Partie 1

Data Warehouse

Entrepôts de données

Chapitre 2 : Modélisation des Data Warehouses

Dr H. EL BOUHISSI Epse BRAHAMI

Plan

- Introduction
- Modélisation multidimensionnelle
- Dimension
- Fait
- Modélisation en étoile
- Modélisation en flocon de neige
- Modélisation en constellation

Rappel

Un entrepôt de données est une base de données qui permet de stocker les données qui sont utilisées dans le cadre de la prise de décision.

Il est approvisionné par les données qui proviennent des bases de données opérationnelles grâces aux outils d'ETL (Extract , Transform et Load).

Modélisation multidimensionnelle

Différents niveaux de modélisation : Niveau conceptuel, Niveau logique, Niveau physique

- **Niveau conceptuel** :

Description de la base multidimensionnelle indépendamment des choix d'implantation.

- **Les concepts:**
  - Dimensions et hiérarchies.
  - Faits et mesures

Modélisation multidimensionnelle

Dans le cadre de la conception des modèles conventionnels de données(MCD) des bases de données classiques on parle des tables et des relations entre elles.

Dans le concept du Business Intelligence ou Informatique décisionnelle nous parlons des dimensions et des faits . Les dimensions sont les axes sur lesquels on veut faire l'analyse.

Modélisation multidimensionnelle : Principes

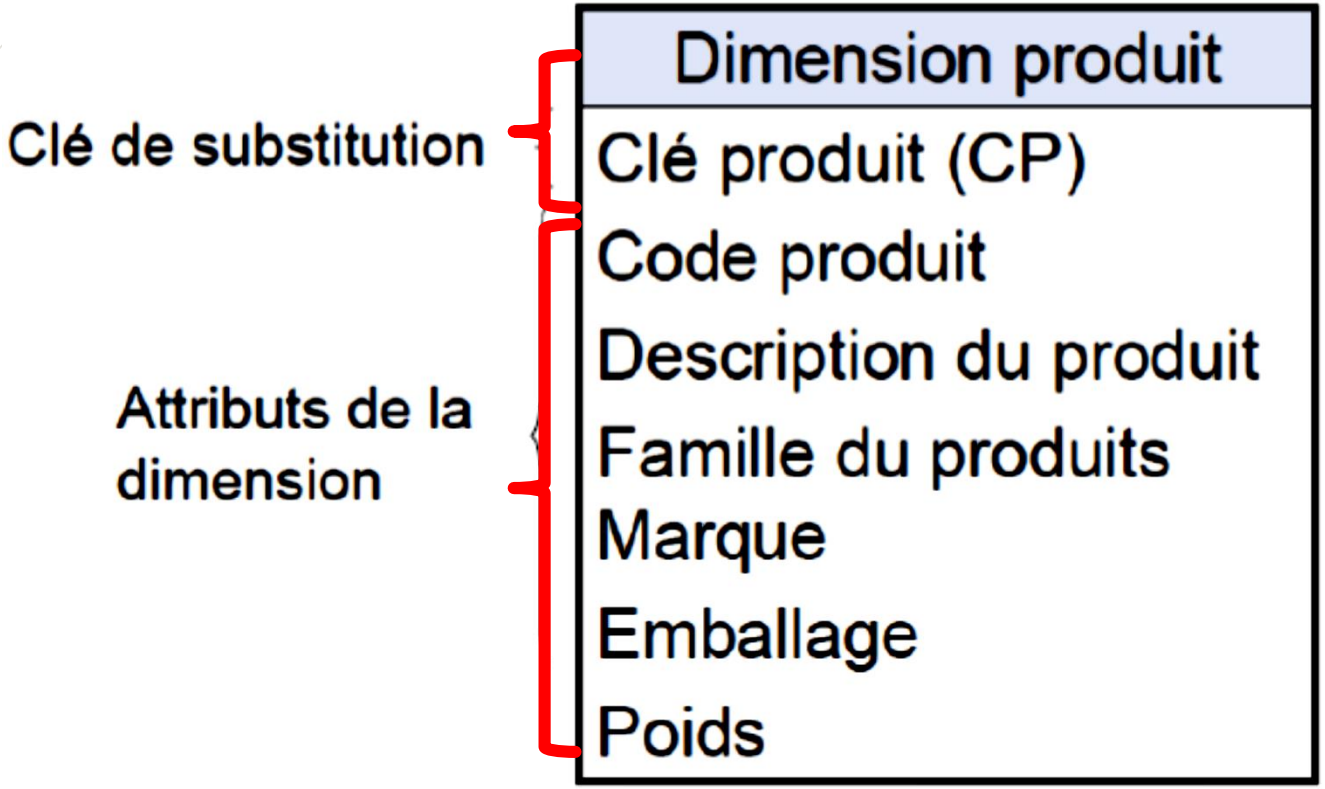
- Intérêt porté sur une partie des données /pas sur la totalité.
- Seulement les données nécessaires à la prise de décision.
- Modèle de données intuitif → vision que portent les analystes et décideurs sur les données.
- Tolérance de violation de certains principes de modélisation classiques (formes normales) en renforçant les contrôles d'intégrité.

