# Master EISIS

## Expertise et Ingénierie des Systèmes d'Information en Santé

Unité d'Enseignement :

IME-EDAD

Entrepôts de données et aide à la décision

Thème :

# Constitution des entrepôts de données

Auteur :

Dr Jean-Charles DUFOUR

jean-charles.dufour@univ-amu.fr

Document mis à jour le 04 mars 2014

# Préambule

Ce diaporama s'inspire :

- du polycop « Découverte de connaissances à partir de données » (R.Gilleron et M. Tommasi, Lille).
- du tutorial T17 animé par J.H. Holmes lors de l'AMIA 2007 (Chicago)

Les autres références utilisées sont mentionnées individuellement dans les diapositives.

# Plan du cours

- Introduction / rappels
- Les étapes de la constitution d'un entrepôt
  - Etude préalable
  - Modélisation des données
  - Alimentation (ou chargement)
  - Publication et mise à disposition

# Introduction / rappels

- Informatique de production:
  - Traitement d'opération individuelles pouvant impliquer différents métiers
  - Pas (ou peu) de compilation, d'historisation, de synthèse
- Informatique décisionnelle
  - Analyse par métiers / sujet
  - Suivi dans le temps d'indicateurs calculés et agrégés

➔Les objectifs sont différents, les schémas de données sont différents

# Introduction / rappels

- Informatique de production :
  - Les bases de données sous-jacente sont complexes est donc difficilement appréhendables par tout utilisateur
  - Le système de production ne doit pas être interrompu/mobilisé pour effectuer des traitements
  - Les données opérationnelles ne sont pas toutes disponibles au sein du même système de production
  - Les données opérationnelles sont rarement dans un format propice à leur analyse

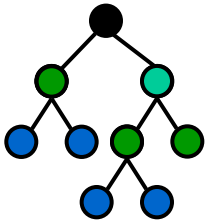
➔Il est donc nécessaire d'avoir d'autres systèmes orientés vers la décision

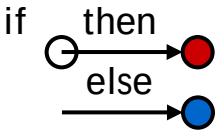
# Introduction / rappels

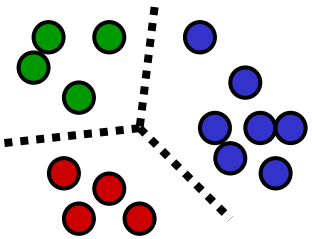
- Un entrepôts de données (*datawarehouse*) est un système orienté vers la décision
  - Historisation des données
  - Restructuration des données de production
  - Récupération d'informations tierces (démographiques, géographique, sociologiques,…)
- Un contexte propice à leur avènement
  - Progrès technologiques (stockage, temps de traitement, temps d'accès, technique de parallélisme,..)
  - Maturités des techniques issues de l'IA et de l'analyse statistique → fouille de données (*Data Mining*)

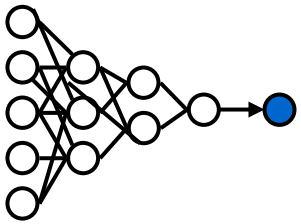
# Méthodes de fouille de données

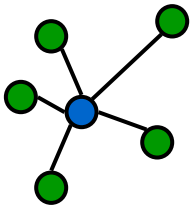
- Il en existe plusieurs, la première difficulté est de choisir celle adapté à la situation et aux objectifs de l'analyse :
  - Arbres de décision
  - Règles d'association
  - Algorithme de segmentation
  - Méthode des plus proches voisins
  - Réseaux de neurones
  - ...









# Les étapes de la constitution d'un entrepôt de données

1. Etude préalable :
   - étude des besoins
   - estimation des coûts
   - estimation des bénéfices attendus
2. Modélisation des données
3. Alimentation (ou chargement)
4. Publication et mise à disposition

# Entrepôts de données : généralités

- Ils accueillent un flot important de données
- Sont alimentés par les données des serveurs de production
- Leur architecture est optimisé pour l'extraction d'informations en un temps minimal
- Nécessite des matériels spécifiques

