Entrepôts de données

Mohamed Bouguessa

Plan

1. Introduction
2. Les entrepôts de données
3. Architecture d’un entrepôt de données
4. Modélisation d’un entrepôt de données

Importance de l'information

- Les grands magasins et les entreprises de vente en ligne
  - conservent les achats de leurs clients (reçus de caisse, commandes en ligne);
  - collectent des informations sur leurs clients grâce à
    - des systèmes de cartes de fidélité ou de crédit,
    - et achètent des bases de données géographiques et démographiques.
- Un autre exemple, les différents service web (réseaux sociaux, forums de discussions, blogues) conservent des traces de connexions des utilisateurs

Importance de l'information

- Cela permet le stockage, la manipulation et le transfert de quantités importantes de données.
- À titre d'exemple,
  - la page d'accueil Yahoo! qui reçoit 166 millions de visiteurs par jour et collecte environ 48 Gb de clickstream par heure!
  - AT&T collecte 100 Gb de données réseau par jour.

Extraction de l'information

- Ces mégabases de données, qui ne cessent d'augmenter jour après jour, cachent des informations décisives face au marché et à la concurrence.
- Le besoin d'extraire de l'information pertinente de ces données est alors un enjeu d'actualité.

➢ Extraire de l'information pertinente → Exploitation de l'information.

Exploitation de l'information

« *Les entreprises qui gèrent leurs données comme une ressource stratégique et investissent dans la qualité de celles-ci sont en avance sur leurs compétiteurs, au niveau de la réputation et de profitabilité* »

– Sondage *PricewaterhouseCoopers Global Data Management* (2001)

- Métro / Loblaws / Super C:
  - Entreprises qui vendent de la nourriture **OU;**
  - Entreprises qui exploitent des connaissances sur:
    - Les préférences des clients;
    - Les biais géographiques;
    - La chaîne logistique;
    - Le cycle de vie des produits;
    - Les informations sur les ventes des compétiteurs.

Exploitation de l'information

Le principe *Google:*

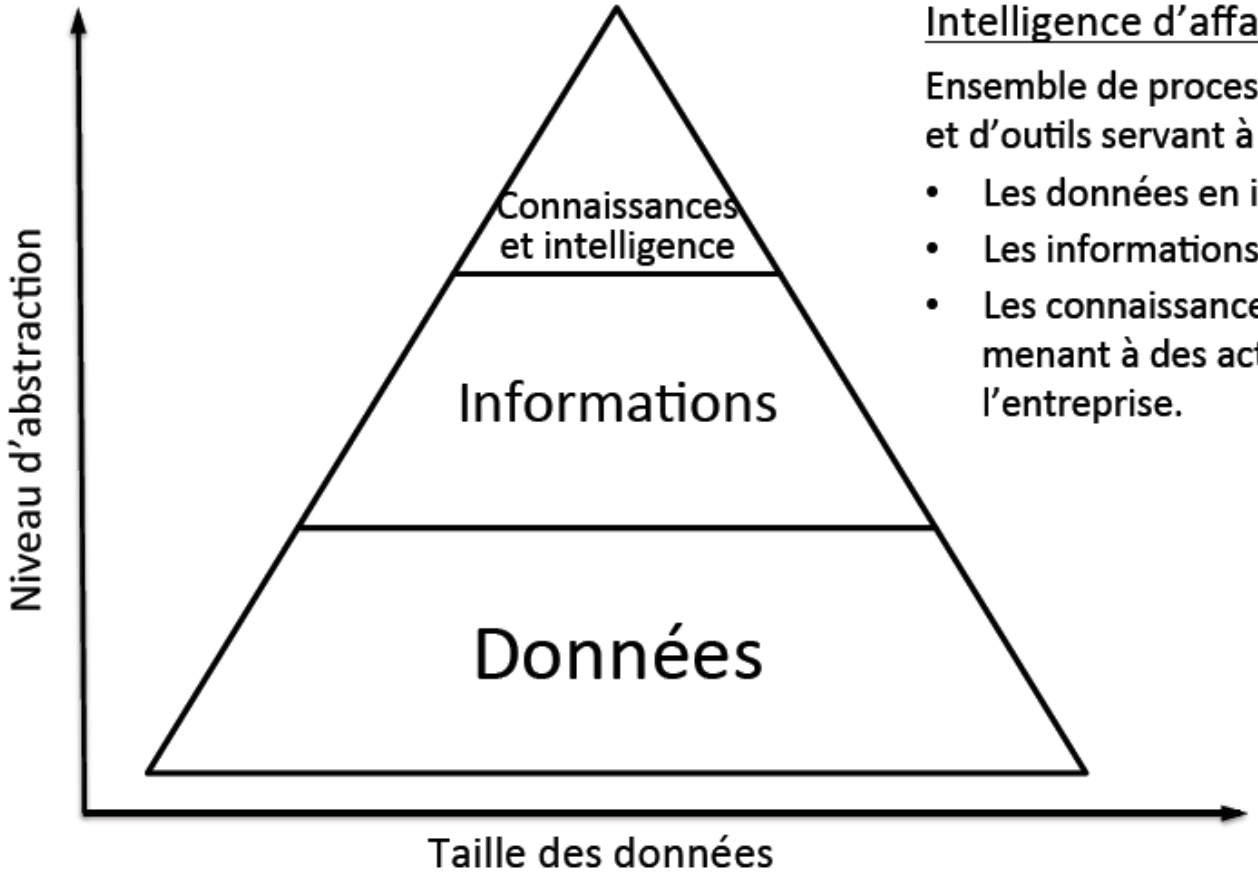
- Toute information a un prix;
- *Google* utilise ses services pour acquérir gratuitement de l'information sur ses usagers:
  - Analyse syntaxique des courriels (*Gmail)*;
  - Profil et liste des contacts (*Google Groups)*;
  - Emploi du temps (*Google Calendar*);
  - *etc.*
- Cette information est utilisée pour envoyer de la publicité **ciblée** aux usagers.

Données →Informations →Connaissances

- Données:
  - Collection d'éléments de valeur brute ou de faits servant à calculer, raisonner et mesurer;
  - Peuvent être collectées, stockées ou traitées;
  - Ne possèdent pas de contexte ni de sens.
- Informations:
  - Proviennent de l'organisation des données, mettant en valeur les relations entre les différents éléments de ces données;
  - Fournissent un contexte et un sens aux données.
- Connaissances:
  - Viennent de la compréhension de l'information dans son contexte;
  - Sont utiles au processus de décision.

Données → Informations → Connaissances

- Pyramide d'abstraction:



Intelligence d'affaires:

Ensemble de processus, de technologies et d'outils servant à transformer:

- Les données en informations;
- Les informations en connaissances;
- Les connaissances en stratégies menant à des actions profitables à l'entreprise.

Exemple d'applications du BI

Application	Exemple de question	Valeur pour l'entreprise
Segmentation des clients	Quels sont les segments du marché et les caractéristiques correspondant à mes clients ?	Personnaliser les relations avec les clients pour avoir un meilleur taux satisfaction et rétention
Tendances d'achat	Quels clients ont le plus de chances de répondre à ma promotion ?	Augmenter la rentabilité des campagnes de promotion en ciblant les bons clients et produits
Rentabilité des clients	Quels profits puis-je obtenir d'un client ?	Prévoir les revenus de l'entreprise et planifier en conséquence
Détection de fraudes	Quelles transactions sont susceptibles d'être frauduleuses ?	Éviter les coûts reliés aux fraudes en éliminant celles-ci
Attrition des clients	Quels clients risquent de quitter ?	Prévenir la perte de clients de valeur important et laisser partir les autres
Choix des canaux	Quels sont les meilleurs canaux pour rejoindre mes clients dans chaque segment du marché ?	Augmenter le nombre de clients et les ventes

Approches du BI

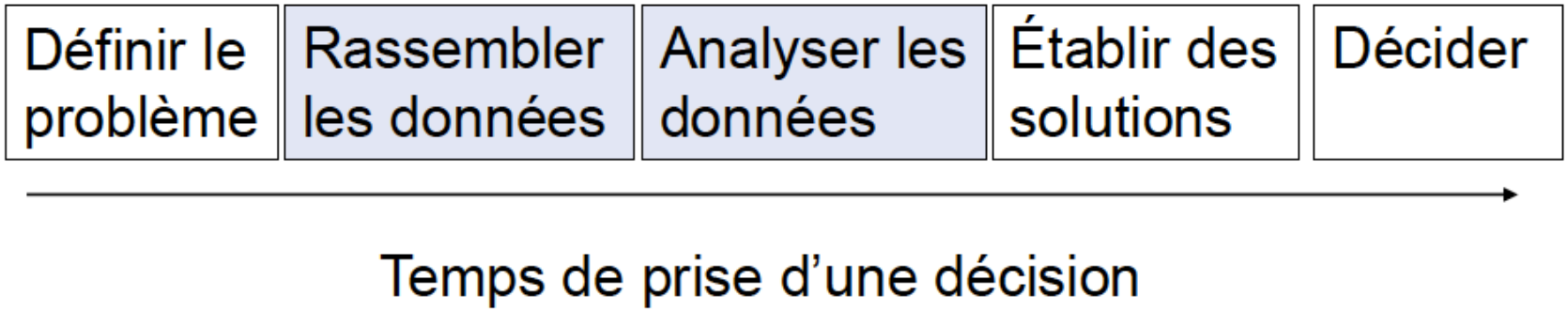
- **Entrepôts de données**
- Forage de données (data mining)
  - Analyse d'association
  - Analyse prédictive
  - Analyse descriptive
- Les requêtes

Plan

1. Introduction
2. **Les entrepôts de données**
3. Architecture d'un entrepôt de données
4. Modélisation d'un entrepôt de données

Contexte

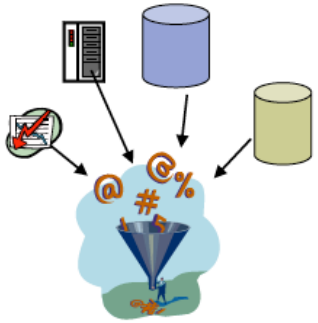
- L'environnement d'affaires est en constante évolution
- Les entreprises doivent répondre rapidement aux changements et innover dans leurs manières d'opérer
- **La prise de décisions** stratégiques et opérationnelles complexes requiert une quantité considérable de données a analyser.