Modélisation multidimensionnelle : Dimension

Table qui représente des axes d'analyse avec lesquels on veut faire l'analyse Géographique, temporel, produits, des activités menées au sein d'une société, etc.

- Chaque dimension comporte un ou plusieurs attributs/membres.
- Chaque membre de la dimension a des caractéristiques propres et est en général textuel.

Modélisation multidimensionnelle : Dimension



Structure de base d'une dimension

Modélisation multidimensionnelle : Fait

Les faits sont ceux sur quoi va porter l'analyse. Ceux sont des tables qui contiennent les informations opérationnelles et relatent la vie d'une entreprise.

Par exemple : on peut avoir une table des faits pour la vente qui permet d'évaluer le chiffre d'affaire net, quantités et montants commandés et quantités facturées. Aussi une table des faits pour les ressources humaines qui permet d'évaluer le nombre d'exemplaires d'un produit en stock, niveau de remplissage du stock.

Un fait est ce que nous voulons analyser.

Modélisation multidimensionnelle : Fait

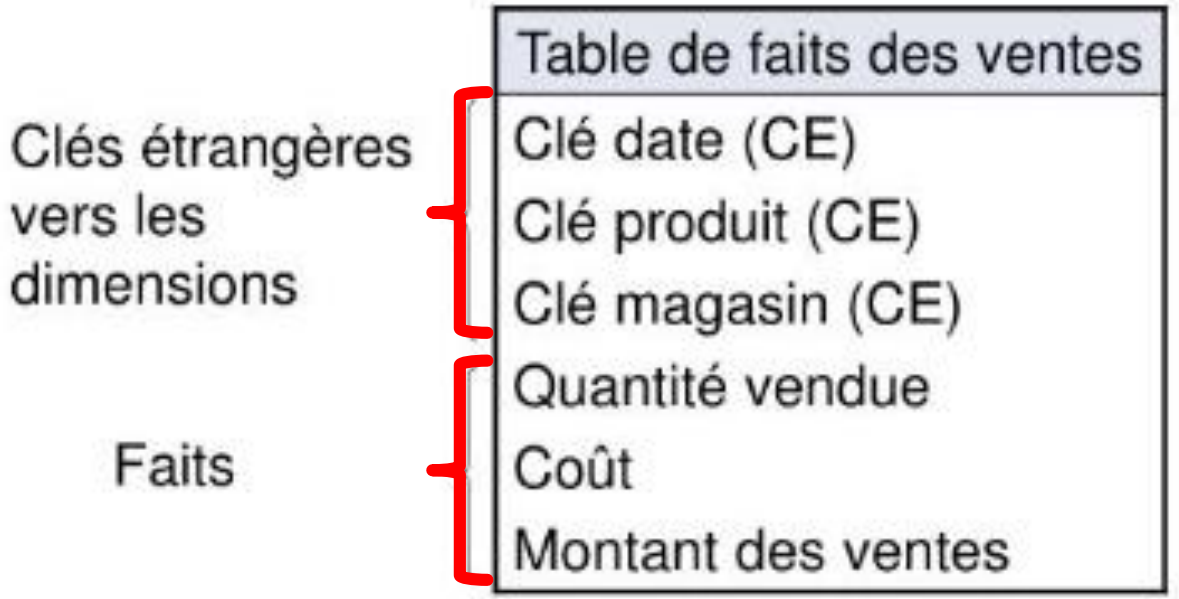
Un fait est la plus petite information analysable. C'est une information qui contient les données observables (**les faits**) que l'on possède sur un sujet et que l'on veut étudier, selon divers axes d'analyse (**les dimensions**).

Les « faits » dans un entrepôt de données, sont normalement numériques, puisque d'ordre quantitatif. Il peut s'agir du montant en argent des ventes, du nombre d'unités vendues d'un produit, etc.