# Entrepôts de données : généralités

- Entrepôt des données ≠ Base de données

| Caractéristique | Base de données | Entrepôt de données |
|---|---|---|
| Opération | gestion courante, production | analyse, support à la décision |
| Modèle de données | entité/relation | étoile, flocon de neige |
| Normalisation | fréquente | plus rare |
| Données | actuelles, brutes | historisées, parfois agrégées |
| Mise à jour | immédiate, temps réel | souvent différée |
| Niveau de consolidation | faible | élevé |
| Perception | bidimensionnelle | multidimensionnelle |
| Opérations | lectures, mises à jour, suppressions | lectures, analyses croisées, rafraîchissements |
| Taille | en gigaoctets | en téraoctets |

Emprunté à fr.wikipedia.org le 20/11/2006

# Entrepôts de données : généralités

- Les besoins du décisionnel
  - Variété des questions
    - prévisibles → tableau de bord
    - imprévisible → outil de requêtage utilisable par l'utilisateur (nécessite : la simplicité du modèle des données et la performance malgré les grands volumes)
  - Modélisation simple, adaptée aux notions directement perçues par l'utilisateur et qu'il souhaite explorer
    - « Faits » (ex : ventes, communications, diagnostics,…)
    - « Dimensions » (ex : temps, région, symptômes,…)

# Entrepôts de données : généralités

- Structure logique doit tenir compte de la nécessité d'optimiser les temps de réponse. Pour cela, la **redondance** dans les informations est souvent utilisées !!
- Objectif est d'assurer la **cohérence globale** des données
- Alimentations **planifiées**
- Transferts préalablement **contrôlés**
- Les informations **ne sont pas modifiées** après leur introduction

# Constitution d'un entrepôt de données :
# 1. étude préalable

## Étude des besoins – définition des objectifs

- Sujet ?
- Population cible ? (ex : tous les malades, seulement les malades curables, …)
- Entité statistique étudiée ? (ex : la personne, un foyer, un quartier,…)

➔ Groupe projet :

Maitrise d'ouvrage
(directions, experts métiers, futurs utilisateurs,…)
+
Maitrise d'œuvre
(statisticiens, informaticiens)

# Constitution d'un entrepôt de données : 1. étude préalable

## Étude des besoins – définition des objectifs

- Résultats attendus par les utilisateurs ?
- Type de requêtes qu'ils formuleront ?
- Projets qui ont été définis ?

➔ Types d'analyses ?

- Rétrospectives
- Prédictives

➔ Recenser les données nécessaires

# Constitution d'un entrepôt de données : 1. étude préalable

## Etude des besoins :

- Déterminer les unités des « dimensions » et la granularité des « faits »
- Envisager le découpage de l'entrepôt en plusieurs parties (*data marts*)

# Constitution d'un entrepôt de données : 1. étude préalable

## Coûts de déploiement

- Matériel :
  - Machine(s) puissante(s) dédié(s)
  - Capacité de stockage très importante
- Logiciel :
  - Logiciels d'administration
  - Logiciels d'interrogation et de visualisation
  - Logiciels de data mining
- Logistique
  - Équipe pour maintenir l'entrepôt (administrateurs, développeurs, concepteurs,…)
  - Prévoir la formation des utilisateurs

