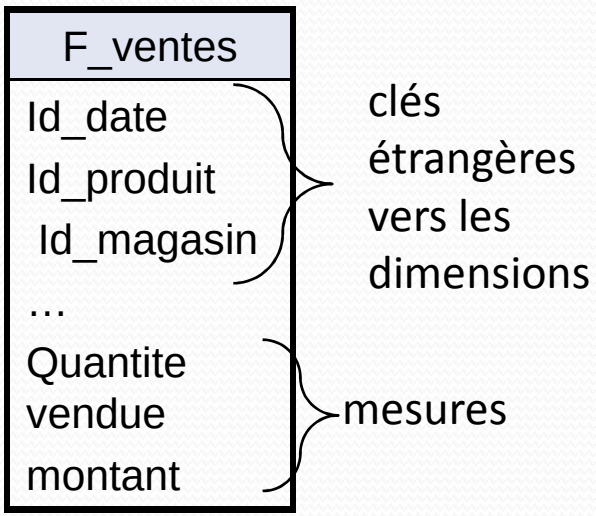
Table de faits

- Objet (sujet) de l'analyse, traduit une activité
- Entité principale du modèle dimensionnel
- Contient les données observables (les mesures) sur le sujet étudié selon divers axes d'analyse (les dimensions)

  == > Grain de mesures de l 'activité : ce que l'on souhaite mesurer (chiffre d 'affaire, nombre de vente, quantités vendue, gain …)

- Une table de fait relie les tables dites de dimension : Contient les clés étrangères des axes d'analyse (dimension)
- Exemple : fait F_ventes

➔ Table de fait contient les valeurs des mesures et les clés vers les tables de dimensions



- On peut avoir plusieurs tables de Fait dans un ED

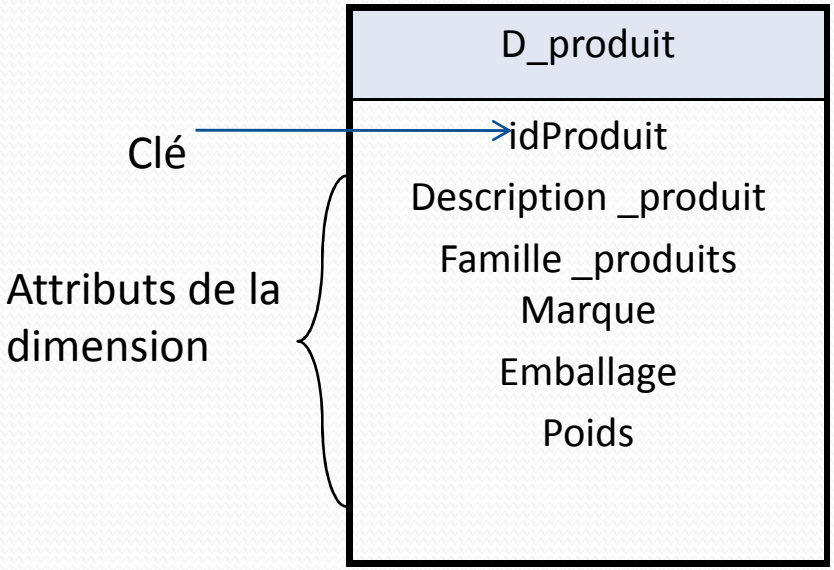
Table de dimension

- Les dimensions sont utilisées pour analyser

les faits. On parle aussi d'axe d'analyse.

== > Axe d'analyse selon lequel vont être étudiées les données observables (mesures)

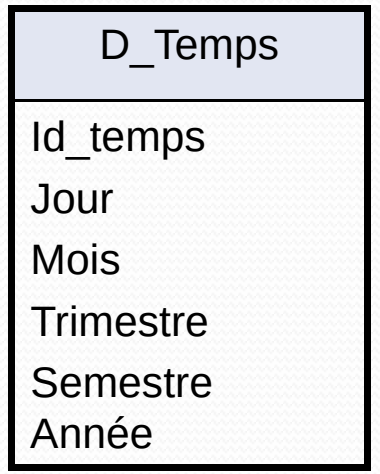
Exemple : Client, produit, période de temps…



Dimension produit

La dimension Temps

- Dans un ED il existe toujours une dimension temps liée à l'historisation
- Commune à l'ensemble du DW
- Reliée à toute table de faits
- Une date peut servir de clé de la table de dimension temps



Dimension Temps

Dimension Dégénérée)

- Une dimension dégénérée est une dimension sans attribut pas de table la clé de la dimension est dans la table de faits