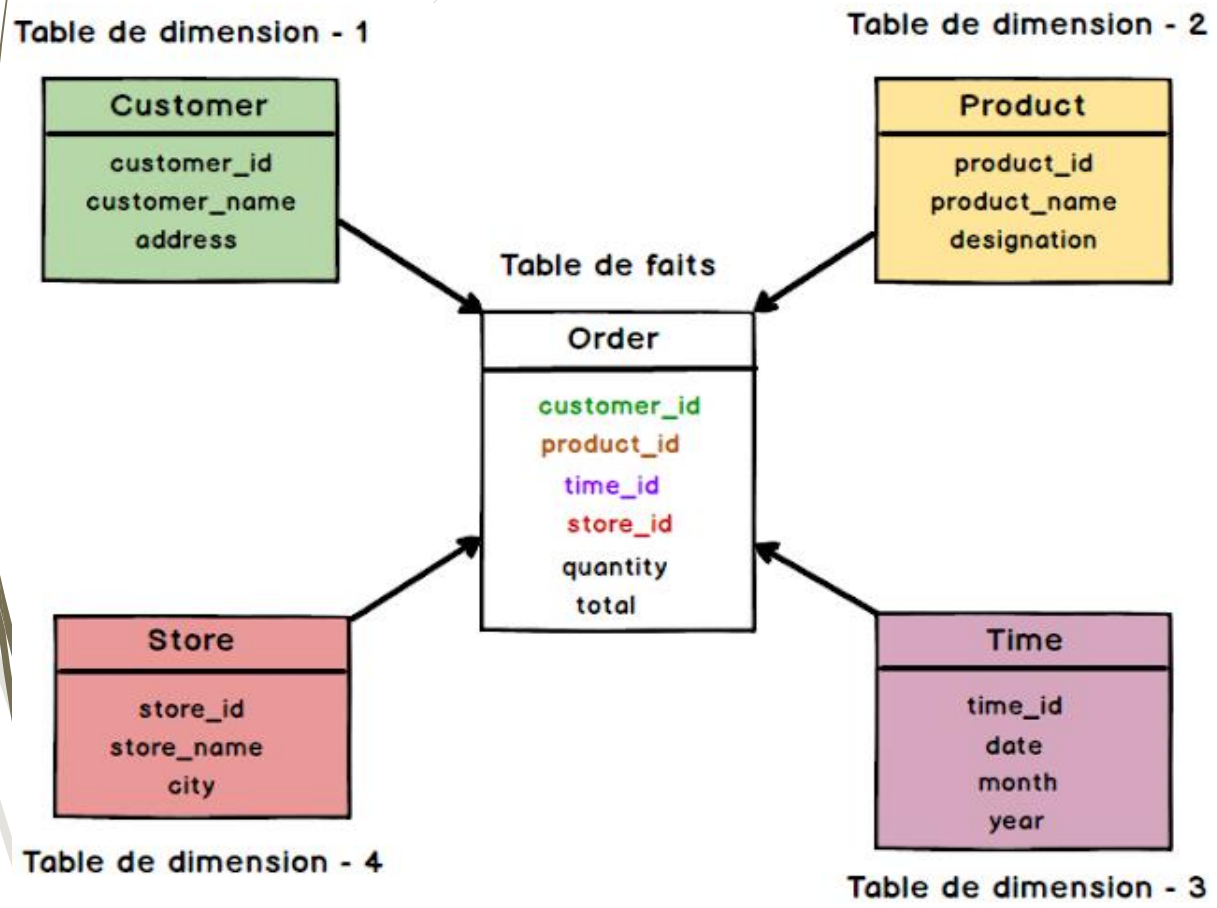
C’est quoi un fait ?

Contient les données observables (faits) sur le sujet étudié selon divers axes d’analyse (les dimensions)

Structure de base d'une "table“ de faits



Dimension VS fait



Les deux sont très importants pour la création d'un schéma, mais **la table de dimension** doit être enregistrée avant **la table de faits**. Comme il est impossible de créer une **table de faits** sans dimensions.

La différence clé entre la table de faits et la table de dimension est que la table de dimension contient des attributs avec lesquels les mesures sont prises dans la table de faits.

Dimension VS Fait

	Table de faits	Table de dimensions
Clé primaire	La table de faits contient une clé primaire qui est une concaténation de clés primaires de toutes les tables de dimensions.	Chaque table de dimension contient sa clé primaire.
Signification	La table de faits contient les mesures avec des attributs d'une table de dimension.	La table de dimension contient les attributs avec lesquels la table de faits calcule la métrique.
Taille de la table	La table de faits se développe verticalement.	La table de dimensions se développe horizontalement.
Attribut & Records	La table de faits contient moins d'attributs et plus d'enregistrements.	La table de dimension contient plus d'attributs et moins d'enregistrements.
Création	La table de faits peut être créée uniquement lorsque les tables de dimensions sont complétées.	Les tables de dimension doivent être créés en premier.
Schéma	Un schéma contient moins de tables de faits.	Un schéma contient plus de tables de dimension.
Les attributs	La table de faits peut contenir des données au format numérique et au format textuel.	La table de dimension contient toujours des attributs au format textuel.

Exercice

On veut construire un entrepôt de données afin de stocker les informations sur les consultations d'un pays. On veut notamment connaître le nombre de consultations, par rapport à différents critères (personnes, médecins, spécialités, etc. Ces informations sont stockées dans les relations suivantes :

PERSONNE (id, nom, tel, adresse, sexe)
MEDECIN (id, tel, adresse, spécialité)
CONSULTATION (id_med, id_pers, date, prix)

1. Quelle est la table des faits?
2. Quels sont les faits?
3. Combien de dimensions ont été retenues? Quelles sont-elles?

