INSEA

MASTER M2SI

BASE DE DONNÉES DÉCISIONNELLE

Pr. HILAL

2020/2021

INTRODUCTION

Positionnement Technologique

- organiser les données dans des formes papiers dans des classeurs
- mieux organiser les données des organismes
- stocker les données pour fournir un historique détaillé de l'activité
- utiliser les données pour la prise de décision

Problématiques

- Stockage des gigas de données, mais difficile à exploiter et à accéder
- Besoin d'analyser les données dans tous les sens
- Besoin d'obtenir les informations plus facilement

Utilisation de l'information pour appuyer une prise de décisions plus fondées sur des faits

PROBLÉMATIQUES

Quelles sont les promotions les plus efficaces ?

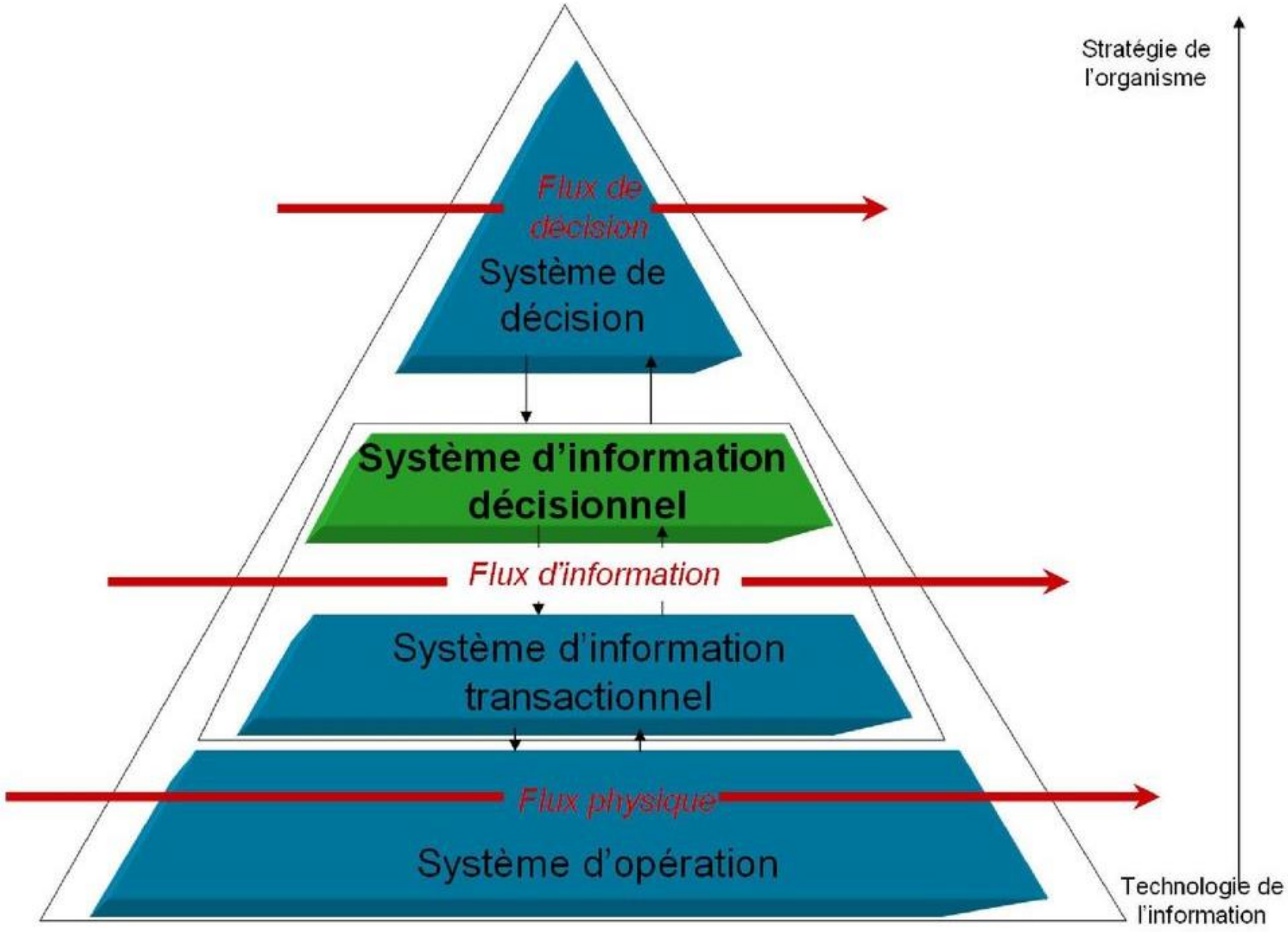
Qui sont mes clients? Quels produits achètent-ils?

