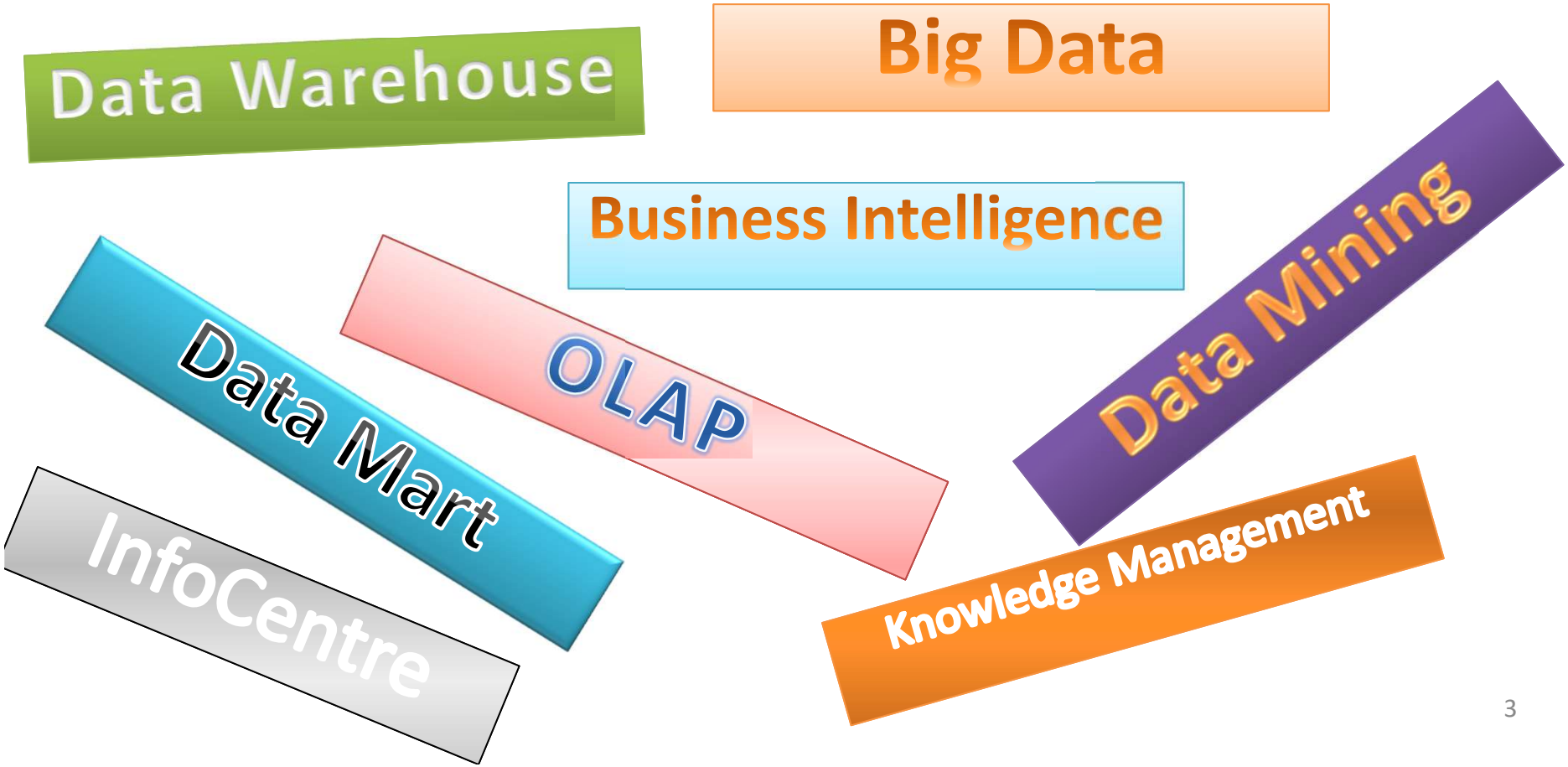
Data WareHouse

Plan

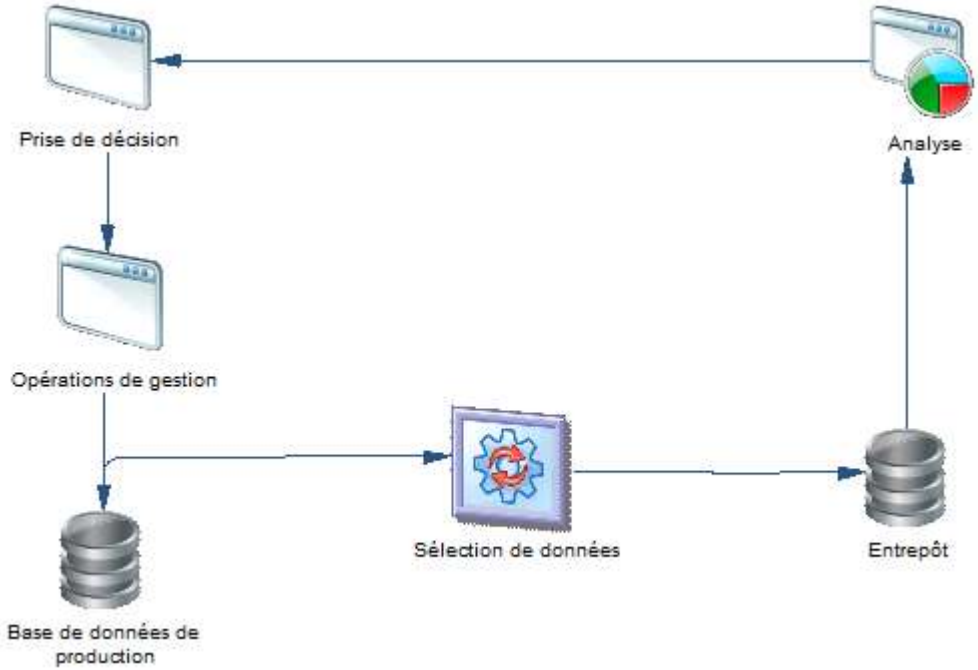
- Introduction
- Les entrepôts de données (Datawarehouse)
- Les datamarts
- Architecture
- Modélisation.

Présentation

- **Besoin:** prise de décisions stratégiques et tactiques
- **Quoi:** productivité de l'entreprise, réactivité des hommes, clients
- **Qui:** le système de pilotage de l'entreprise (Décideurs)



Le processus de prise de décision



Définition d'un DW

- Le Data warehouse (entrepôt de données) est Une collection de données orientées sujet, intégrées, non volatiles et qui varie dans le temps, organisées pour le support d'un processus d'aide à la décision (Définition: [W. H. Inmon])

  - Sujet
    - Les données sont structurées par **sujet** ou par thème (clients, produits, personnel…)
  - Données intégrées
    - Les données sont issues du SIO de l'entreprise et éventuellement de sources externes à l'entreprise.
    - Les différentes données provenant de sources différentes (BDR, XML, fichiers plats,…) et hétérogènes sont **intégrées** et homogénéisées dans une structure unique.

Définition d'un DW

- Homogénéisation:
  - Synonymie :Par exemple deux attributs nom_salarié et nom_employe dans deux sources différentes désignent la même entité.
  - Homonomie: deux noms identiques qui désignent des entités différentes.
  - Une même information peut être exprimée dans deux sources avec des types ou des unités différentes.
- Les données sont **non volatiles et historisées**: la portée temporelle des données dans un DW et plus longue que celle des BDO.
  - BDO: valeur courante des données . Les autres données sont soit détruites soit archivées.
  - DW: les données sont historisées
  - En général , dans un DW chaque donnée fait référence au temps.