Solution

PERSONNE (id, nom, tel, adresse, sexe)
MEDECIN (id, tel, adresse, spécialité)
CONSULTATION (id_med, id_pers, date, prix)

1.Quelle est la table des faits? : **Consultation**
2.Quels sont les faits? **Le prix et le nombre de consultation**
3.Combien de dimensions ont été retenues? Quelles sont-elles?
Trois dimensions : Médecin, Personne Temps

Modélisation multidimensionnelle : Mesure

Élément de donnée sur lequel portent les analyses, en fonction des différentes dimensions.

Ces valeurs sont le résultat d'opérations d'agrégation sur les données.

Exemple :

- Coût des travaux.
- Nombre d'accidents.
- Ventes.
- …

Modélisation multidimensionnelle : Clés

Tables de dimension

- Clé primaire

Tables de fait

Clé composée

- Clés étrangères des tables de dimension

Modélisation multidimensionnelle

Concevoir un Datawarehouse, au niveau conceptuel, on utilise deux modèles :

- **Schéma en étoile (star schema)**: Au milieu, une table de faits connectée à un ensemble de tables de dimensions.
- **Constellation de faits (Factflaked schema):** Plusieurs tables de faits partagent quelques tables de dimension (constellation d'étoiles).

Au niveau logique, il existe 1 modèle :

- **Schéma flocon de neige (snowflake schema):** Un raffinement du précédent où certaines tables de dimensions sont normalisées (donc décomposées). Il provient de la normalisation des tables de dimension.

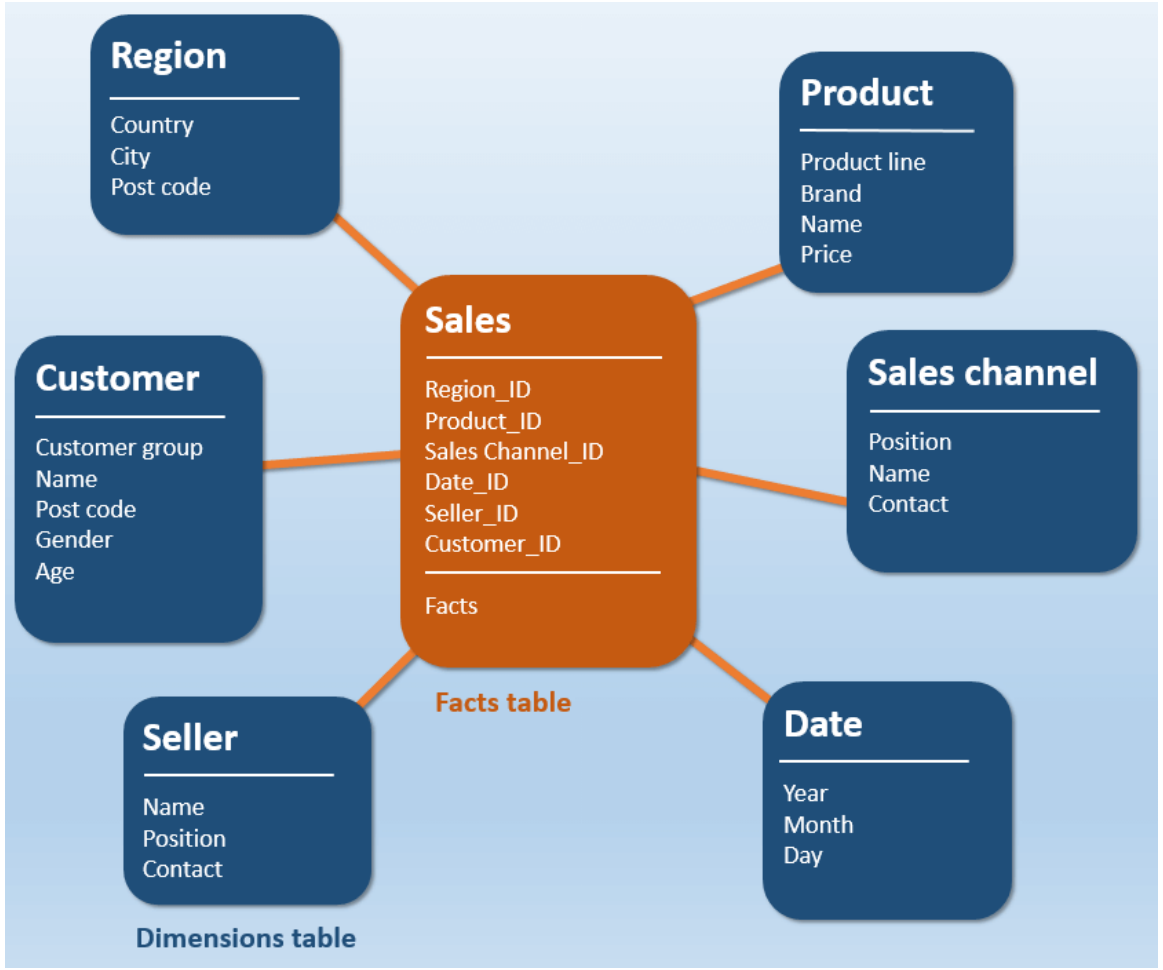
Les modèles en étoile et flocon de neige sont les plus utilisés en entreprises. Les deux composants principaux de ces modèles sont les dimensions et les faits.