# Constitution d'un entrepôt de données : 1. étude préalable

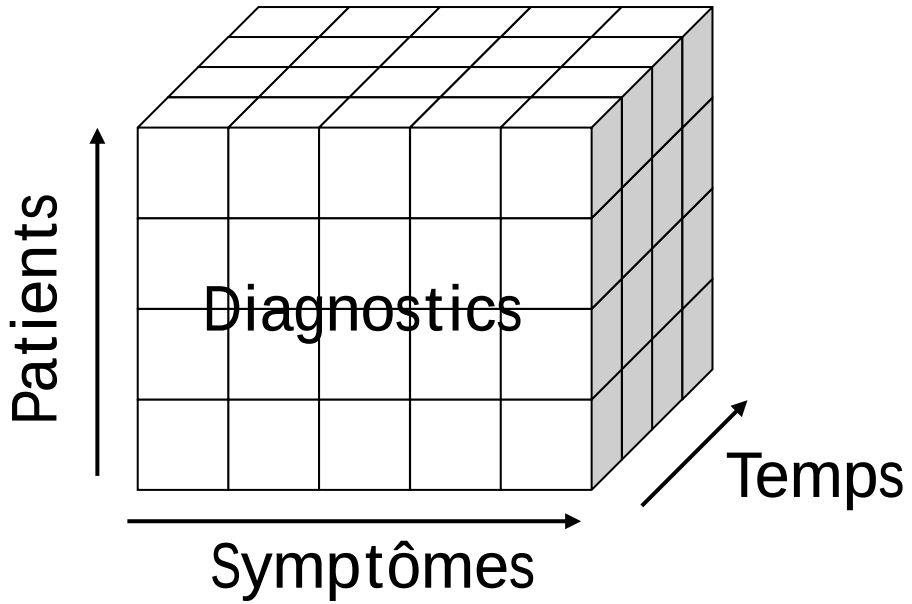
## Bénéfices attendus

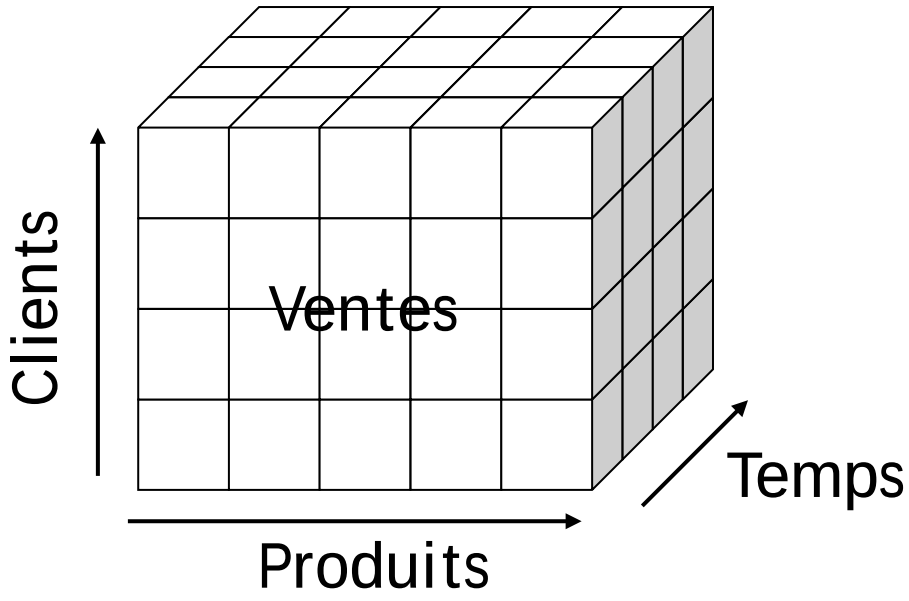
- Gains apportés à l'entreprise par le système ? :
  - Impact de la découverte de nouvelles connaissances
  - Impact sur l'organisation
  - Impact sur les processus de l'entreprise
  - Etc..

➔ Faire un bilan de retour sur investissement (ROI)

# Constitution d'un entrepôt de données :
# 2. Modélisation des données

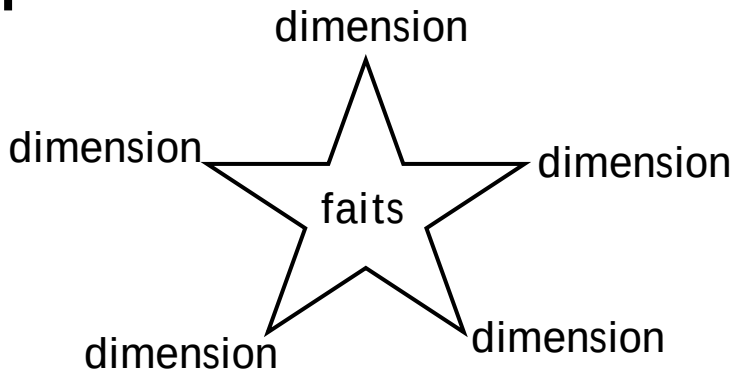
- Modèle conceptuel :
  - Simple : appréhendable par les utilisateurs
  - Multidimensionnel (cube 3D, 4D, ...)
    - « faits »
    - « dimensions »