Les données pertinentes

- Sources de données
  - Sources internes
    - Bases de données de production
    - Bases créées par les utilisateurs (bases relationnelles, fichiers plats).
  - Sources externes
    - Internet.
    - Organismes
- Caractéristiques de ces données:
  - Dispersées et hétérogènes
  - Détaillées
  - Peu/pas adaptées à l'analyse
  - Volatiles: pas d'historisation systématique

- Données pertinents
  - informations dont la variation permet de dévoiler des dysfonctionnements ou même prévoir des problèmes futurs
  - Types d'indicateurs
    - Indicateurs internes: produits, services, fonctionnement, Personnel
    - Indicateurs entrants/sortants: relations clients/fournisseurs
    - Indicateurs externes: pouvoir d'achat des consommateurs, Réglementation, conjoncture du marché, concurrence, tendance technologique…

Domaines d'applications

- Déterminer et contrôler la performance de l'entreprise
- Mesurer et gérer les risques financiers.
- Planifier la stratégie Achat.
- Banque
  - Risques d'un prêt, prime plus précise
- Assurance
  - Risque lié à un contrat d'assurance (voiture)
- Santé
  - Épidémiologie
  - Risque alimentaire
- Marketing
  - Améliorer la connaissance client
  - Ciblage de clientèle
  - Déterminer des promotions
- Logistique
  - Adéquation demande/production

Data Marts ou magasins de données

- C'est un DW spécialisé dans un sujet ou un métier particulier (Finance, Marketing,…).
- Intérêt d'un DataMart
  - Moins de données à gérer
  - Amélioration des temps de réponse
  - Plus simple à mettre en œuvre qu'un DW

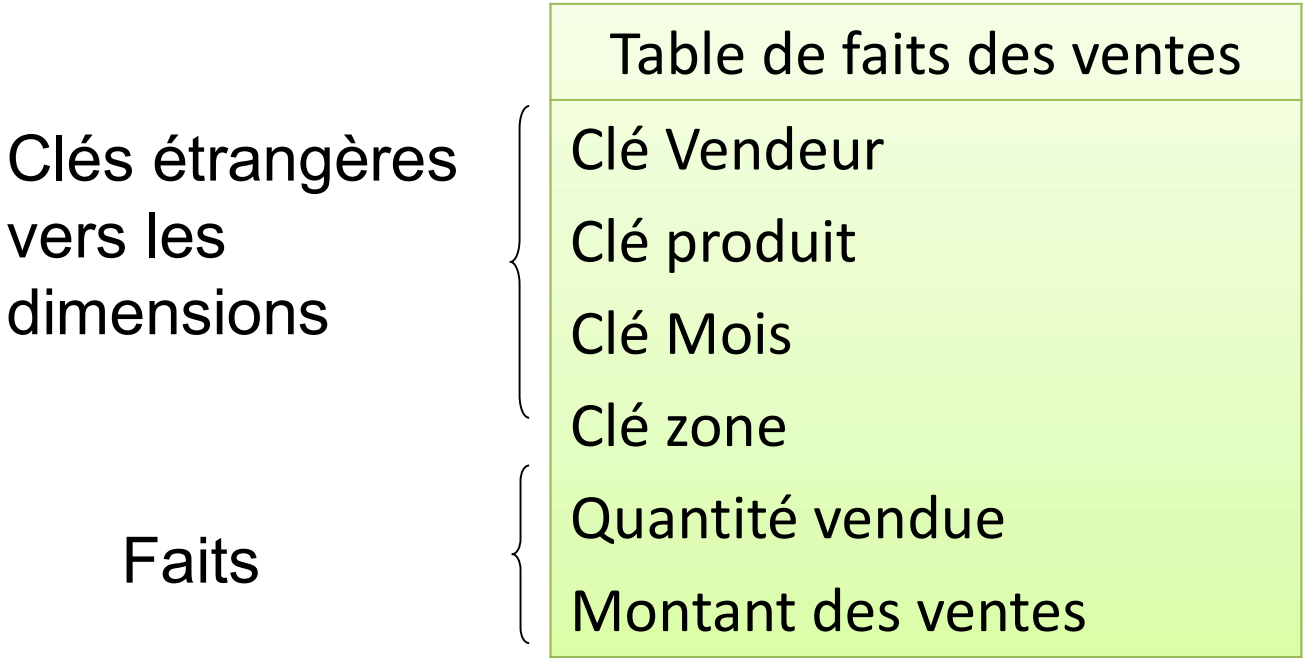
Modélisation d'un DW

- Inconvénients du modèle Entité/Relation
  - Schéma très/trop complexe pour l'analyse des données
  - Inapproprié pour l'analyse