➢ Quelle sont les données utilisables par les décideurs?

Les données utilisables par les décideurs

- Données opérationnelles (de production)
  - Bases de données (Oracle, SQL Server)
  - Fichiers, …
  - Paye, gestion des RH, gestion des commandes…



- Caractéristiques de ces données:
  - Distribuées: systèmes éparpillés
  - Hétérogènes: systèmes et structures de données différents
  - Détaillées: organisation des données selon les processus fonctionnels, données surabondantes pour l'analyse
  - Volatiles: pas d'historisation systématique

Problématique

- Comment répondre aux demandes des décideurs?
  - En donnant un accès rapide et simple à l'information stratégique
  - En donnant du sens aux données

➢ Mettre en place un outil pour répondre aux exigences d'un système capable de supporter la prise de décision, et l'intégration de données provenant de sources multiples.

  ➢ Entrepôt de données (Data Warehouse - DW)

Définition d'un entrepôt de données

« Un entrepôt de données est une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le processus de décision. »

-- Bill Inmon

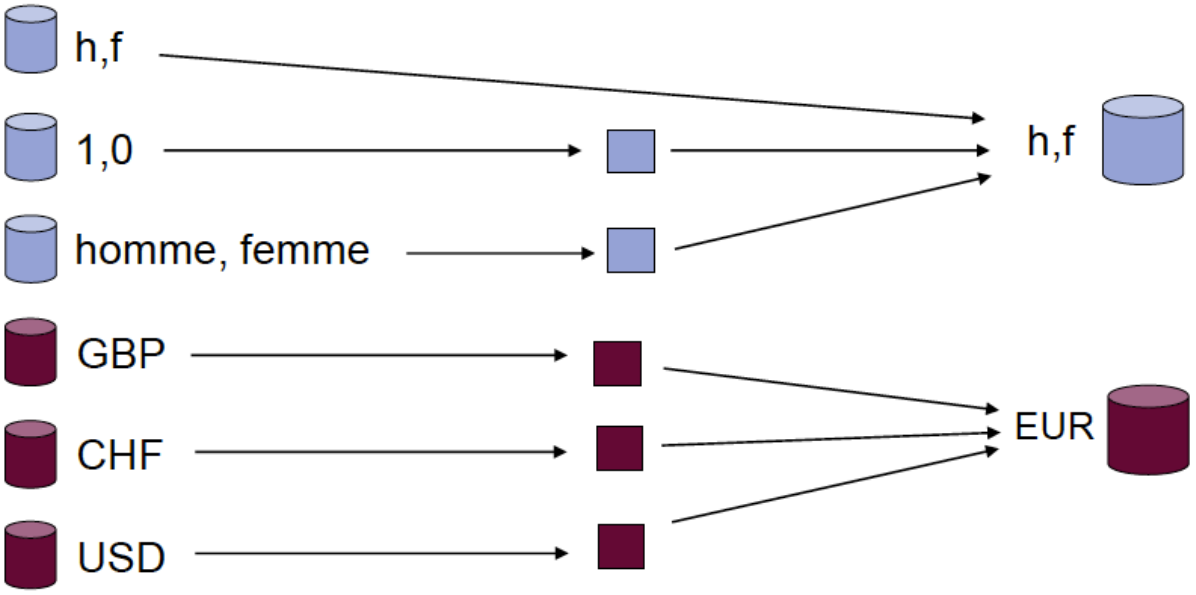
Orientées sujet

- Un entrepôt de données rassemble et organise des données associées aux différentes structures fonctionnelles de l'entreprise, pertinentes pour un sujet ou thème et nécessaire aux besoins d'analyse

➢ Les données sont donc organisées par sujet (ex: clients, produits, ventes, etc.)

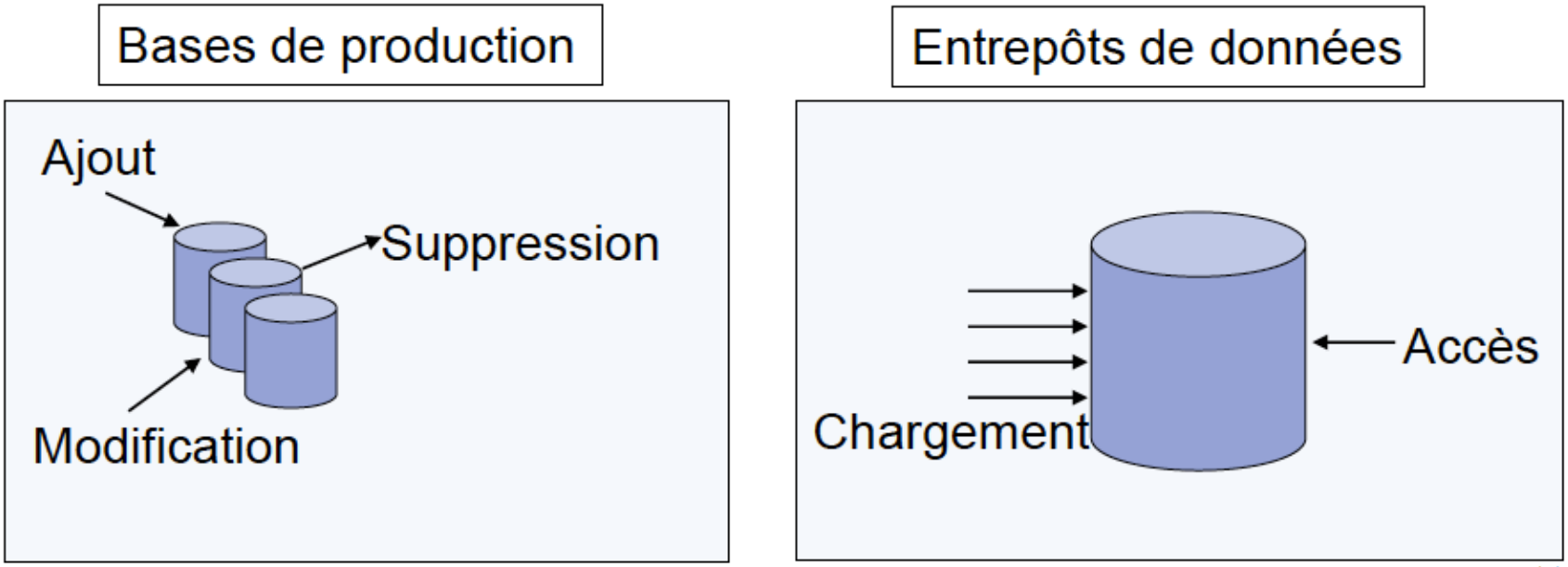
Intégrées

- Les données, qui proviennent de diverses sources hétérogènes, sont consolidées et intégrées dans l'entrepôt.
  - Normalisation des données
  - Définition d'un référentiel unique



Non volatiles

- Une fois insérées dans l'entrepôt, les données ne sont jamais modifiées ou effacées; elle sont conservées pour des analyses futures.



Historiées

- Données datées
  - Les données persistent dans le temps
  - Mise en place d'un référentiel temps