Les types de mesures

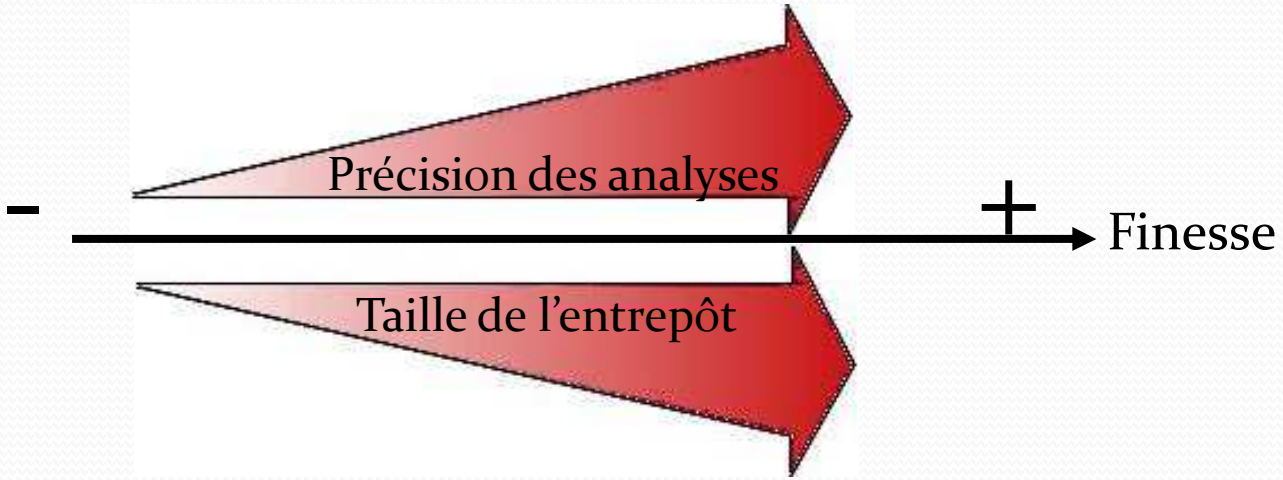
- Trois types de mesures:
  - Additif
  - Semi additif
  - Non additif

Les types de mesures

- Additif : additionnable suivant toutes les dimensions
  - Quantités vendues, chiffre d'affaire
  - Peut être le résultat d'un calcul:
    - Bénéfice = montant vente - coût
- Semi additif : additionnable suivant certaines dimensions
  - Solde d'un compte bancaire:
    - Pas de sens d'additionner sur les dates car cela représente des instantanés d'un niveau
    - Somme sur les comptes: on connaît ce que nous possédons en banque
- Non additif : fait non additionnable quelque soit la dimension
  - Prix unitaire : l'addition sur n'importe quelle dimension donne un nombre dépourvu de sens

Granularité de la table de faits

- Répondre à la question :
  - Que représente un enregistrement de la table de faits?
- La granularité définit le niveau de détails de la table de faits:
  - Exemple: le chiffre d'affaire par produit, par client et par jour

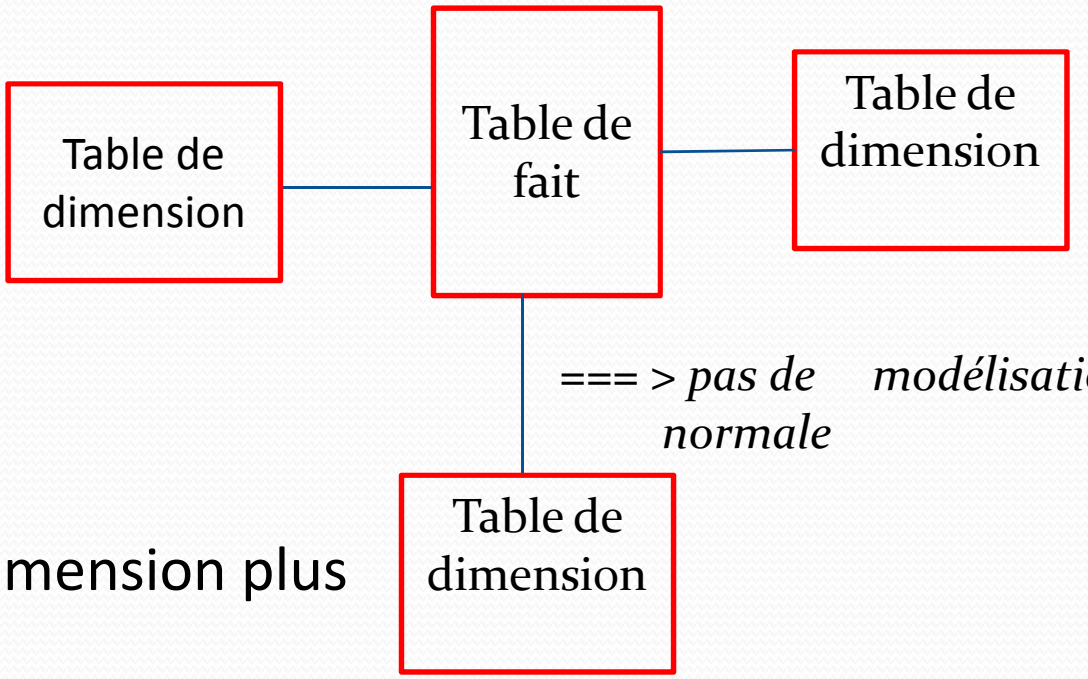


Hiérarchie d'une dimension
Granularité d'une dimension

- 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
  - Exemple :
    - Dimension temps : mois → trimestre → semestre →
    - Dimension Géographie : Ville → Département → Région

Modèle en étoile

- Un schéma sous forme d'étoile : Une table de fait centrale entourée par des tables de dimensions dé-normalisés
- Les dimensions n'ont pas de liaison entre elles
- Les mesures sont stockées dans la table de faits



=== > *pas de modélisation en forme normale*

== > Taille de dimension plus grosse

Modèle en étoile

- Exemple :

On souhaite avoir le CA des ventes d'un produit, par date, client, magasin ainsi que toutes les sommations possibles de chiffre d'affaires dans une année donnée.