DWH – Modèle en étoile

C'est une manière de relier une **dimension** à un **fait** dans un entrepôt de données.

Dans le modèle en étoile, on a une table de fait centrale qui est liée par les tables de dimensions dénormalisées.

Les **dimensions** ne sont pas liées entre elles.



DWH – Modèle en étoile

Avantages

- Facilité de navigation.
- Performances : nombre limité de jointures ; gestion des données creuses.
- Gestion des agrégats.
- Fiabilité des résultats.

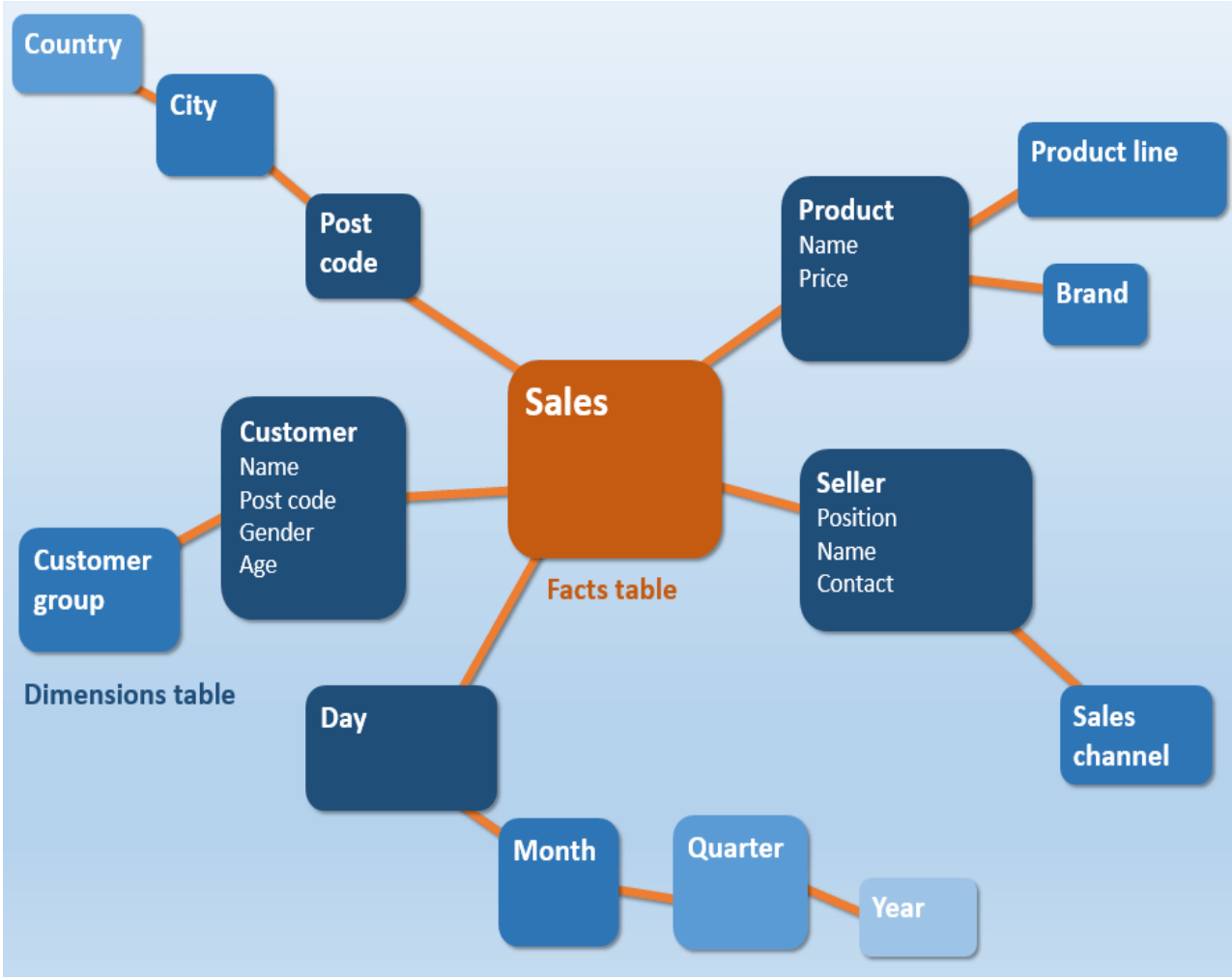
Inconvénients

- Toutes les dimensions ne concernent pas les mesures.
- Redondances dans les dimensions.
- Alimentation complexe.