- Le modèle multidimensionnelle
  - Concepts
    - Les faits: mesurent l'activité (exemple: quantité vendue)
    - Dimensions: Axes d'analyse
    - Attributs des dimensions
  - Opérations sur les données
    - Drill Down: une donnée agrégée est visualisée à un niveau de détail plus fin
    - Consolidation: les données sont visualisées à un niveau plus agrégé
    - Slicing and Dicing : visualisation des données selon différentes perspectives.
  - Principe
    - Ne pas trop normaliser les tables

Table de faits

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



Types des faits

- Fait additif: additionnable suivant toutes les dimensions (ex: chiffre d'affaire)
- Fait semi additif: additionnable seulement suivant certaines dimensions
  - Exemple : nombre de clients, dimension produit (un même client peut acheter plusieurs produits) .
- Fait non additif: non additionnable quelque soit la dimension (comptage des faits ou affichage 1 par 1, ex: prix unitaire d'un produit).

Granularité ou finesse de la table de faits

- La granularité définit le niveau de détails de la table de faits
  - mois, jour, heure du jour
  - région ,magasin , rayonnage

Table de dimension

- Axe d'analyse selon lequel vont être étudiées les faits
- Contient le détail sur les faits
- Dimension = axe d'analyse
  - Client, produit, temps…
- Granularité d'une dimension : nombre de niveaux hiérarchiques (ex: continent, pays, région, ville)

- La clé primaire d'une table de dimension est appelée clé de substitution (Surrrogate key)
- « Code produit » est la clé de production (clé utilisée dans la base de données de production

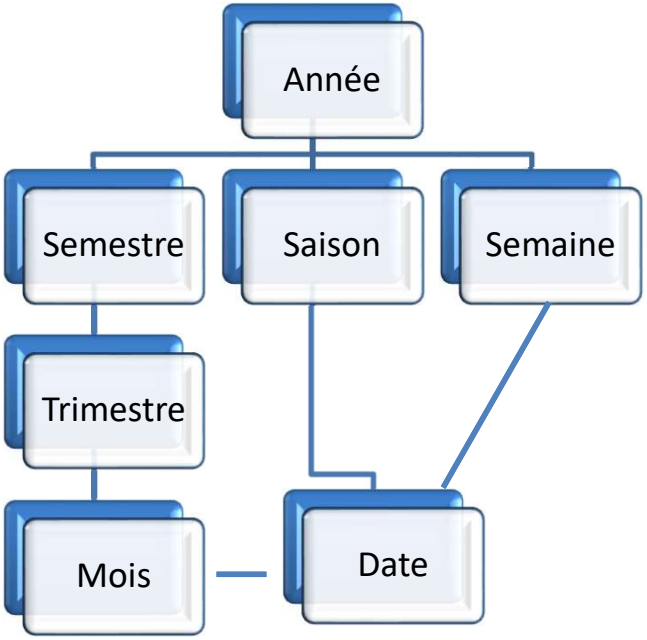
Clé primaire

Attributs de la dimension

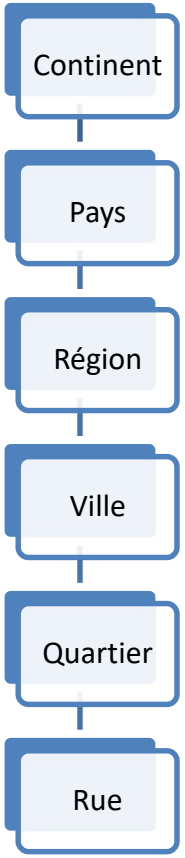
Dimension produit
Clé produit (CP)
Code produit
Description du produit
Famille du produit
Marque
Emballage
Poids