== >

- une table de faits **vente** caractérisée par**:**

idproduit, idclient, idvendeur, iddate, montant

- et des tables dimensions **:**

Axes de mesures

- **produit :** idproduit, nom, desc, poids, famille_prod
- **client :** idclient, nom, …
- **magasin :** idmagasin, description
- **temps :** iddate**,** jour, semaine, mois , année

Modèle en étoile

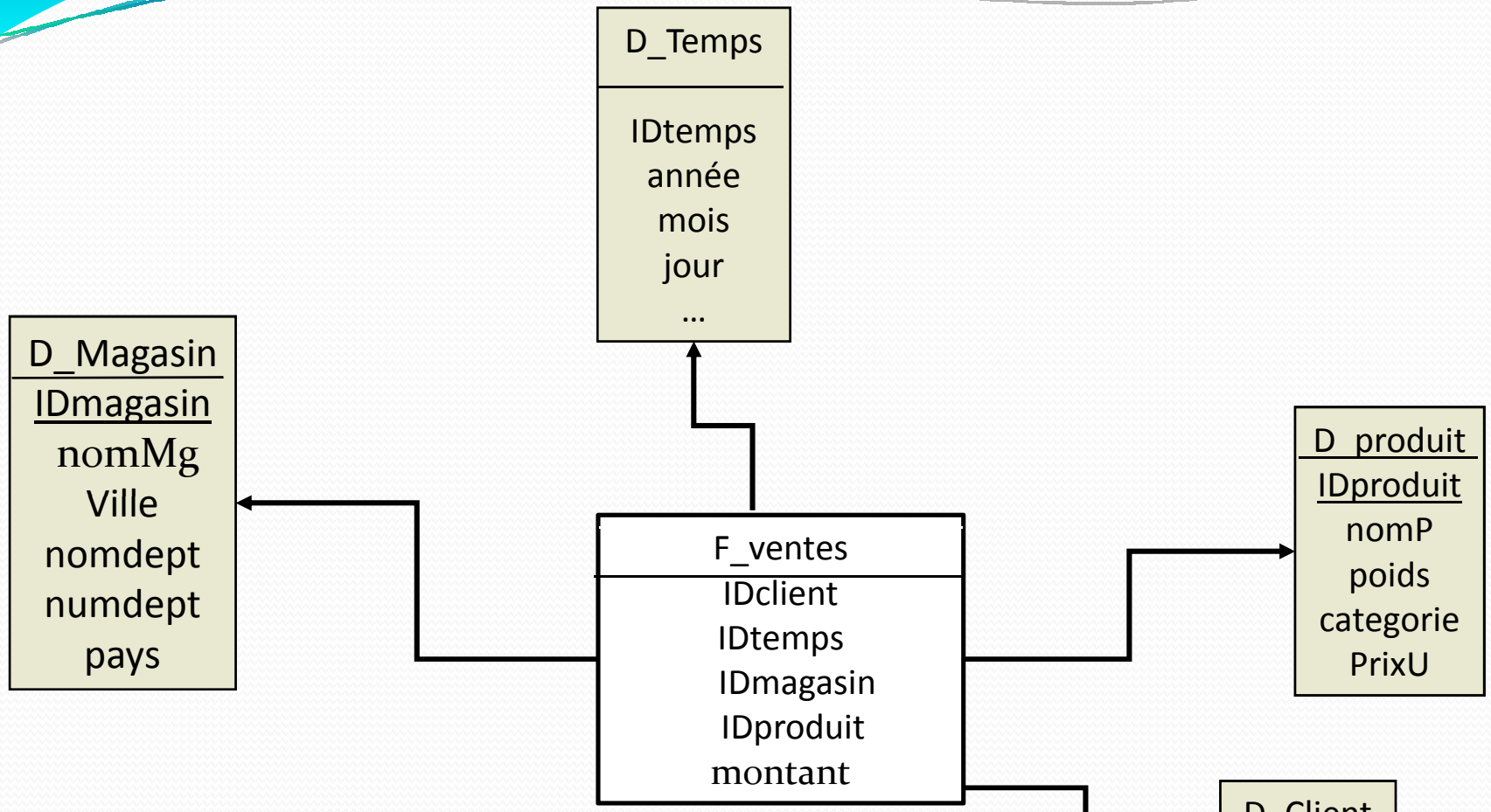


Table de faits vente

Table de faits : F_ventes
Les dimensions : D_Magasin,D_produit,D_client, D_temps

Modèle en étoile

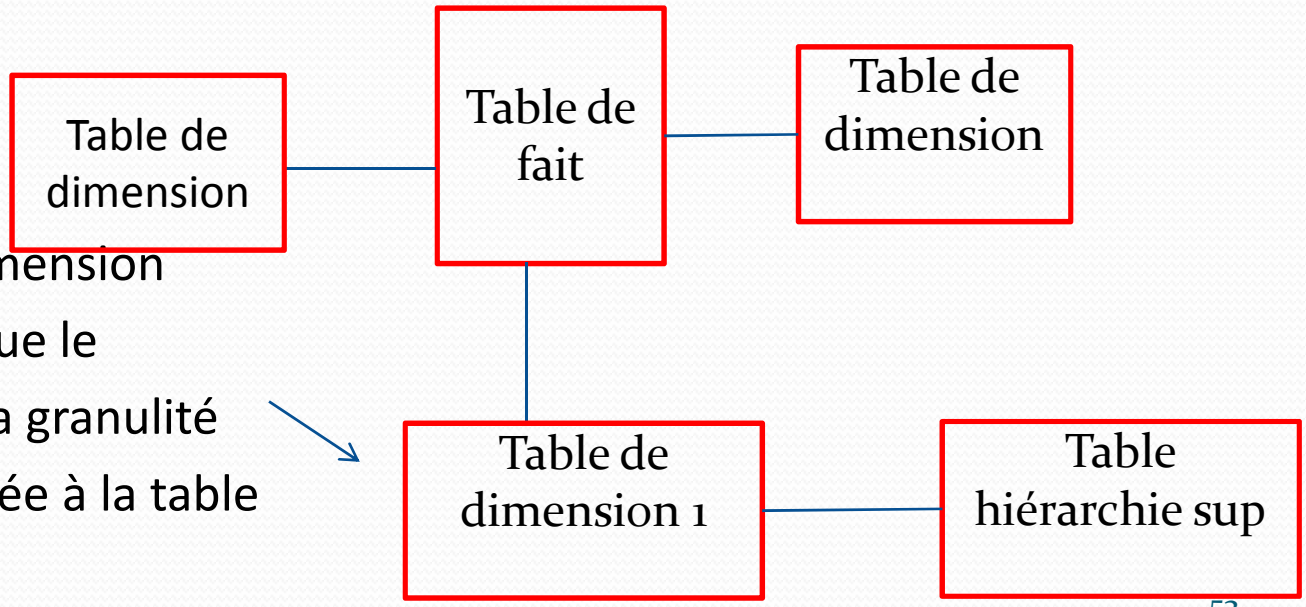
- Avantages:
  - Facilité de navigation
  - Performances : nombre de jointures limité
  - gestion des agrégats
- Inconvénients:
  - Redondance dans les dimensions
  - Alimentation complexe.

Modèle en flocon

- Il est dérivé du schéma en étoile.
- La table des faits reste inchangé.
- Certains tables de dimensions sont normalisées :

  Les dimension en question sont décomposée selon leurs hiérarchies . On obtient ainsi plusieurs niveaux pour ces dimensions. Chaque niveau est représenté dans une table différente

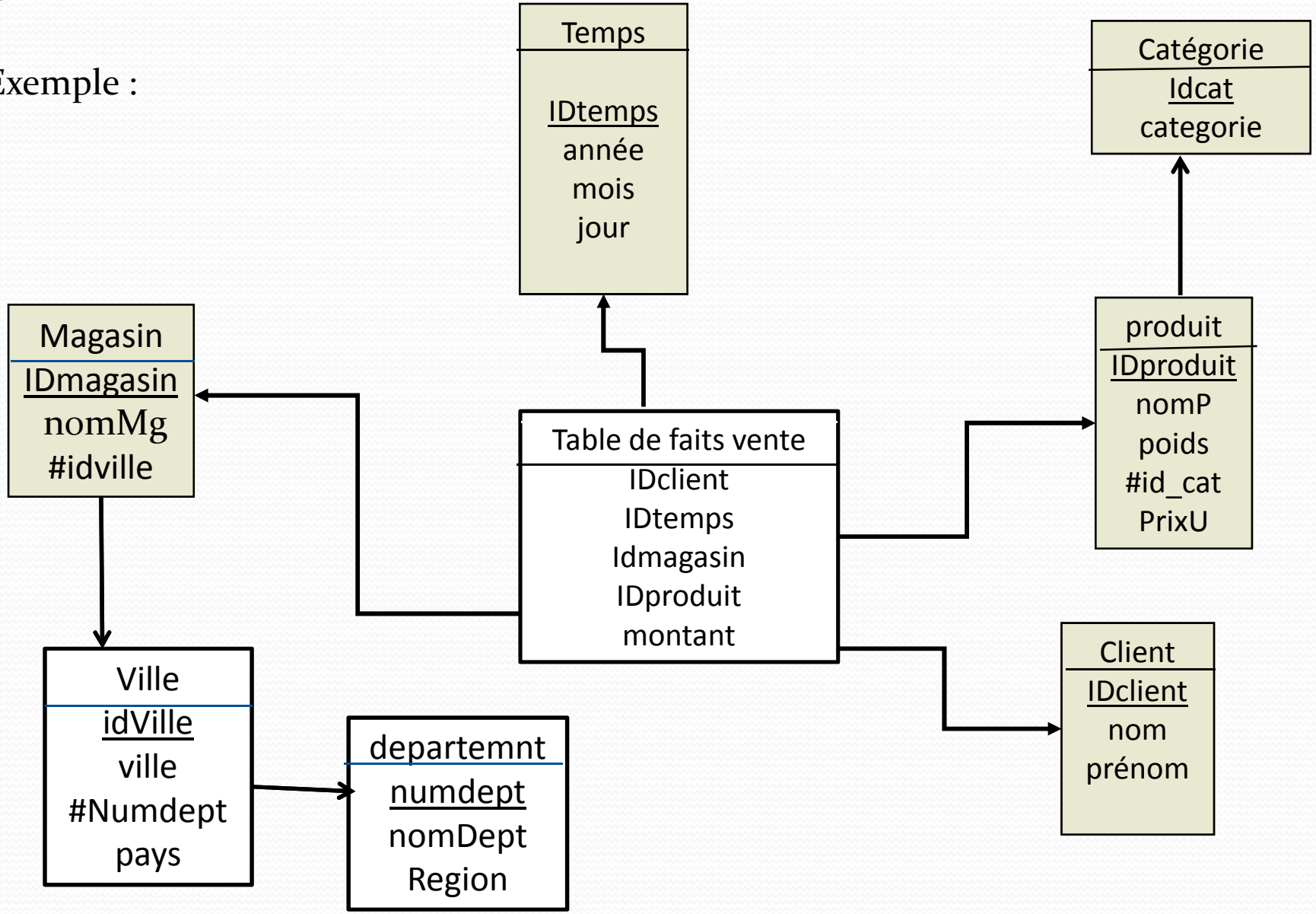
  Exemple : Commune, Département, Région, Pays