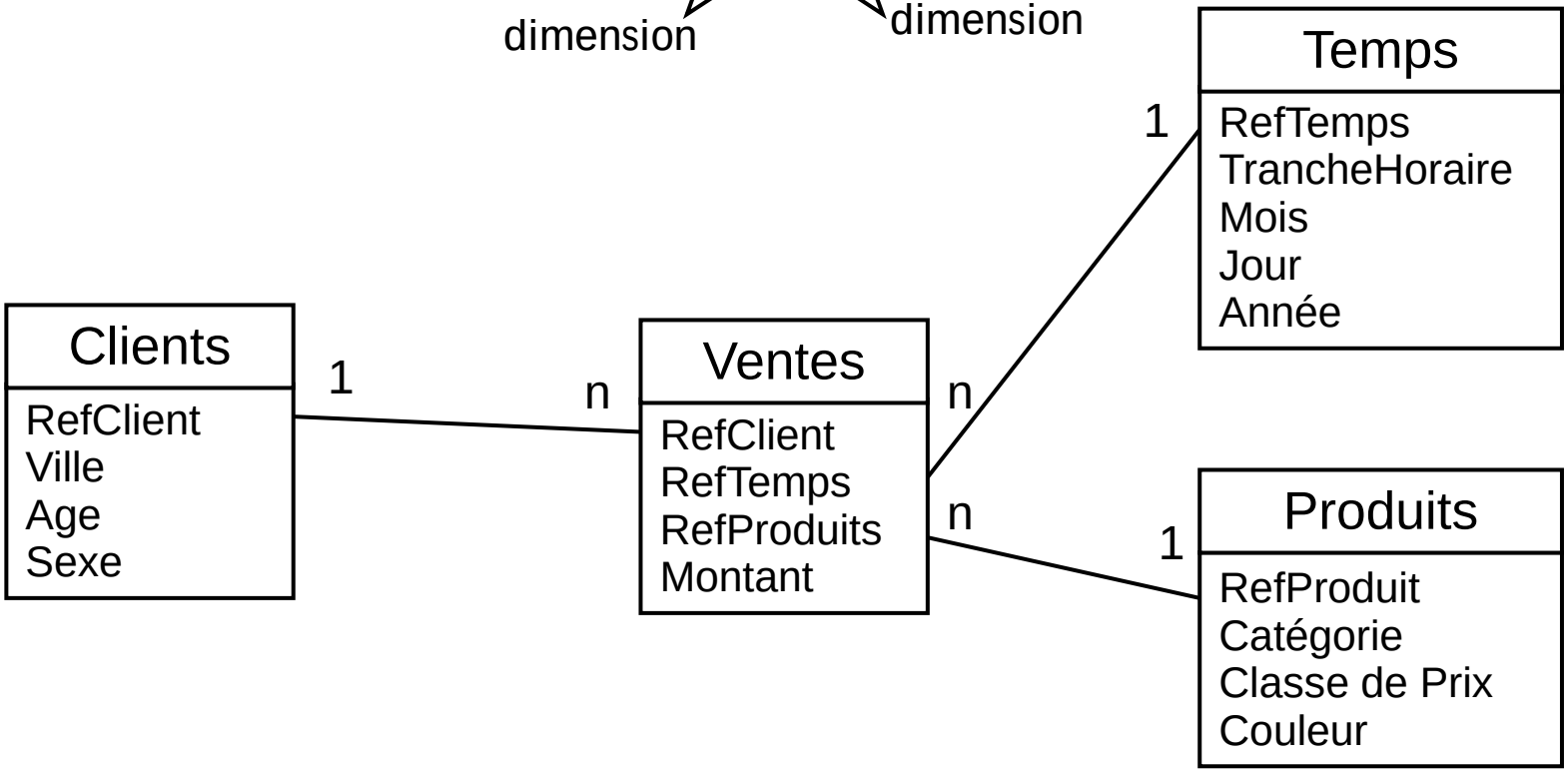


# Constitution d'un entrepôt de données :
# 2. Modélisation des données

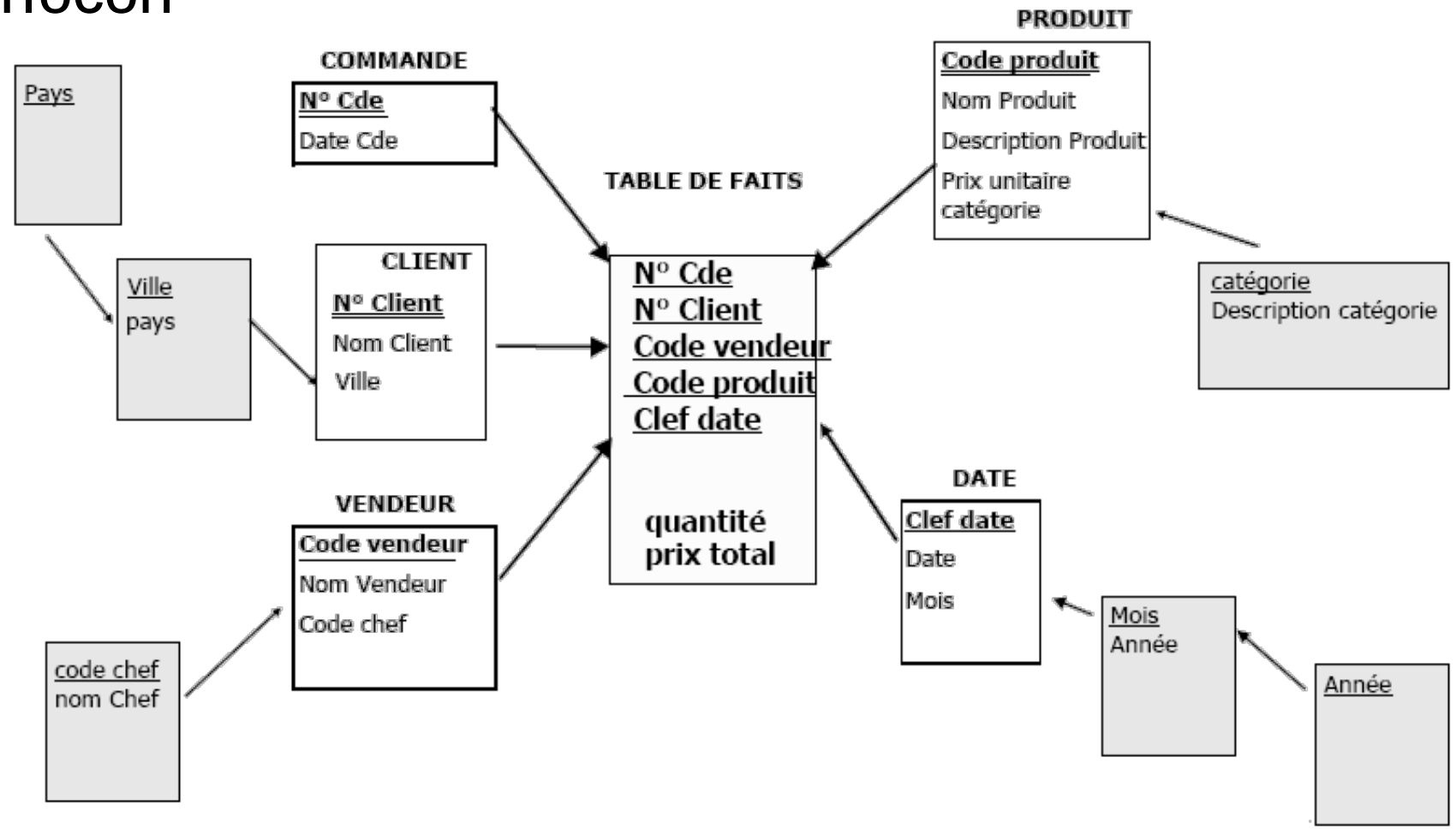
- Modèle logique :
  - en étoile



# Constitution d'un entrepôt de données :
# 2. Modélisation des données

- Modèle logique :
  - en flocon




Emprunté à C. Vangenot, LBD, EPFL

# Constitution d'un entrepôt de données : 3. Alimentation

- Transfert des données du/des système(s) opérationnel vers l'entrepôts
- L'adaptation des données est nécessaire
- Ajout de faits ++
- Ajout de dimensions +/-
- Chargement → impact physique sur l'entrepôt et son utilisation (recalcul des clés, mise à jour des index,…)

# Constitution d'un entrepôt de données : 3. Alimentation

- Transformation et vérification sont nécessaire :
  - Formatage
    - Format physique des données
    - Transformation de type, de nom
  - Consolidation
  - Uniformisation d'échelle
  - Autres

# Constitution d'un entrepôt de données : 3. Alimentation

- Nettoyage des données
  - Vérifier les erreurs de saisies dans les bases de production (ex : doublons passés inaperçus)
  - Contrôle des domaines des valeurs et des valeurs douteuses
  - Gestion des information manquantes
  - Gestion des données incohérentes :
    - Valeurs hors-limites
      *(ex: t° corporelle de 67°C, Pression Artérielle de 600mmHg)*
    - Données contradictoires
      (*ex: date naissance différente selon la source, pathologie féminine chez un homme, non-fumeur consommant 20 cigarettes/jours, etc.)*
  - ...