Hiérarchie des dimensions

Hiérarchie multiple



Hiérarchie simple

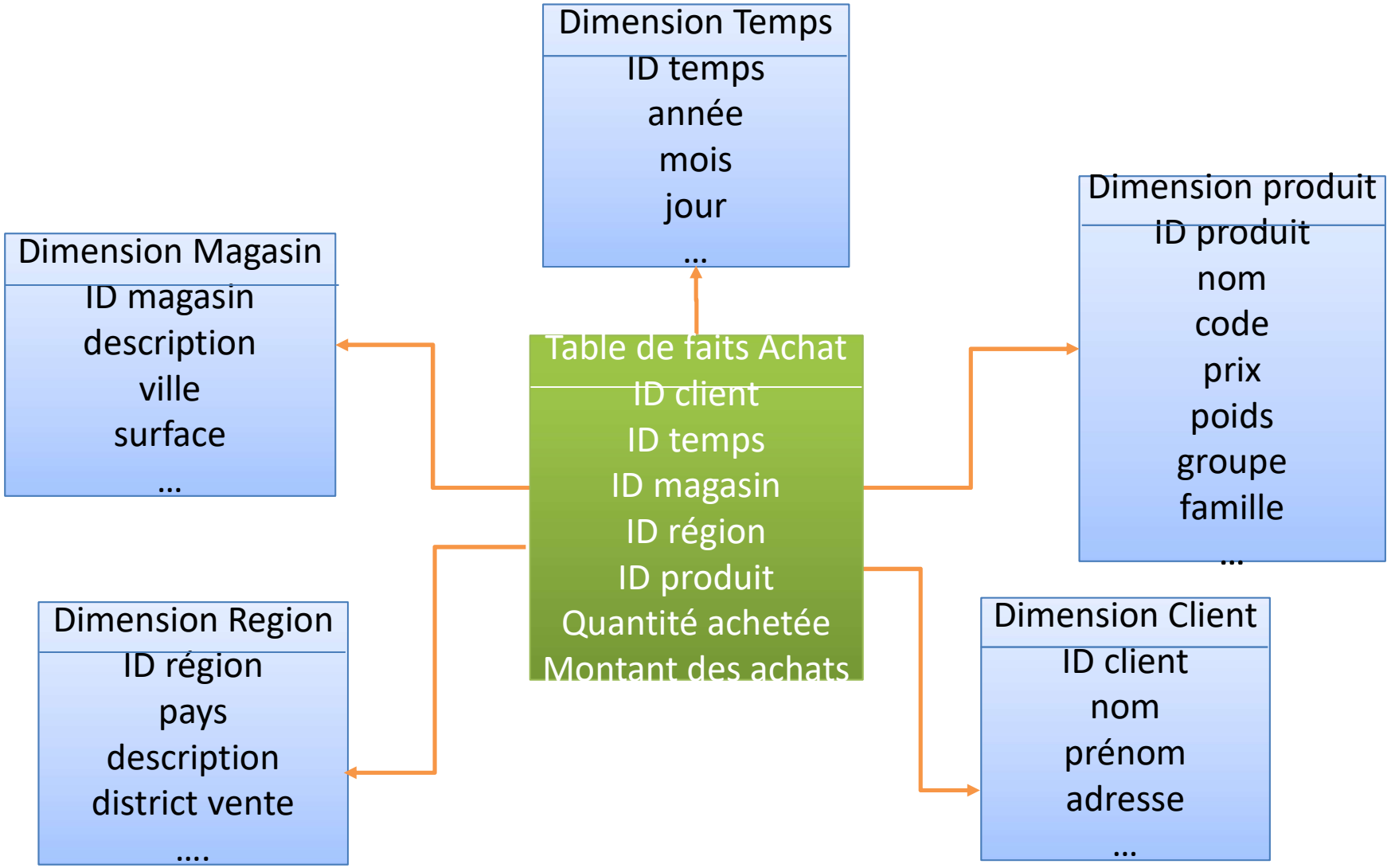


La dimension Date

- Commune à l'ensemble du DW
- Reliée à toute table de faits

Dimension Date
ID Date (CP)
Jour de la semaine
Jour du mois
Mois
Trimestre
Semestre
Année
Num_jour_dans_année
Num_semaine_ds_année