- Seul la table de dimension de niveau hiérarchique le plus bas (contenant la granulité la plus faible) est reliée à la table de fait

Modèle en flocon

Exemple :



Modèle en flocon

- Avantages:
  - Normalisation des dimensions
  - Réduction du volume
  - Permet des analyses par pallier (drill down) sur la dimension hiérarchisée.

- Inconvénients:
  - Modèle plus complexe : nombreuses jointures
  - Requêtes moins performantes

Modèle en constellation

On parle de modèle en constellation lorsqu'on dispose de plusieurs tables de faits qui se partagent des dimensions

Processus de normalisation des dimensions (3NF)

- Utilisation de modélisation conceptuelle des hiérarchies :

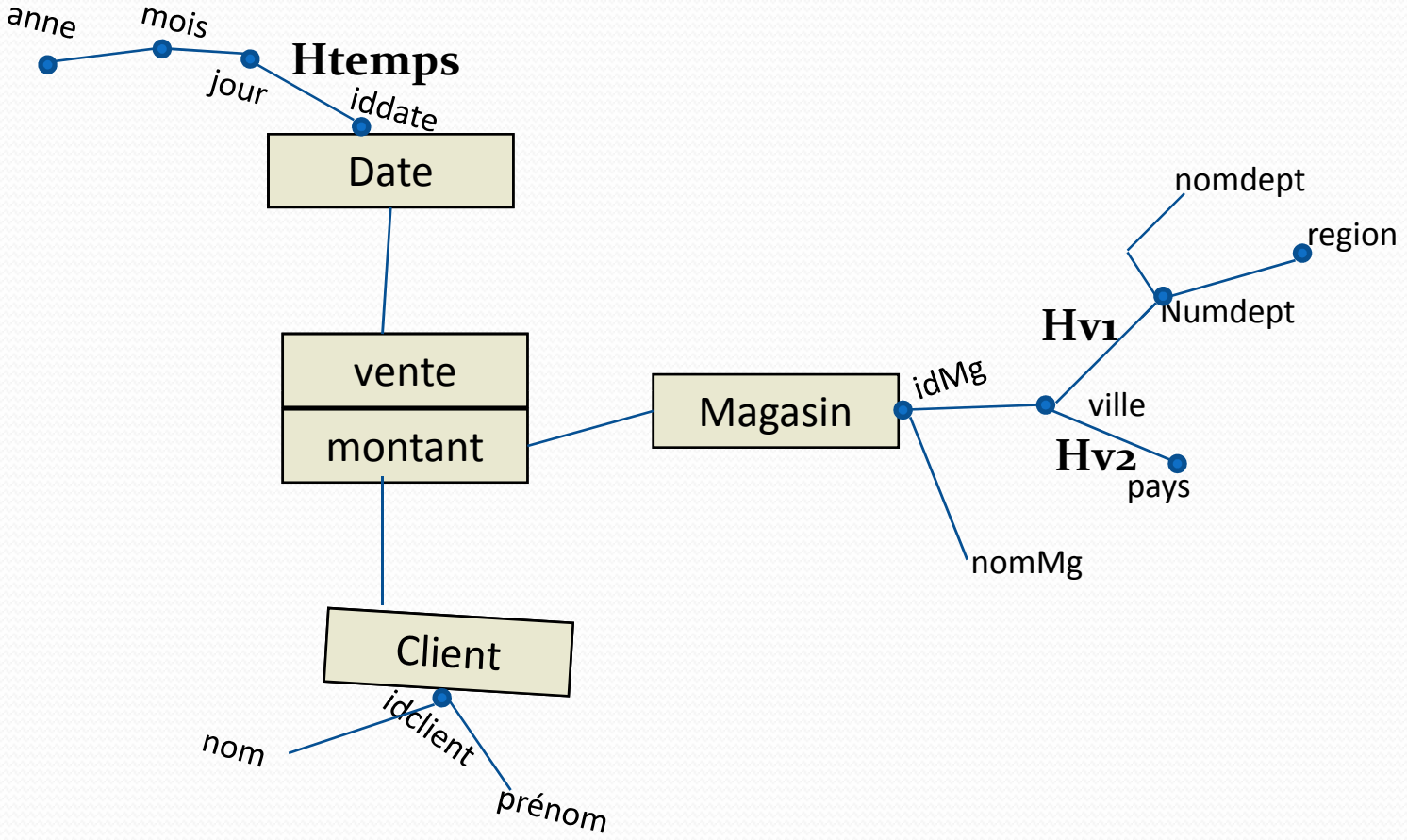
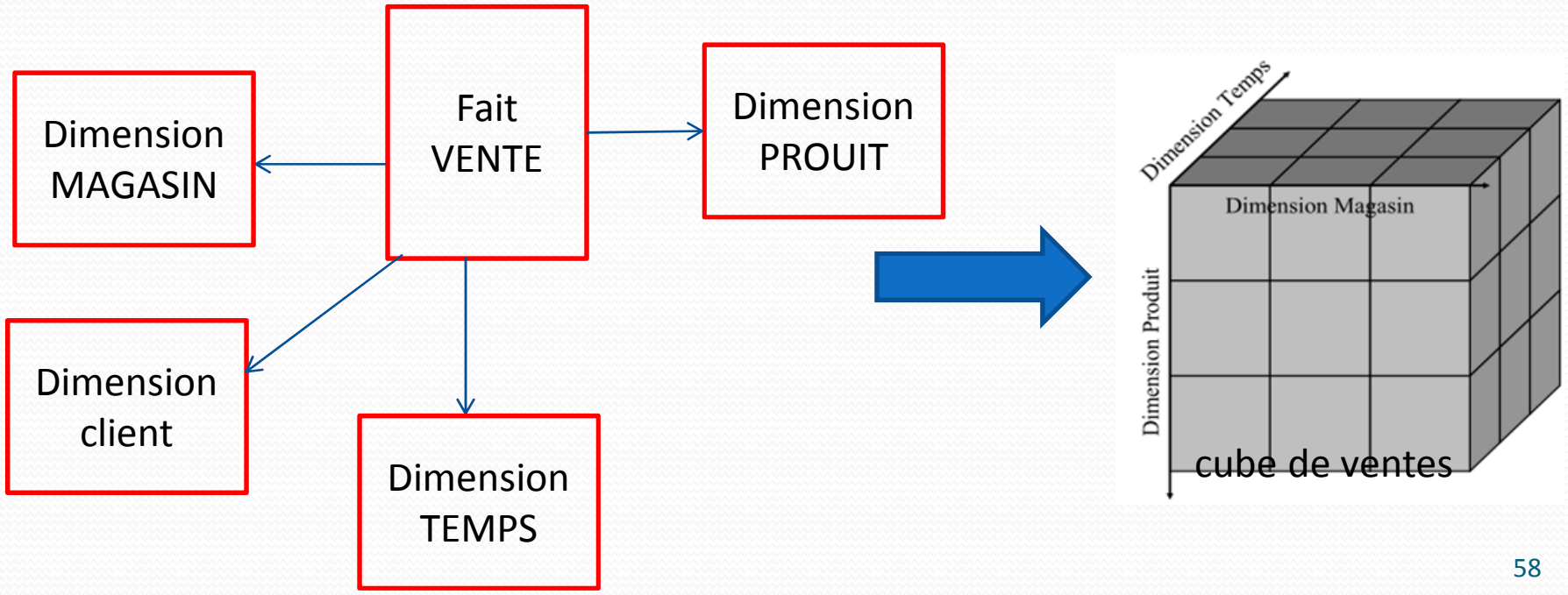


Schéma conceptuel d'une BDM

Cube de données

- Lorsque le nombre de dimensions est de 3, la base de données peut être représentée par un **cube.** Lorsqu'il est supérieur à 3 c'est un **hypercube.** Pour simplifier, on parle dans tous les cas d'un **cube de données**
- Tout se passe comme si les données sont stockées sous forme de cube ou hyperube

➔ La donnée se trouve à l'intersection de n dimensions



Cube de données

Cube :

Une structure permettant de croiser des dimensions pour stocker des variables (les mesures).