Base de production

Image de la base en Mai 2005

Répertoire

Nom	Ville
Dupont	Paris
Durand	Lyon

Image de la base en Juillet 2006

Répertoire

Nom	Ville
Dupont	Marseille
Durand	Lyon

Entrepôt de données

Calendrier

Code	Année	Mois
1	2005	Mai
2	2006	Juillet

Répertoire

Code	Année	Mois
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

De l’entrepôt à la décision

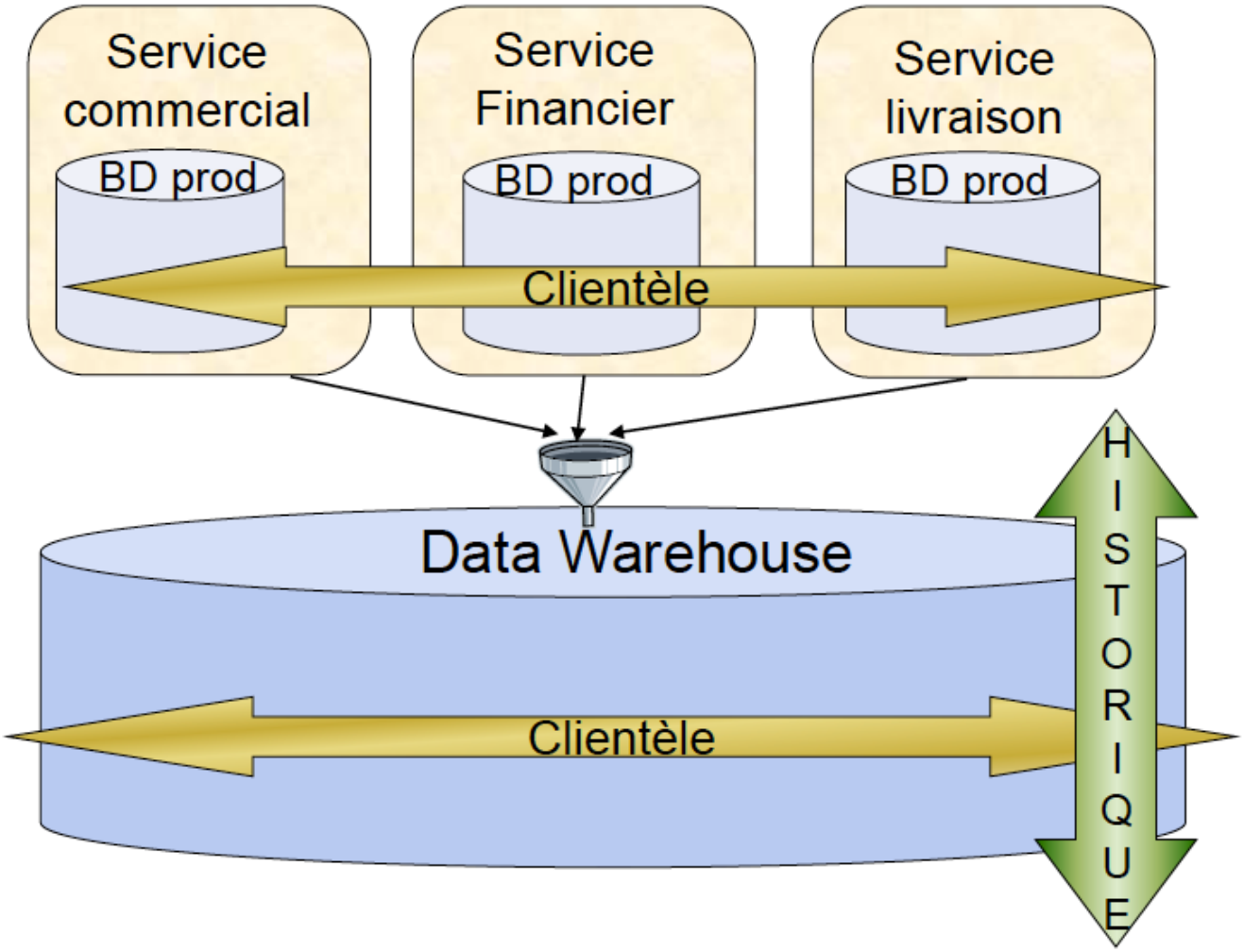
- **Entreposage des données** : avant d’être chargées dans l'entrepôt, les données sélectionnées doivent être :
  - extraites des sources (internes : BD opérationnelles, externes : BD et fichiers notamment issus du Web)
  - soigneusement épurées afin d'éliminer des erreurs et réconcilier les différentes sémantiques associées aux sources)

De l'entrepôt à la décision

- **Exploitation des données de DW** : systèmes décisionnels
  - A partir des données d'un DW diverses analyses peuvent être faites, notamment par des techniques « On-Line Analitical processing » (OLAP) ou de forage de données (data mining).
  - Notons que les informations et connaissances obtenues par exploitation de DW ont un impact direct sur les bénéfices de l'entreprise (augmentation des ventes par un marketing plus ciblé, amélioration de la rotation des stocks, …)

SGBD et DW

OLTP: On-Line Transactional Processing



OLAP: On-Line Analitical Processing

OLTP vs OLAP

	OLTP	**OLAP**
utilisateurs	employé	décideur
fonction	Operations journalières	Aide à la décision
Conception de la BD	orientée application	Orientée sujet
données	courante, à mettre à jour detaillée, relationelle isolée	historique, résumée, multidimensionnelle integrée, consolidée
usage	répété	ad-hoc
accès	Lecture écriture Index sur clé primaire	Lecture seule Différentes analyses
unité de travail	transaction simple	Requête complexe
# enr. utilsés	dizaines	millions
#users	milliers	centaines
Taille de la BD	100MB-GB	100GB-TB

Intérêts des l'entrepôts de données

- Intégration des différents bases
  - Rassembler des données hétérogènes
  - Les homogénéiser et les restructurer
- Fournissent une vue consolidée des données de l'entreprise
- Données non volatiles (pas de suppression)
- Historisation
- Organisation vers prise de décision
- Procurent un avantage concurrentiel à l'entreprise

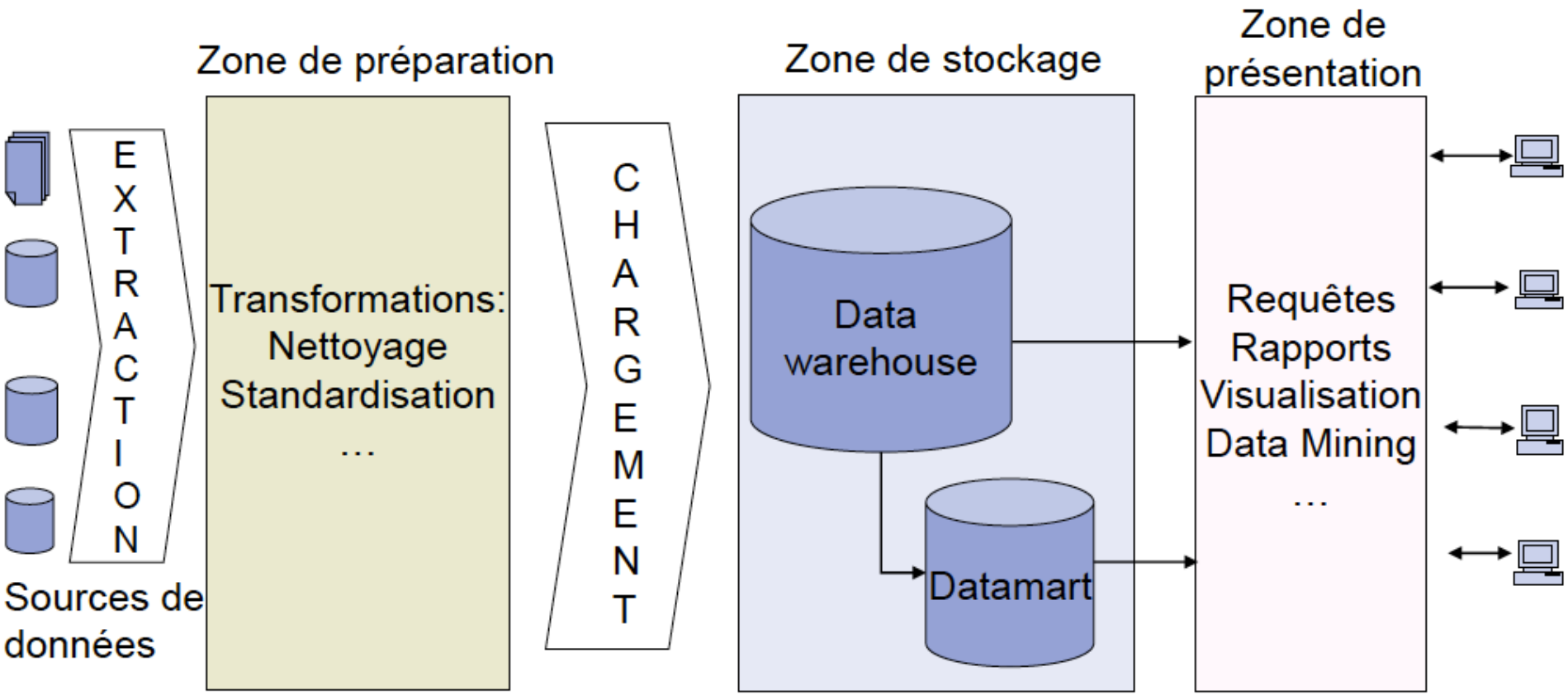
Intérêts des entrepôts de données

- Intégration des différents bases
  - Rassembler des données hétérogènes
  - Les homogénéiser et les restructurer
- Fournissent une vue consolidée des données de l'entreprise
- Données non volatiles (pas de suppression)
- Historisation
- Organisation vers prise de décision
- Procurent un avantage concurrentiel à l'entreprise

Plan

1. Introduction
2. Les entrepôts de données
3. **Architecture d'un entrepôt de données**
4. Modélisation d'un entrepôt de données