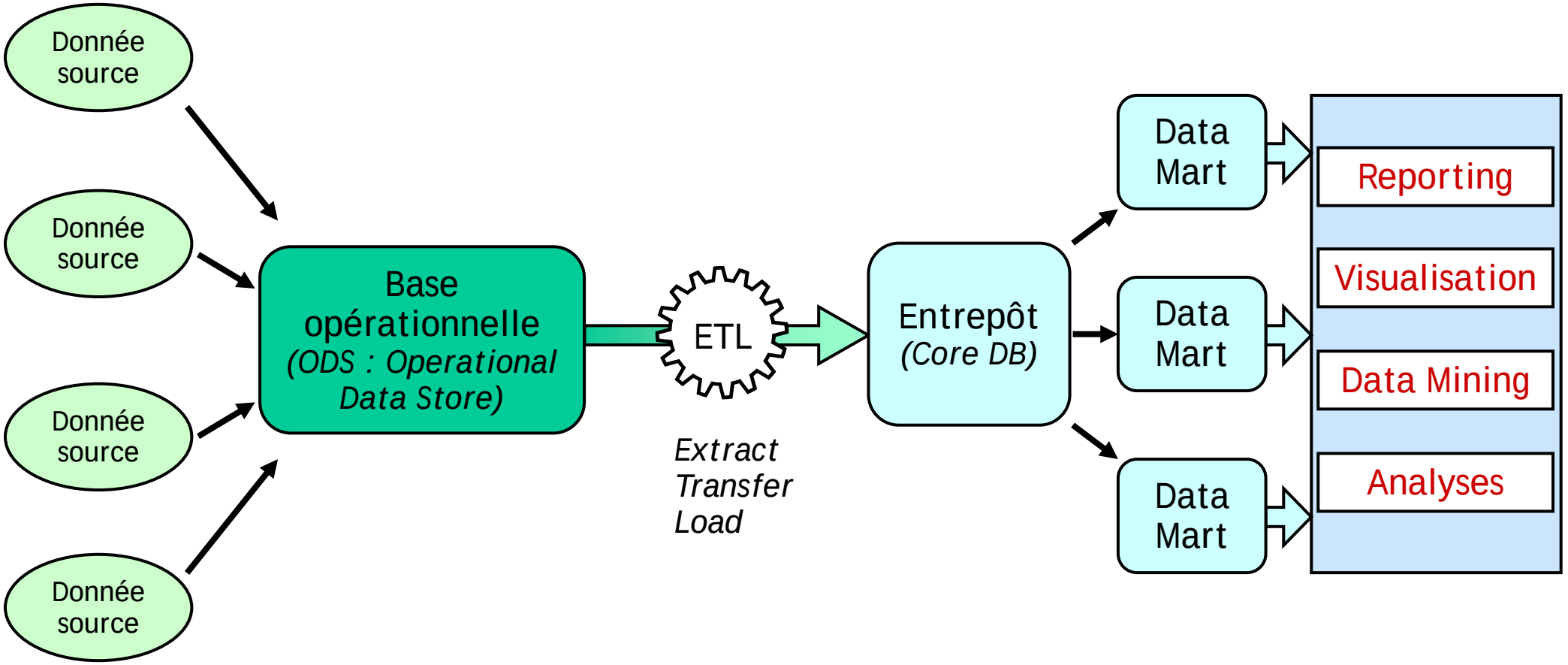
# Constitution d'un entrepôt de données : 3. Alimentation

- Préparation des données
  - Standardisation/homogénéisation des codes et valeurs utilisées
  - +/- normalisation des valeurs
    *(ex : transformation logarithmique, bornage, etc.)*
  - +/- « dénormalisation » des base de données pour obtenir des données « à plat »
  - +/- discrétisation de valeurs

# Constitution d’un entrepôt de données : 4. Publication

- Avant la publication effective il faut vérifier la cohérence globale des données chargées
- Les procédures de certification de la qualité des données chargées dépend de la nature des données, de l’organisation et du domaine de l’entrepôts

# Schéma logique du traitement des données




*[d'après O. Brazhnik et J.F. Jones, J Biomed Inform 2007, 40(3):252-69]*