DWH – Modèle en flocon de neige

C’est une manière de relier une dimension à un fait dans un entrepôt de données. C’est le modèle en étoile avec une normalisation des dimensions.

Il peut exister des hiérarchies des dimensions pour diviser les tables de dimensions lorsqu’elles sont trop volumineuses.
Les informations redondantes sont stockées dans des tables de dimensions distinctes, classifiées et hiérarchisées.



DWH – Modèle en flocon de neige

- Un seul niveau hiérarchique par table de dimension.
- La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait (elle a la granularité la plus fine)

Avantages :

- Normalisation des dimensions.
- Economie d'espace disque (réduction du volume).

Inconvénients :

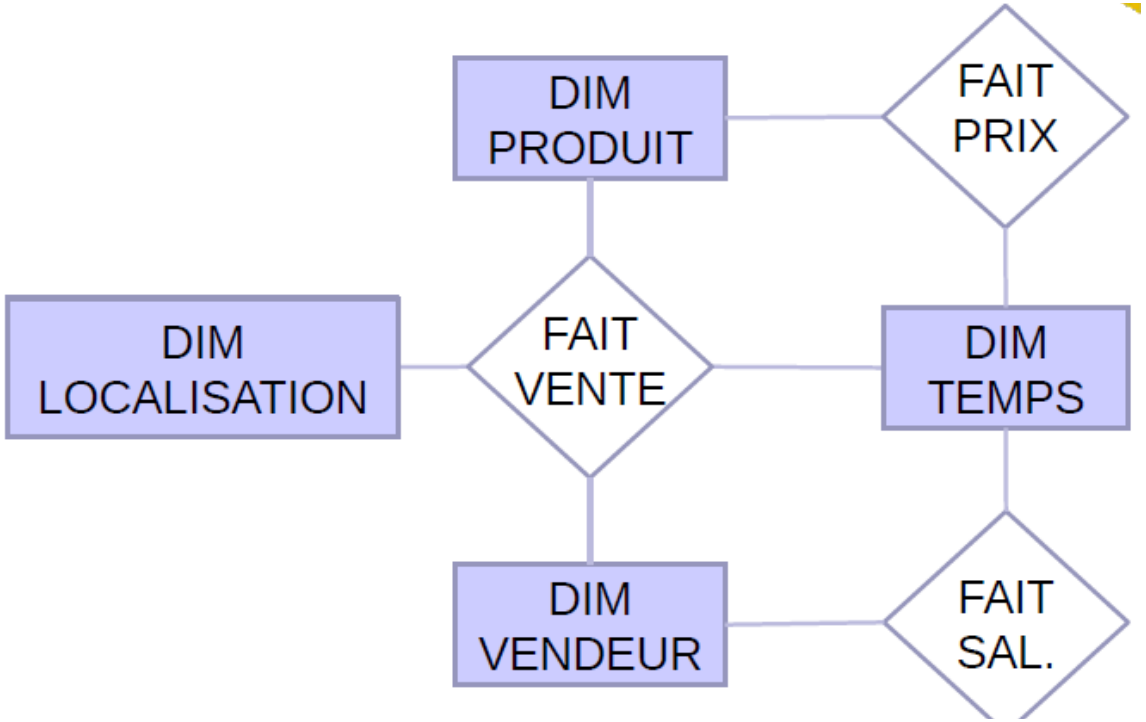
- Modèle plus complexe (nombreuses jointures).
- Requêtes moins performantes.

DWH – Modèle en constellation

Une série d'étoiles :

- Fusion de plusieurs modèles en étoile qui utilisent des dimensions communes.

- Plusieurs tables de fait et tables de dimensions éventuellement communes.



Modèle en étoile – Modèle en flocon de neige

Les schémas en flocon se caractérisent par une consommation d'espace de stockage plus faible que les schémas en étoile. Ceci résulte d'un stockage de données normalisé. **La normalisation se réfère au transfert des colonnes vers de nouvelles tables dans le but d'éviter les doublons**.

L'élimination des redondances réduit aussi le coût de maintenance et de gestion des données : dans le meilleur des cas, chaque information n'apparaît qu'une seule fois et ne doit donc être placée qu'une fois dans le schéma.

Dans la pratique, la structure de données d'un DWH est généralement basée sur le schéma en flocon, tandis que les data marts individuels sont implémentés en tant que schéma en étoile.

Modèle en étoile – Modèle en flocon de neige

Les schémas en flocon se caractérisent par une consommation d'espace de stockage plus faible que les schémas en étoile. Ceci résulte d'un stockage de données normalisé. **La normalisation se réfère au transfert des colonnes vers de nouvelles tables dans le but d'éviter les doublons**.

L'élimination des redondances réduit aussi le coût de maintenance et de gestion des données : dans le meilleur des cas, chaque information n'apparaît qu'une seule fois et ne doit donc être placée qu'une fois dans le schéma.