Permet de faciliter l'analyse d'une mesure selon différentes dimension

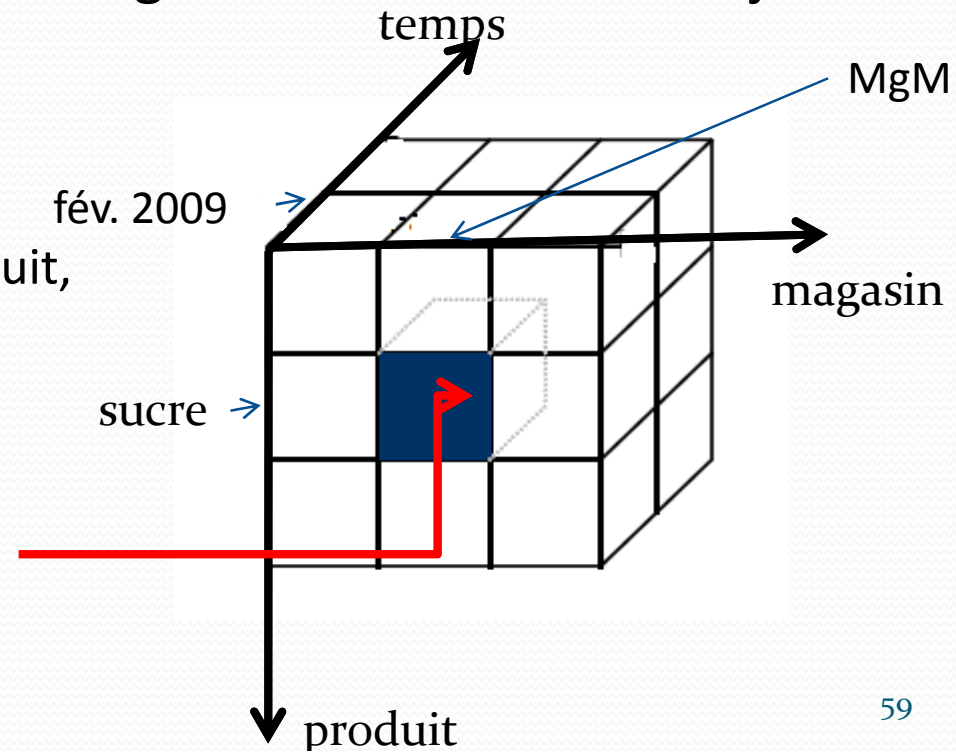
Les calculs sont réalisés lors du chargement ou de la mise à jour du cube

- exemple : un cube de ventes

  qui comprend :

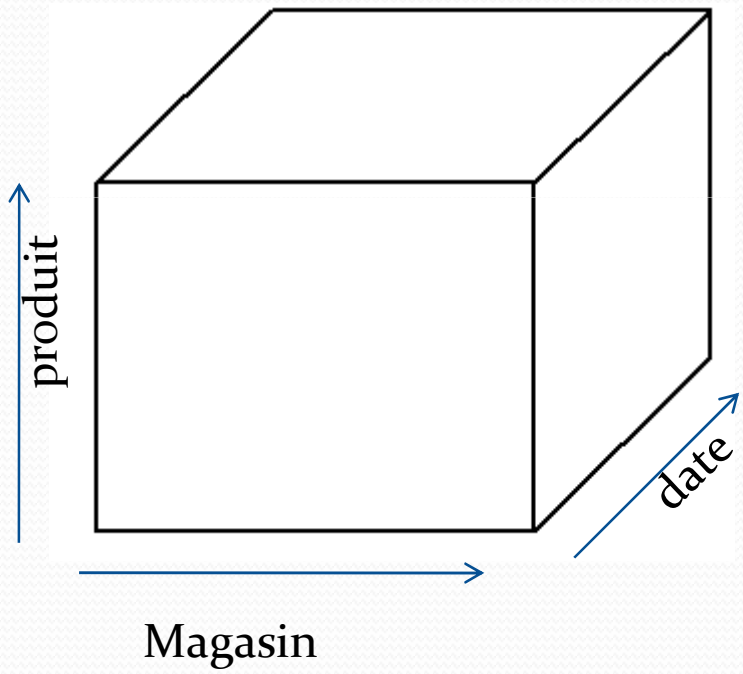
  - Les dimensions temps, produit, magasin
  - La mesure Ventes