Quels sont mes meilleurs clients?

Quelle est la période la plus lucrative?

Qui sont les clients insatisfaits? Pourquoi?

Positionnement Organisationnel



Business Intelligence

L'informatique décisionnelle (***Business Intelligence (BI)***), également appelée « intelligence d'affaires » , désigne les solutions informatiques apportant une aide à la décision avec, en bout de chaîne, des rapports et des tableaux de bord à la fois **analytiques** et **prospectifs**.

Le but est de consolider les **informations** disponibles au sein des bases de données de l'entreprise.

Business Intelligence

- **Traitement opérationnel:** Les opérations quotidiennes comme la capture, le stockage et la manipulation des données.

- **Traitement décisionnel :** L'analyse de ces données ou d'autres formes d'informations pour appuyer la prise de décision

Motivations

- Réconciliation sémantique
  - Dispersion des sources de données au sein d'une entreprise
  - Différents codage pour les mêmes données
  - L'entrepôt rassemble toutes les informations au sein d'un unique schéma
  - Conserve l'historique des données
- Performance
  - Les données d'aide à la décision nécessitent une autre organisation des données
  - Les requêtes complexes de décision dégradent les performances des requêtes Transactionnelles.
- Disponibilité
  - La séparation augmente la disponibilité
  - Une bonne façon d'interroger des sources de données dispersées
- Qualité des données

DÉFINITION DATAWAREHOUSE

Un Datawarehouse est une collection de données conçue pour l'**interrogation** et l'**analyse** plutôt que le traitement de transactions. Il contient généralement des données historiques dérivées de données transactionnelles, mais il peut comprendre des données d'autres origines.

Les Datawarehouses séparent la charge d'analyse de la charge transactionnelle. Ils permettent aux entreprises de consolider des données de **différentes origines**. Au sein d'une même entité fonctionnelle, le datawarehouse joue le rôle d'outil analytique.

BASES DE DONNÉES VS DATAWAREHOUSE

Les **SGBD** sont des systèmes conçus pour l'OLTP (On-Line Transaction Processing).

Permet d'**insérer, modifier, interroger** des informations rapidement, efficacement, en sécurité.

Deux **objectifs** principaux :

- ajouter, retrouver et supprimer des enregistrements repérés par **une clef**

"rechercher une aiguille dans une botte de foin"

- ces opérations doivent pouvoir être effectuées très rapidement, et par de **nombreux utilisateurs simultanément**.

Les systèmes OLTP sont mal adaptés à l'analyse de données.

BASES DE DONNÉES VS DATAWAREHOUSE

Les entrepôts sont des systèmes conçus pour **l'aide à la prise de décision**.

Les objectifs principaux sont

regrouper, organiser, coordonner des informations provenant de sources **diverses**,
les **intégrer** et les **stocker** pour donner à l'utilisateur une vue orientée métier,
retrouver et **analyser** l'information facilement et rapidement.

Questions typiques :

Quels sont les produits qui se vendent le mieux dans chaque région, et quel est l'impact des données démographiques sur ces résultats de vente ?