Dans la pratique, la structure de données d'un DWH est généralement basée sur le schéma en flocon, tandis que les data marts individuels sont implémentés en tant que schéma en étoile.

Notion de Hiérarchie

Les dimensions peuvent être déployées en hiérarchies fonctionnelles, organisationnelles, spatiales ou temporelles.

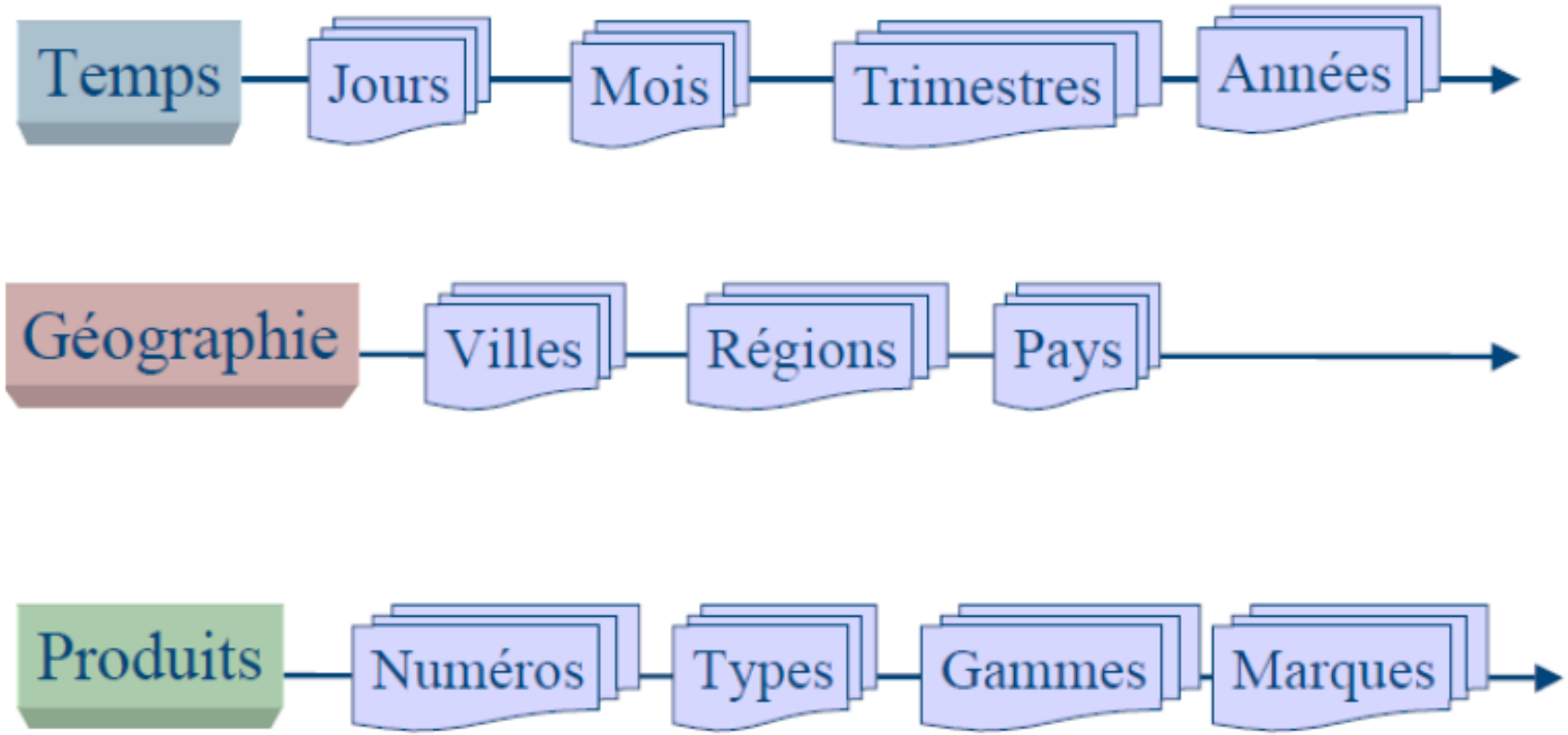
Chaque membre appartient à un niveau hiérarchique (ou niveau de granularité) particulier

Exemples :

- Dimension temporelle: jour, mois, année
- Dimension géographique: magasin, ville, région, pays
- Dimension produit: produit, catégorie, marque, etc.