Exemple de modèle en étoile



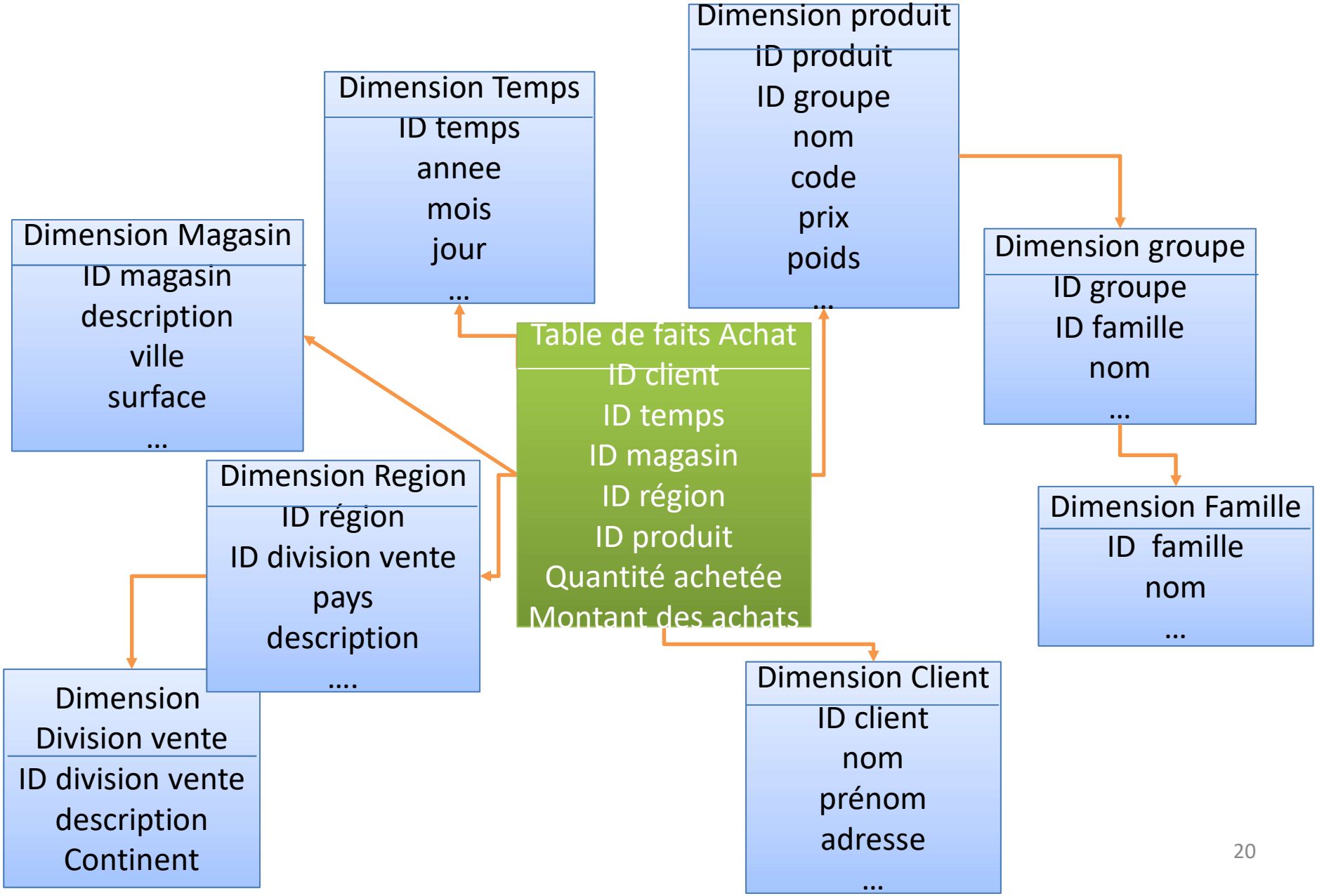
Le modèle en flocon

- Dérivé du modèle en étoile
- Les tables de dimension sont normalisées et les redondances sont éliminées.
- Comparaison étoile/flocon
  - Flocon
    - Le modèle en flocon permet de montrer les hiérarchies entre dimensions
    - La normalisation dans le modèle en flocon permet de réduire la taille des tables.
  - Etoile
    - La dé-normalisation du modèle permet d'améliorer les performances d'exécution des requêtes.
    - Le modèle est plus facile à comprendre par l'utilisateur non informaticien
    - Nombre de jointures limité.