Architecture générale



Source de données

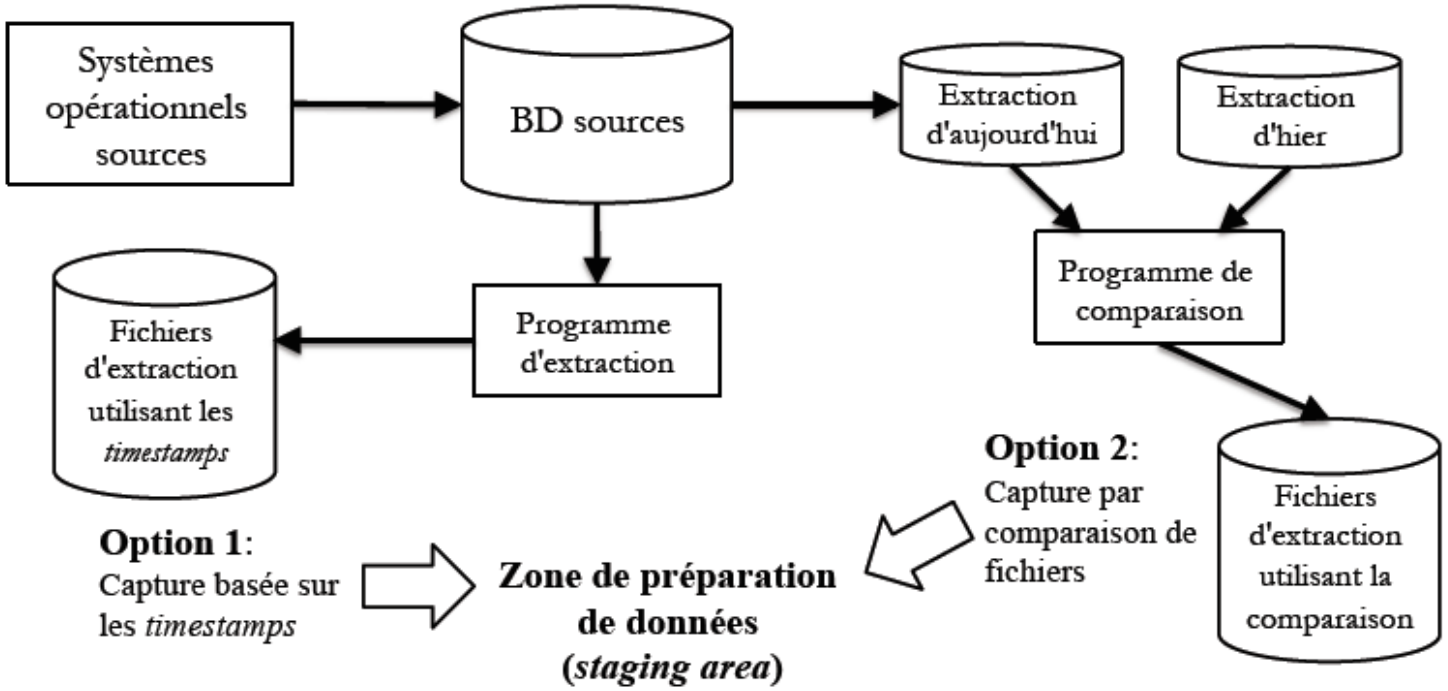
- **Enterprise resource planning (ERP) :**
  - Gèrent les processus opérationnels d'une entreprise (ex: ressources humaines, finances, distribution, approvisionnement, etc.).
- **Customer relationship management (CRM) :**
  - Gèrent les interactions d'une entreprise avec ses clients (ex: marketing, ventes, après-vente, assistance technique, etc.).
- **Point of sale (POS) :**
  - Matériels et logiciels utilisés dans les caisses de sorties d'un magasin.
- **Externes:**
  - Ex: données concurrentielles achetées, données démographiques.

Intégration des données (ETL)

- **Processus Extract, Tranform, Load (ETL):**
  1. Extraire les données des sources hétérogènes:
     - Extraction différée;
     - Extraction temps-réel.
  2. Transformer/consolider les données:
     - Données redondantes / manquantes;
     - Différents noms / types;
     - Incohérences.
     - ➢ But : Rendre cohérentes les données des différentes sources.
  3. Charger les données intégrées dans l'entrepôt:

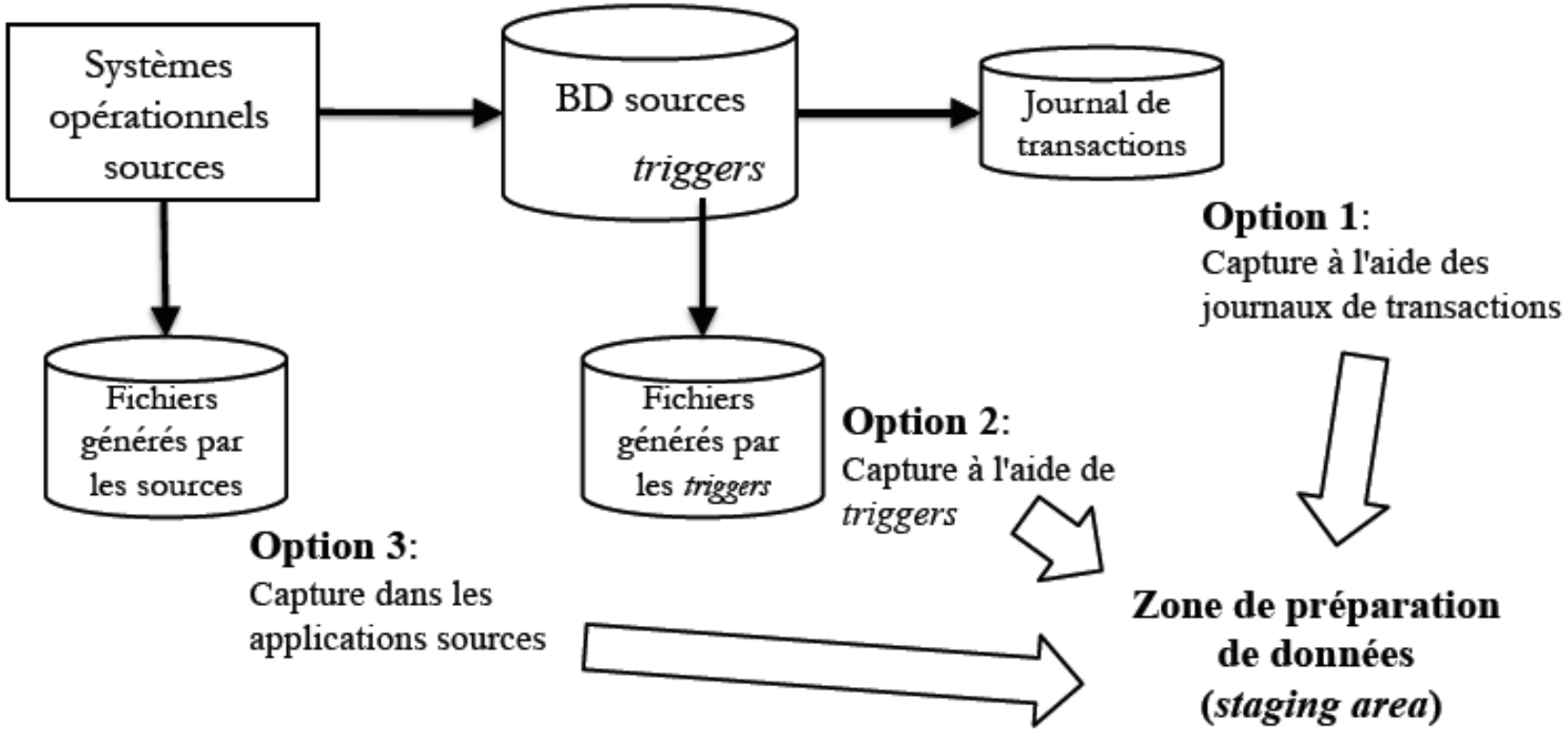
Extraction des données

- **Extraction différée:**
  - Extrait tous les changements survenus durant une période donnée (ex: heure, jour, semaine, mois).



Extraction des données

- **Extraction en temps-réel:**
  - S'effectue au moment où les transactions surviennent dans les systèmes sources.