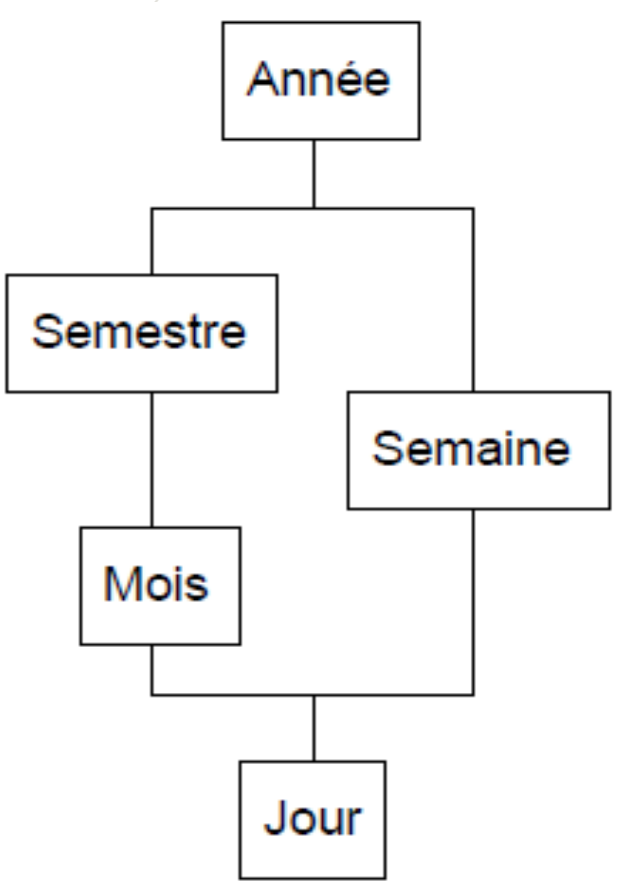
Attributs définissant les niveaux de granularité sont appelés paramètres
Attributs informationnels liés à un paramètre sont dits attributs faibles

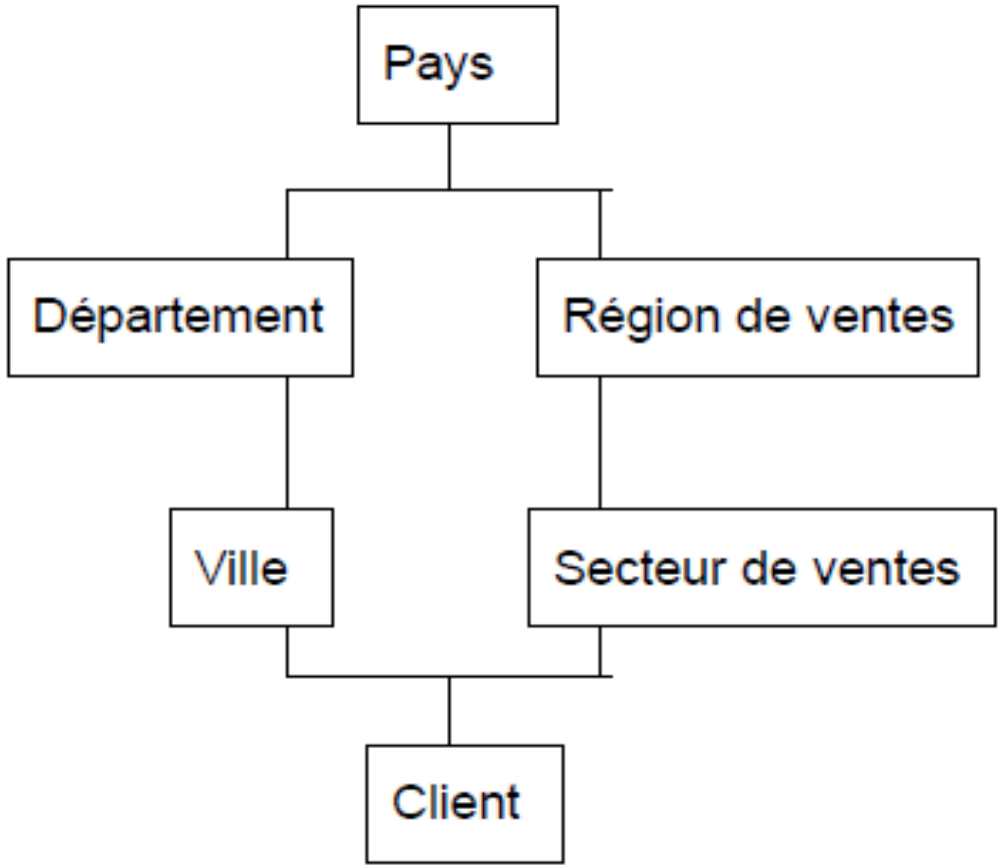
Mono-hiérarchie



Hiérarchies multiples dans une dimension

Plusieurs hiérarchies alternatives pour une même dimension)





Granularité

Niveau de détail de la table de dimension.
Suit la granularité de toutes les dimensions → dépend des besoins d'analyse

Exemples :

- Date: jour, mois, année, décennie
- Produit : produit, catégorie

Exemples de granularité de la table de fait :

analyser le volume de vente par

- Produit, Jour, Ville
- Catégorie, Jour, Ville

La définition de la granularité dépend des besoins d'analyse et de la disponibilité des données détaillées.

Fin

Dr. H. EL BOUHISSI BRAHAMI