Les ventes de sucre pour le magasin « MgM » en fév. 2009

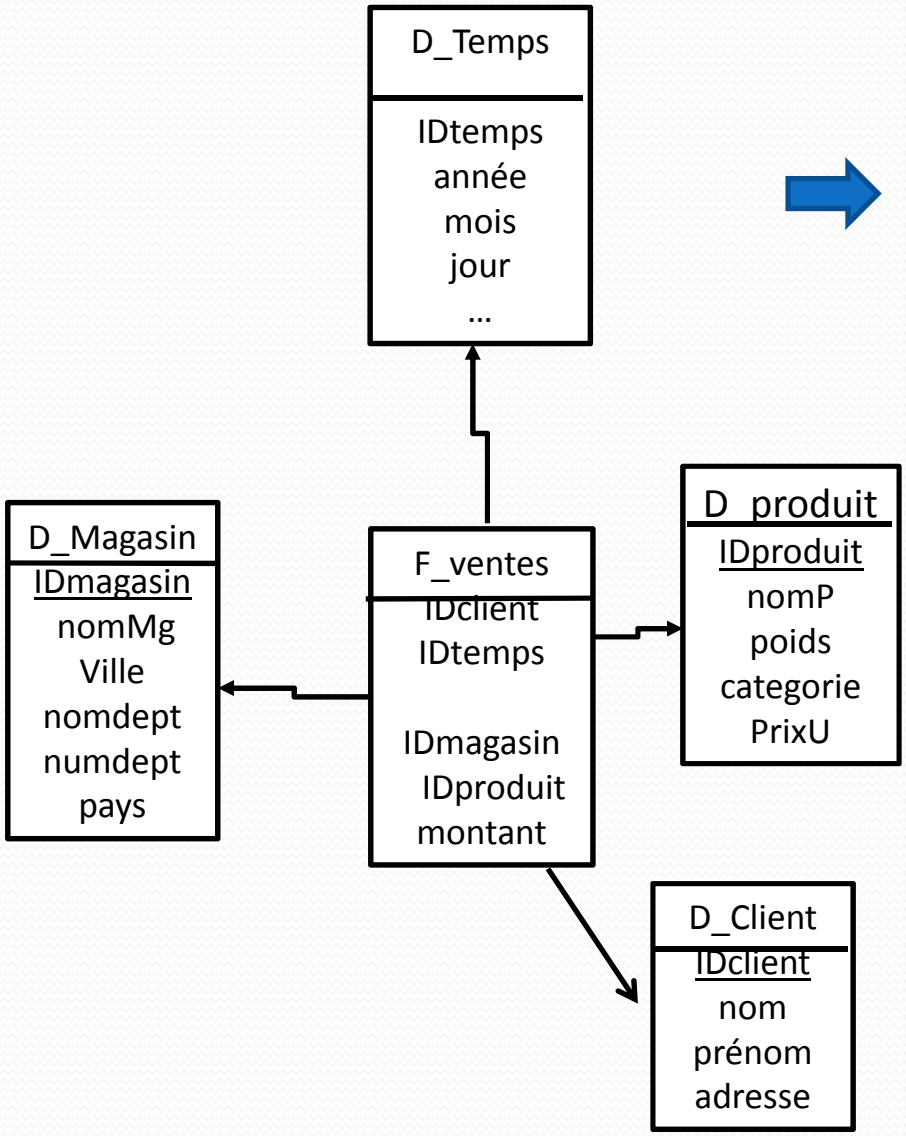


Cube de données

Table Vente			
idMagasin	**idProduit**	**dateVente**	**CA**
1	10	10/01/2000	100
2	20	10/01/2000	200
3	10	10/01/2000	500
1	10	15/01/2000	300
3	40	15/01/2000	100
2	60	16/01/2000	200
4	60	20/02/2000	400
2	10	20/02/2000	200
1	40	25/02/2000	100
4	10	04/03/2000	300
1	20	04/03/2000	200
…			



Cube de données



- Cube à 4 dimensions
- Graphiquement, on peut dessiner en perspective 4 types de cubes à 3 dimensions. (on fixant la valeur d'un axe) . Dans chaque cube, l'élément de base est l'**indicateur « montant ».**

Coupes :
Le but d'un entrepôt de données est la présentation de tableaux de bord.
lorsque le nombre de dimensions du cube de données est n, avec n supérieur
à 2, il faut faire des coupes (en anglais slice and dice) en fixant les valeurs de n-2
dimensions, pour se ramener à un tableau à 2 dimensions, donc affichable.

Agrégation : Au lieu de couper, on peut aussi agréger les données et présenter un tableau à 2 dimensions en *sommant les valeurs de certaines (voire toutes) des n-2 dimensions restantes.*

- l’analyse ascendant

  - exemple : On peut obtenir le CA par vendeur par produit

  - La dimension produit est une hiérarchie car il peut se recomposer en catégorie. La catégorie se décompose

    == > Un produit appartient à une catégorie. On peut donc avoir le CA par vendeur par catégorie de produit : C’est l’analyse ascendant

    L’analyse ascendant/descendant fait partie de l’analyse multi-dimentionnel

Méthodologie: 9 étapes de Kimball

1. **Choisir le sujet ou le processus à modéliser**
2. **Choisir la granularité des faits**
3. **Identifier les dimensions**
4. **Choisir les faits**
5. **Choisir les mesures des faits**
6. Stocker les pré-calculs dans les tables de faits
7. Finaliser les tables de dimensions
8. Choisir la durée de la base
9. Suivre les dimensions lentement évolutives
10. Décider des requêtes prioritaires et des modes de requêtes

Exercices

Exercice 1 :

On souhaite mettre en place une BDR pour une entreprise spécialisée dans la vente des produit. Soit les informations suivante :

- Un produit est caractérisé par son libellé et appartient à une catégorie
- une vente peut concerner plusieurs produit avec une qte spécifique pour chaque produit.
- Une vente est concerné par un seul client qui peut négocier le prix de vente
- une vente est effectuée en une date donnée,
- une vente est réalisé par un seul vendeur appartenant à un service de vente spécialisé dans le produit .

Questions :

-Proposer un schéma entité-association (EA) modélisant cette situation.

- Elaborer le schéma relationnel

Cette dernière entreprise souhaite mettre en place un entrepôt de données pour rassembler toutes les nuits des informations sur les ventes du jour afin de dresser des tableaux de bord sur les ventes.

L'ED à modéliser doit être capable de fournir le chiffre d'affaires des ventes d'un produit, par date, client, et vendeur, ainsi que toutes les sommations possibles de chiffre d'affaires.

Eléments :

Questions :

- Quelle est la table des faits?
- Quels sont les mesures ?
- Quelles sont les dimensions ?. Quelles sont les hiérarchies des dimensions?. Dessinez les
- Donner les schéma conceptuel approprié
- Quel est le schéma relationnel en étoile le plus approprié à cette analyse ?
- De quel type de cube de données dont on dispose ? quel est l'élément de base du cube ?
- Combien de types de cubes à 3 dimension peut-on déduire ?. Faites une représentation d'un du cube OLAP (sans tenir compte des hiérarchies)