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

Modèle en flocon



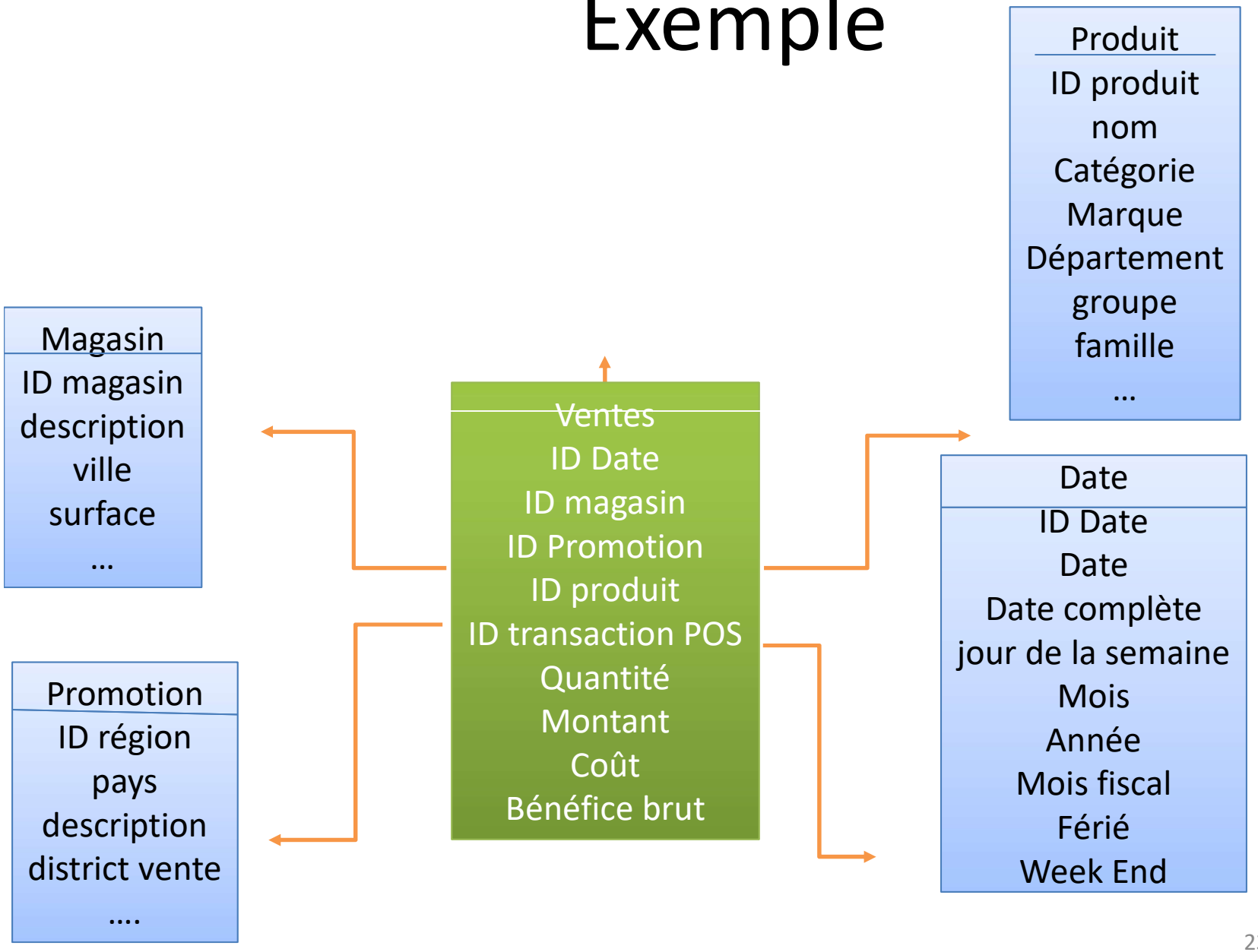
Architecture d'un Datawarehouse

- Deux Approches
  - Approche Inmon: Top-down
  - Approche Kimball: Bottom-Up
- 5 architectures possibles
  - Data marts indépendants: chaque data mart est conçu et alimenté individuellement, pas de données partagées entre les data marts.
  - Bus de data marts: l'approche de kimball avec « conformed dimensions ».
  - Hub and spoke

Etapes de modélisation d'un DW

- Choisir les processus métiers à modéliser :
  - Exemple : le processus "vente" .
- Définir la granularité de chaque processus:
  - Définir ce que représente chaque enregistrement dans la table de faits (exemple : une ligne de ticket de caisse).
- Choisir les dimensions
  - Exemple: date, produit, magasin, promotion
- Identifier les faits numériques:
  - Les faits ayant des granularités différentes doivent appartenir à des tables de faits différentes.

Exemple


Produit
ID produit
nom
Catégorie
Marque
Département
groupe
famille
…
Magasin
ID magasin
description
ville
surface
…
Ventes
ID Date
ID magasin
ID Promotion
ID produit
ID transaction POS
Quantité
Montant
Coût
Bénéfice brut
Date
ID Date
Date
Date complète
jour de la semaine
Mois
Année
Mois fiscal
Férié
Week End
Promotion
ID région
pays
description
district vente
….

Types de dimension

- Dimension dégénérée
- Dimension à évolution lente
- Dimension à évolution rapide

Dimension dégénérée (Degenerate dimension)