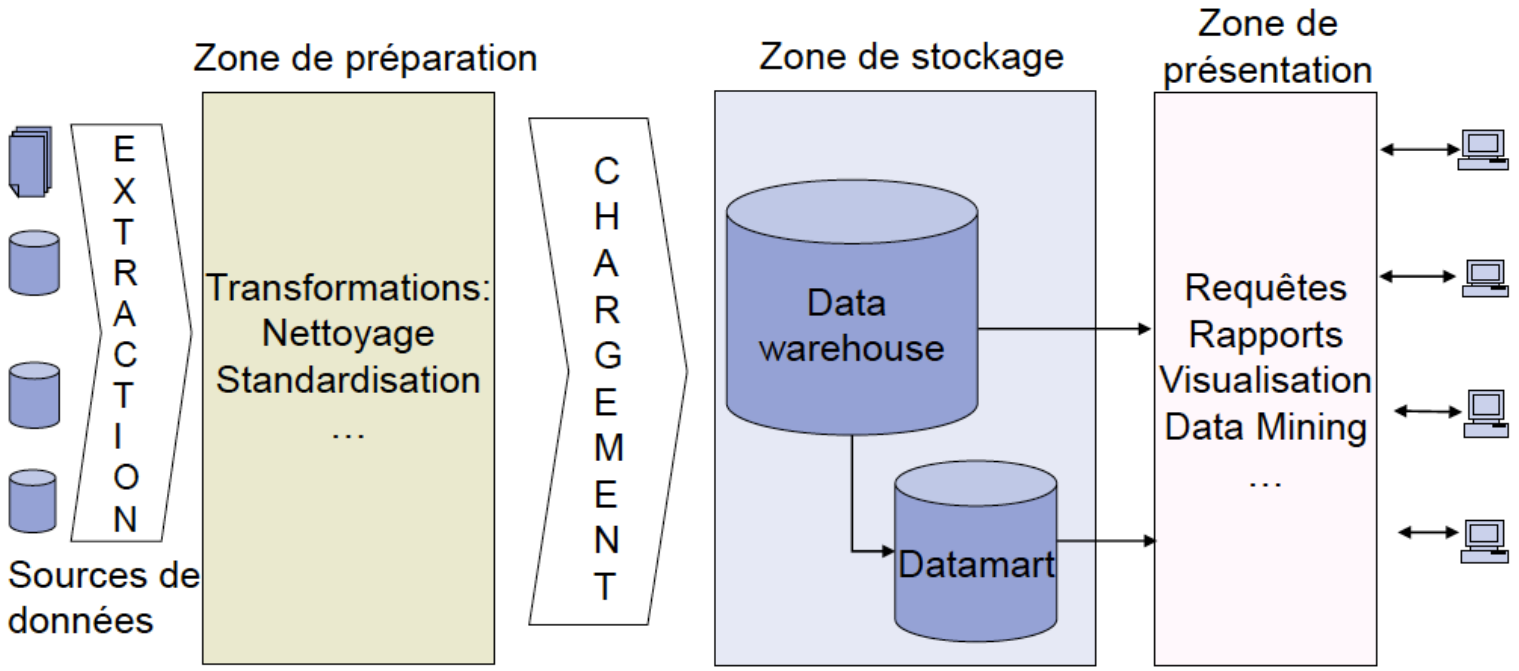
Transformation des données

- **Types de transformation**:
  - **Révision de format**. Ex: Changer le type ou la longueur de champs individuels.
  - **Décodage de champs**. Ex: ['homme', 'femme'] vs ['M', 'F'] vs [1,2].
  - **Découpage de champs complexes**. Ex: extraire les valeurs prénom, secondPrénom et nomFamille à partir d'une seule chaîne de caractères nomComplet.
  - **Pré-calcul des agrégations**. Ex: ventes par produit par semaine par région.

Zone de stockage

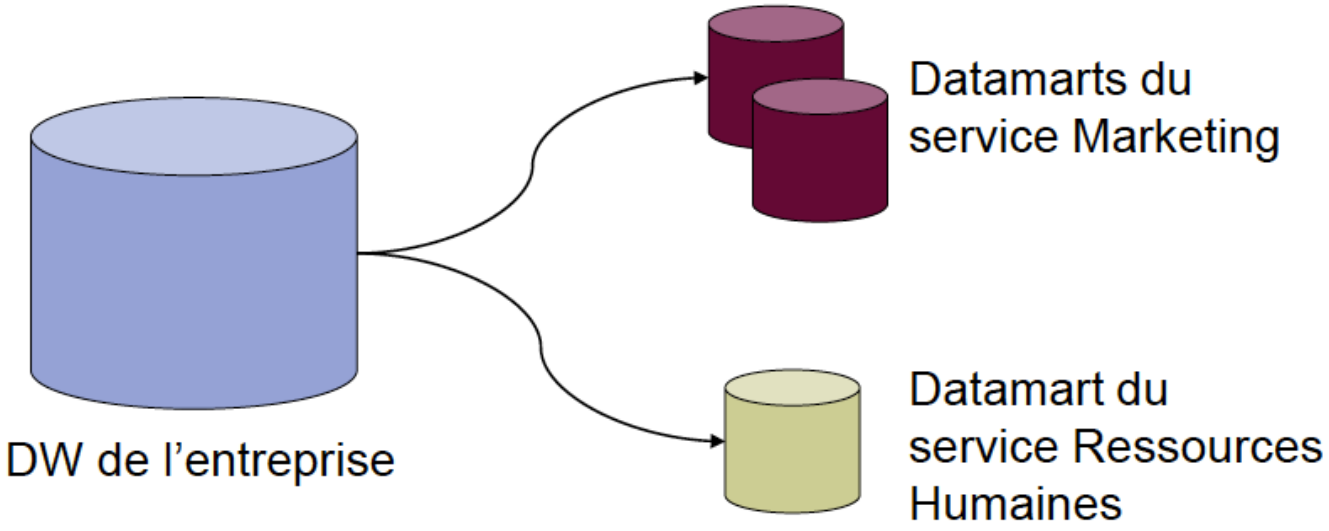
- On y transfère les données nettoyées
- Stockage permanent des données

➢ Data mart?



Data marts ou les magasins des données

- 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

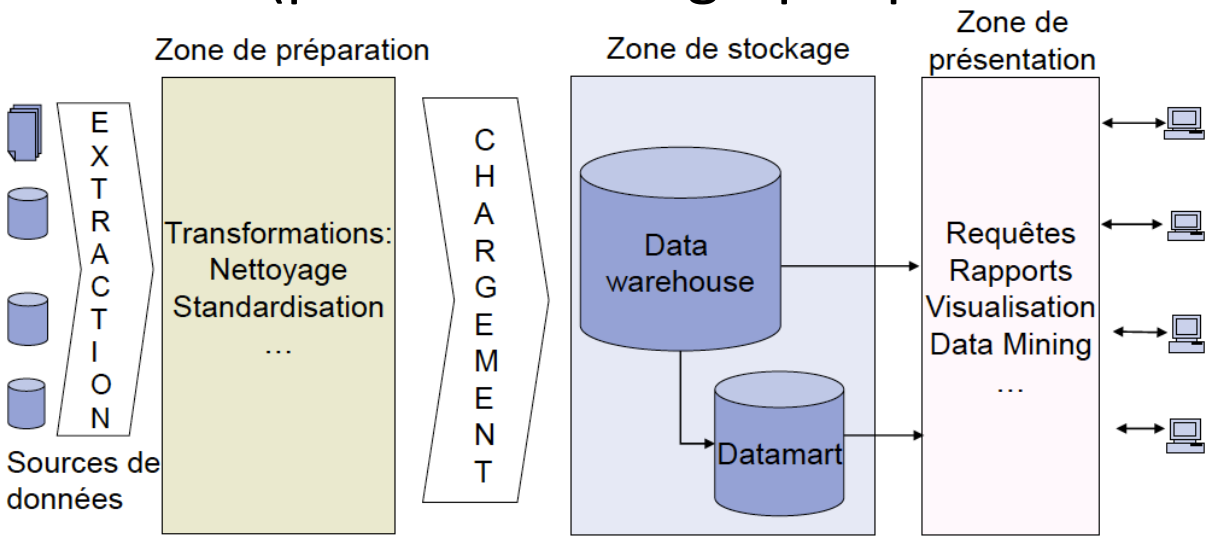


Intérêts des Data marts

- Nouvel environnement structuré et formaté en fonction des besoins d'un métier ou d'un usage particulier
- Moins de données
  - Plus facile à comprendre, à manipuler
  - Amélioration des temps de réponse : fourni un accès rapide aux données les plus souvent analysées.
- Utilisateurs plus ciblés:
  - Data mart plus facile à définir

Zone de présentation

- Donne accès aux données contenues dans le DW pour l'analyse et l'exploration des données entreposées :
  - Formulation de requêtes afin de trouver des faits à étudier;
  - L'analyse de tendance (courbes d'évolution);
  - Découverte de connaissance (data mining)
  - Visualisation (présentations graphiques variées).



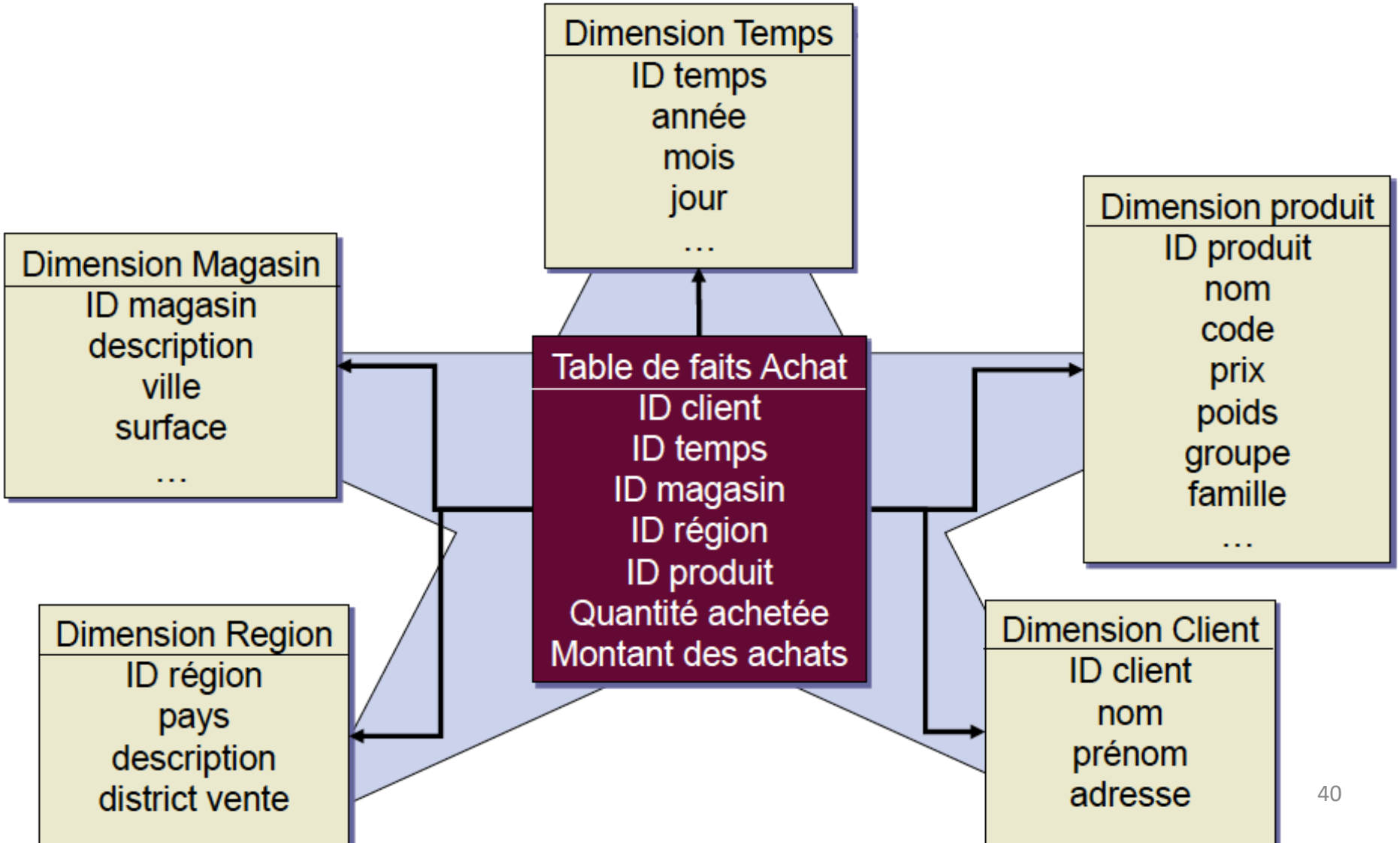
Plan

1. Introduction
2. Les entrepôts de données
3. Architecture d'un entrepôt de données
4. **Modélisation d'un entrepôt de données**

La modélisation dimensionnelle

- **Définition** :
  - Technique de conception logique permettant de structurer les données de manière à les rendre intuitives aux utilisateur d'affaires et offrir une bonne performance aux requêtes.
- **Caractéristiques :**
  - Divise les données en **faits** et **dimensions**;
  - **Les faits** (mesures) sont généralement des valeurs numériques provenant des processus d'affaires;
  - **Les dimensions** fournissent le contexte (qui, quoi, quand, où, pourquoi et comment) des faits;
- **Représentation** :
  - Schéma en étoile: une table de faits entourée de plusieurs tables de dimension .

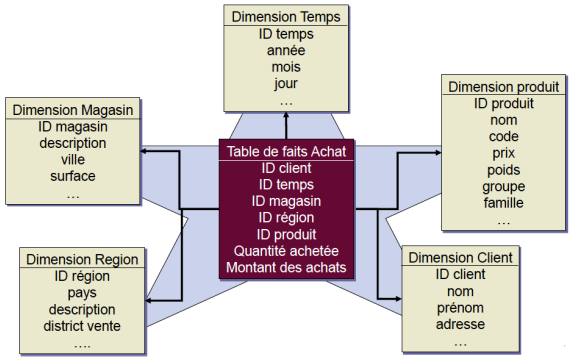
Exemple de schéma en étoile



La modélisation dimensionnelle

- Modèle entité-relation (ER):
  - Représente les données sous la forme d'entités (tables) et de relations
- Modèle dimensionnel :
  - Représente les données comme des faits et des dimensions;
- Avantages du modèle dimensionnel:
  - Compréhensibilité:
    - Les données sont regroupées selon des catégories d'affaires qui ont un sens pour les utilisateurs d'affaires;
  - Performance:
    - Évite les jointures coûteuses;

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)

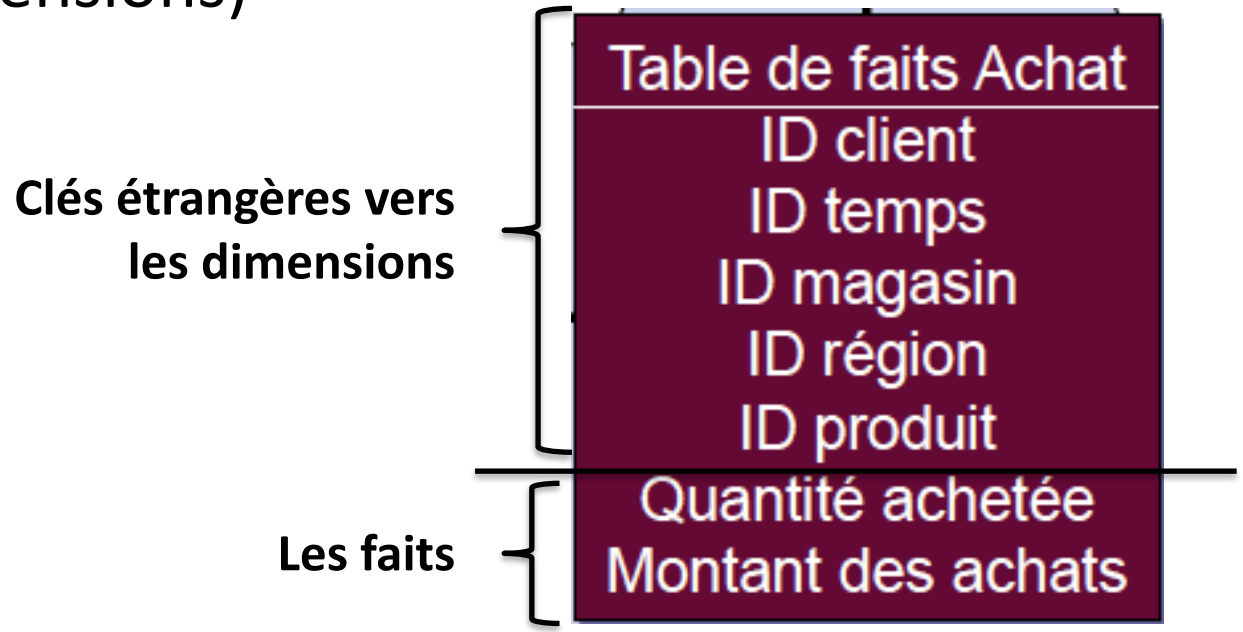
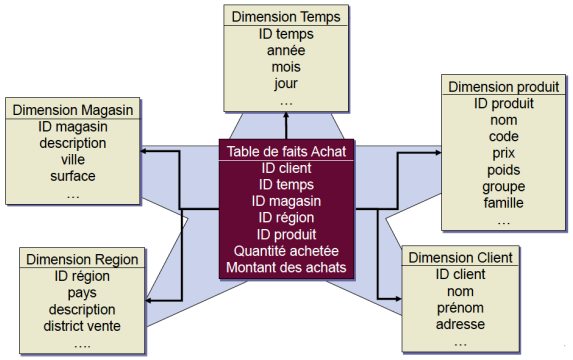


Table de faits



- Contient
  - Des clés étrangères vers les tables de dimension:
    - Ex: ID du client qui fait la commande, ID du produit commandé, etc.
  - Ce que l'on souhaite mesurer :
    - Quantités achetée, montant des achats…

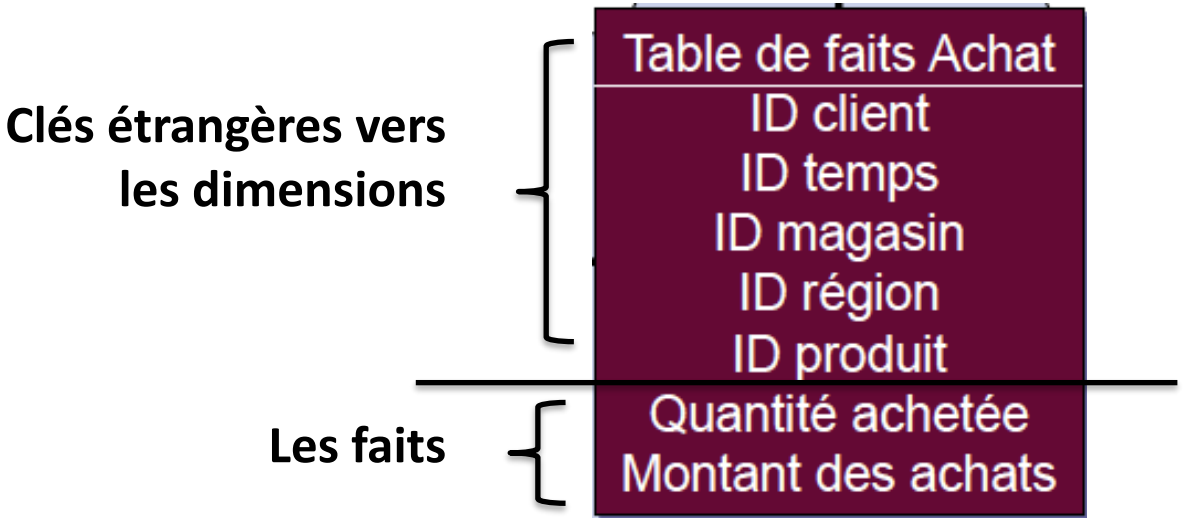
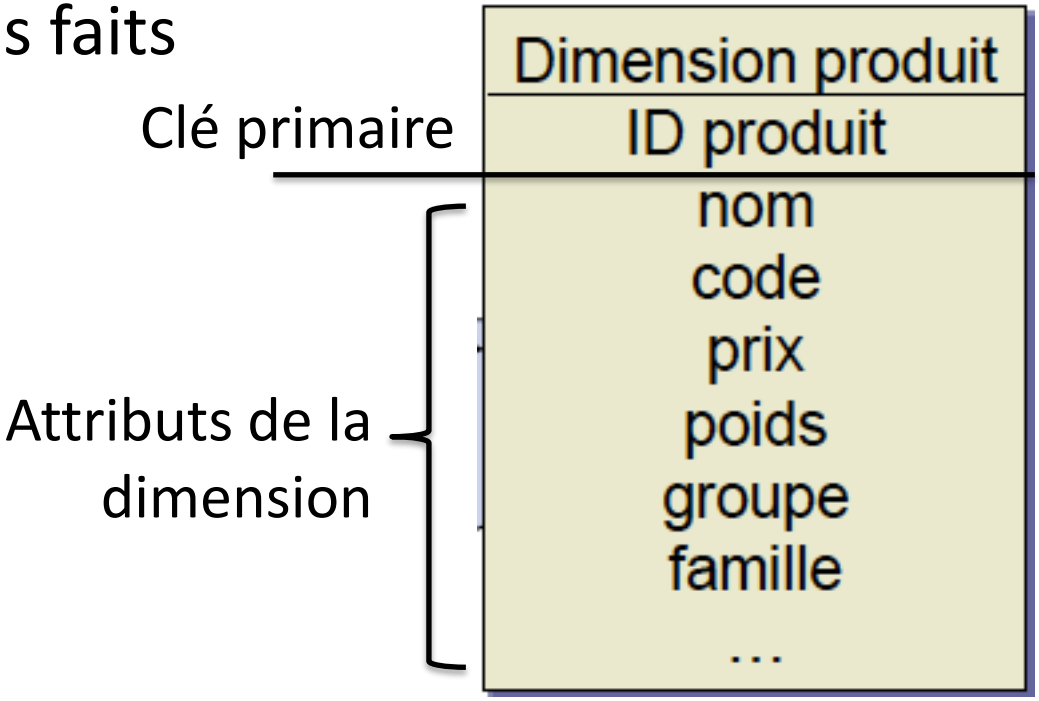