- La dimension dégénérée est une clé de dimension dans la table de faits et qui n'est pas associée à une table dimension (exemples: numéro de POS, numéro de commande).

Dimensions à évolution lente

- Les attributs d'une dimension peuvent subir des changements.
  - Un client peut changer d’adresse, avoir des enfants, …
  - Un produit peut changer de noms, de composition(«Raider» en «Twix»);

3 solutions possibles:

- Écrasement de l’ancienne valeur
- Versionnement
- Valeur d’origine / valeur courante.

Dimensions à évolution lente

Solution 1: Écrasement de l'ancienne valeur

- Avantage:
  - Facile à mettre en œuvre
- Inconvénients:
  - Perte de la trace des valeurs antérieures des attributs

Solution 2: Ajout d'un nouvel enregistrement.

- Avantages:
  - Permet de suivre l'évolution des attributs
  - Permet de segmenter la table de faits en fonction de l'historique
- Inconvénient:
  - Accroit le volume de la table

Solution 3: Ajout d'un nouvel attribut

- Avantages:
  - Avoir deux visions simultanées des données :
- Inconvénient:
  - Inadapté pour suivre plusieurs valeurs d'attributs intermédiaires

Dimension à évolution rapide

- Subit des changements très fréquents (tous les mois) dont on veut préserver l'historique
- Solution: isoler les attributs qui changent rapidement et créer une mini-dimension

Dim client
Clé_client
Nom
Prénom
Adresse
…
Revenus
Nb_enfants

Mini Dimension
Clé
Revenus
Nb_enfants

Dictionnaire de données

- C'est un référentiel de métadonnées destiné aux utilisateurs et à l'administrateur du DW
  - Une métadonnée permet de qualifier une données: sémantique, règle de calcul, provenance, qualité…