Table de dimension

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



- Choix des dimensions:
  - Demande le jugement et l'intuition du modélisateur.

Table de dimension - caractéristiques

- Consistent en un ensemble d'attributs regroupés selon les objets clés d'une entreprise. Ex: produits, clients, installations, etc.
- Propriétés des attributs:
  - Descriptif (ex: chaînes de caractères);
  - De qualité (ex: aucune valeur manquante ou erronée);
  - Valeurs discrètes (ex: jour, âge d'un client);
- La puissance analytique de l'entrepôt est proportionnelle à la richesse et la qualité des attributs dimensionnels.

La dimension temps

- Commune à l'ensemble de l'entrepôt
- Reliée à la table de faits

- Définition des attributs: Pré-générer toutes les valeurs pour un certain historique (ex: 10 ans) pour faciliter la référence et éviter les mises à jour

Granularité d'une dimension

- Une dimension contient des membres organisés en hiérarchie : Un ensemble d'attributs ayant une relation hiérarchique (x est inclus dans y).
- Exemple :
  - **Temps** : année → mois → semaine → jour → heure
  - **Produit** : famille → catégorie → marque → nom du produit
  - **Lieu** : pays → province → région → ville → code postale

Évolution des dimensions

- Dimensions à évolution lente
- Dimensions à évolution rapide

Dimensions à évolution lente

- Un produit peut changer de noms ou de formulation:
  - « yogourt à la vanille » en « yogourt saveur vanille »
- Gestion de la situation, 3 solutions:
  - Écrasement de l'ancienne valeur avec la nouvelle
  - Ajout d'un nouvel enregistrement
  - Ajout d'un nouvel attribut

Écrasement de l'ancienne valeur

Clé produit	Description du produit	Groupe de produits
12345	Intelli-Kids	~~Logiciel~~ Jeux éducatifs

- **Avantage**:
  - Facile à mettre en ouvre
- **Inconvénients**:
  - Perte de la trace des valeurs antérieures des attributs : Impossible de faire des analyses sur l'ancienne valeur
  - À utiliser seulement lorsque l'ancienne valeur n'est pas significative pour les besoins d'affaires;
  - Exige de mettre à jour les données agrégés avec l'ancienne valeur.

Ajout d'un nouvel enregistrement

Clé produit	Description du produit	Groupe de produits
12345	Intelli-Kids	Logiciel
25963	Intelli-Kids	Jeux éducatifs

- Avantages:
  - Permet de suivre l'évolution des attributs faire des analyses historiques
- Inconvénient:
  - Accroit le volume de la table

Ajout d’un nouvel attribut

Clé produit	Description du produit	Groupe de produits	Nouveau groupe de produits
12345	Intelli-Kids	Logiciel	Jeux éducatifs

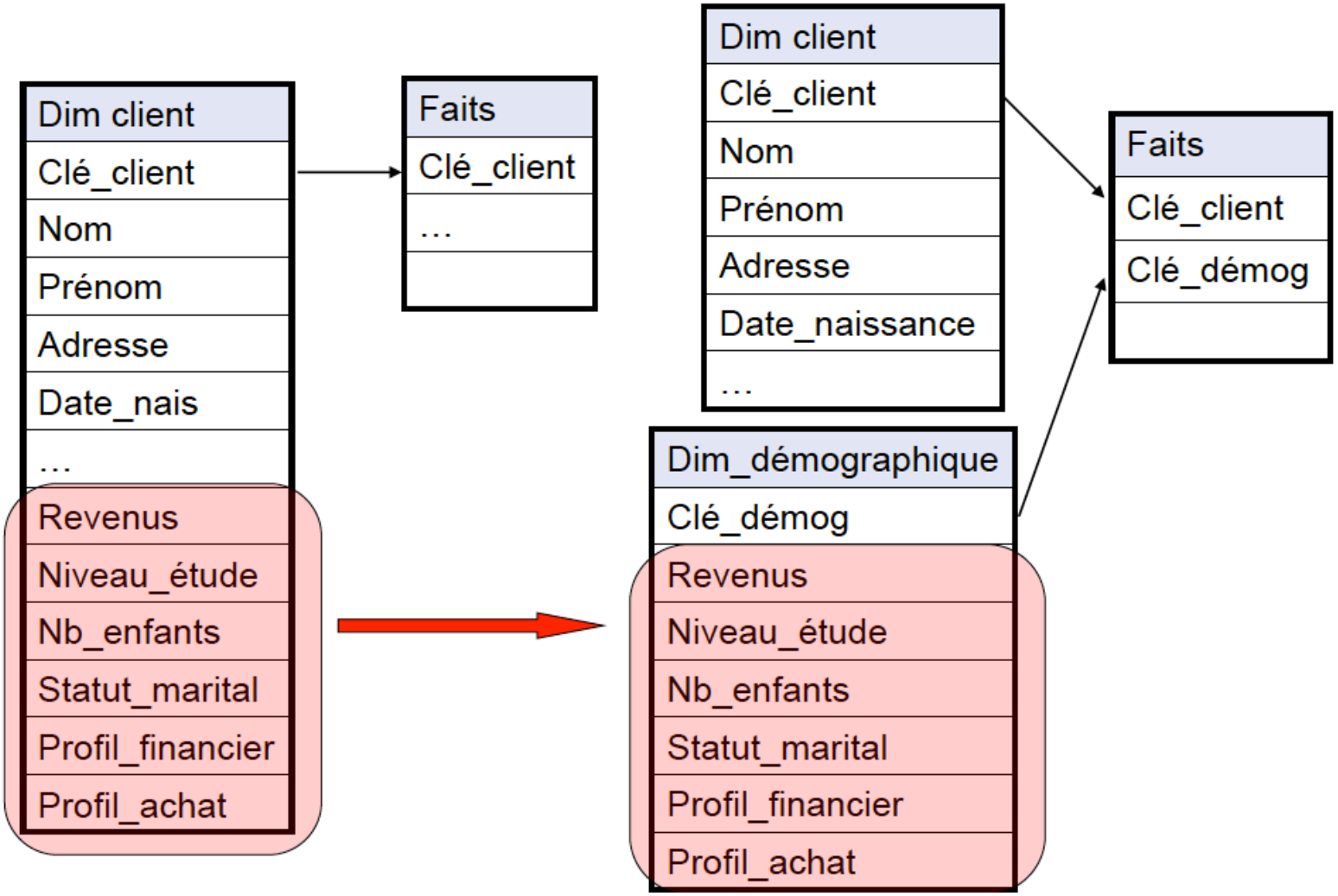
- Avantages:
  - Voir les données (enregistrements) comme si le changement n’avait pas eu lieu.
- Inconvénient:
  - Explosion des nombre d’attributs

➢ À utiliser lorsque la valeur change peu souvent et il n'est pas nécessaire d'avoir une historique profonde. Ex: changement annuel et besoin d'une profondeur de deux ans.

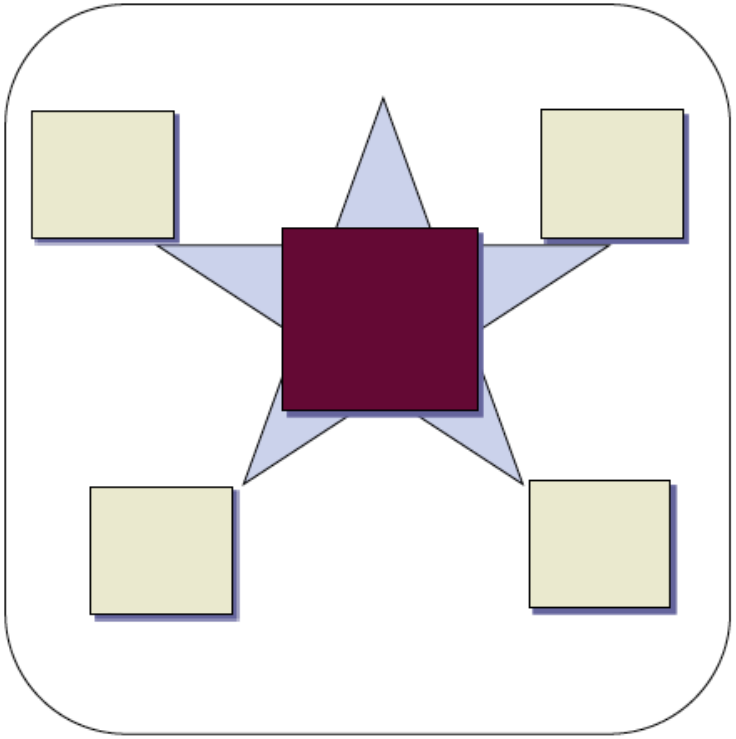
Évolution des dimensions

- Dimensions à évolution lente
- **Dimensions à évolution rapide:**
  - Subit des changements fréquents dont on veut préserver l'historique
  - Solution: isoler les attributs qui changent rapidement

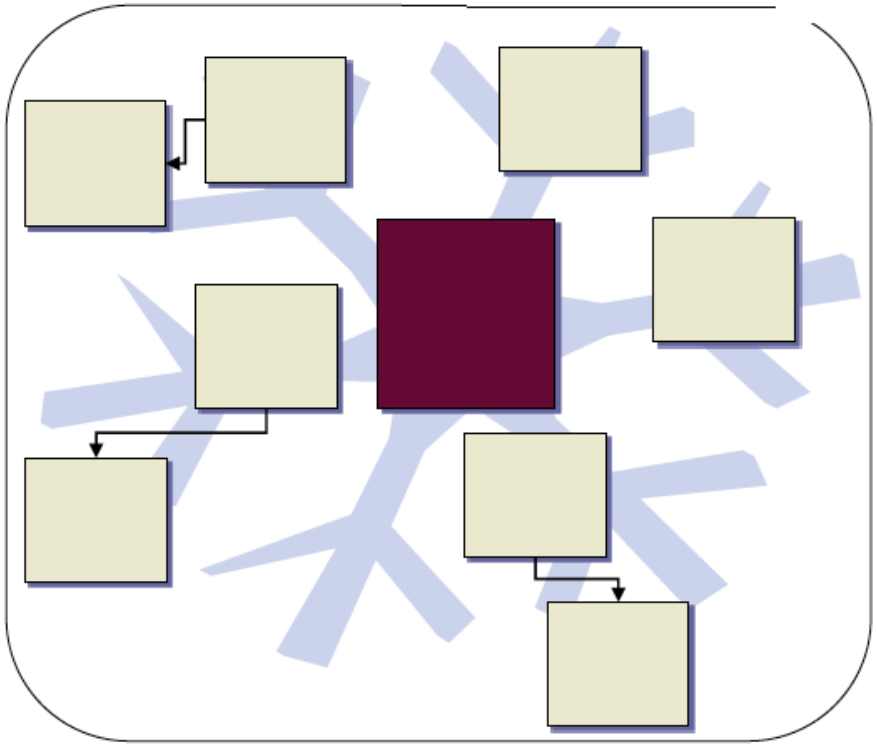
Dimensions à évolution rapide



Les types de modèles



Modèle en étoile

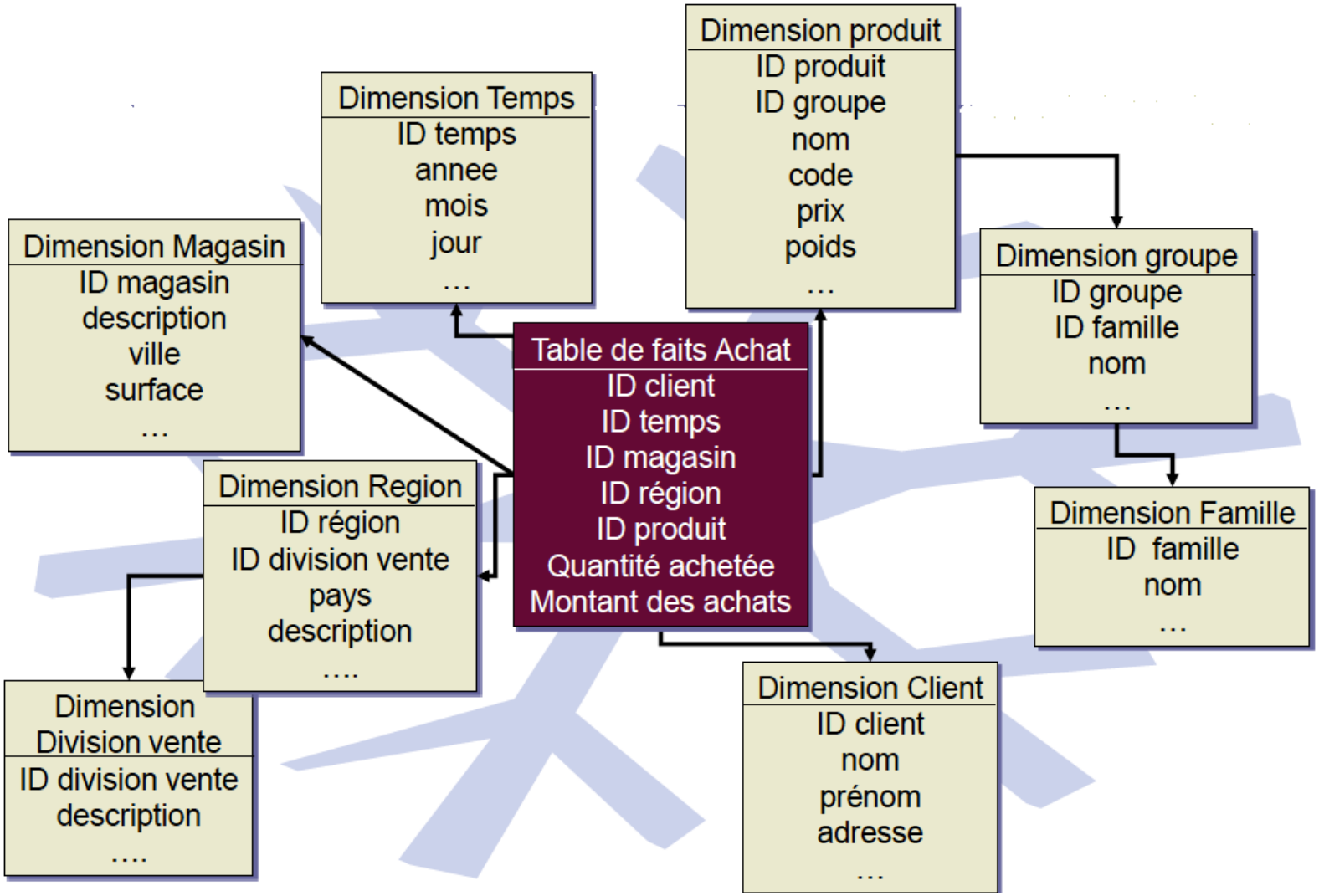


Modèle en flocon

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

Modèle en flocon – Exemple



Modèle en flocon

- Avantages:
  - Plus facile de mettre à jour les dimensions en cas de changement.
- Désavantages:
  - Modèle complexe : schéma moins intuitif aux utilisateurs d'affaires.
  - Dégradation de la performance à cause des jointures additionnelles.

Outils d'analyse

- OLAP (prochain cours)
- Forage de données (data mining)
- Requêtes

Domaines d'application

➢ Domaine bancaire

- Pour une banque, il est important de pouvoir regrouper les informations relatives à un client afin de répondre à ses demandes de crédit par exemple.
- Des mailing ciblés doivent aussi être rapidement élaborés à partir de toutes les informations disponibles sur un client lors de la Commercialisation d'un nouveau produit
- L'utilisation de cartes de crédit nécessite des contrôles *a posteriori*, par exemple pour la recherche de fraudes : la mémorisation des mouvements peut rendre de grands services
- Les échanges d'actions et de conseils de courtages sont facilités par une mémorisation de l'histoire et une exploitation par des outils décisionnels avancés par exemple pour déterminer des tendances de marchés

Domaines d'application

- **Domaine de la grande distribution**

- intéressant de regrouper les informations de ventes pour déterminer les produits à succès, mieux suivre les modes, détecter les habitudes d'achats, les préférences des clients par secteur géographique
- La fouille de données (data mining) a permis de développer des techniques sophistiquées d'exploitation de données qui aident à mettre en évidence les règles de consommation
- Explorer le panier de la ménagère est devenu un exercice d'école : il s'agit de trouver à partir de l'enregistrement des transactions quelles sont les habitudes d'achats, plus précisément quels sont les produits achetés en même temps.

- **Apports constatés dans la grande distribution :**

- Augmentation des ventes grâce à un meilleur marketing
- Amélioration des taux de rotation de stocks
- Élimination des produits obsolètes
- Meilleure négociation des achats

Plan

1. Introduction
2. Les entrepôts de données
3. Architecture d'un entrepôt de données
4. Modélisation d'un entrepôt de données