BASES DE DONNÉES VS DATAWAREHOUSE

	BD- OLTP	Entrepôts
Objectif	collecte de données opérations au jour le jour	consultation et analyse
Utilisateurs	un département (Employé)	transversal (Gestionnaire)
Types de données	données de gestion (données courantes)	données d'analyse (données historiques)
Informations	détaillées	détaillées + agrégées
n-uplets accédés	dizaines	millions
Opérations	requêtes simples, pré-déterminées sélections et mises à jour nombreuses transactions transactions courtes temps réel recherche d'enregistrements détaillés	requêtes complexes, ad-hoc sélections peu de transactions transactions longues batch agrégations et group by

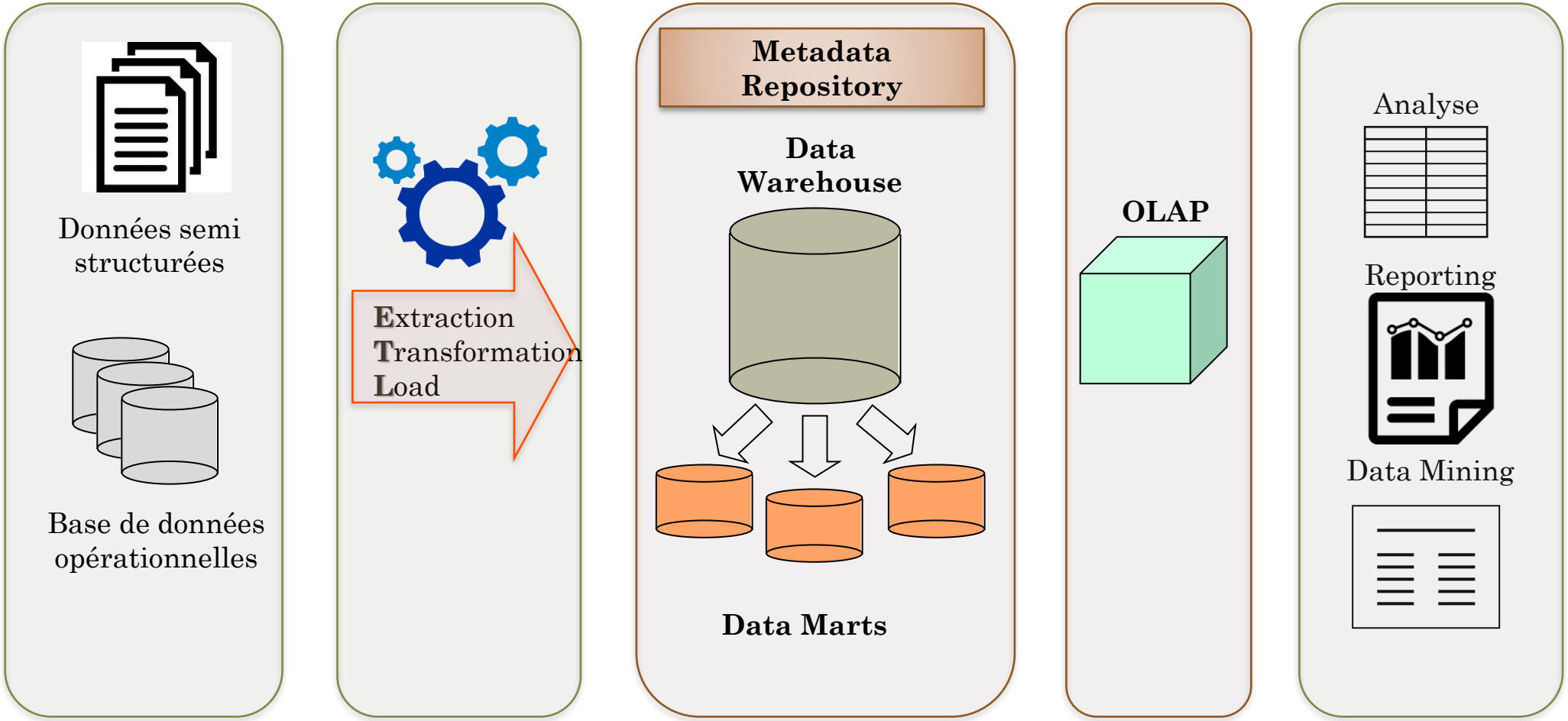
DATAWAREHOUSE VS DATAMINING

- L'**analyse multidimensionnelle** consiste à modéliser des données selon plusieurs axes: **Datawarehouse**

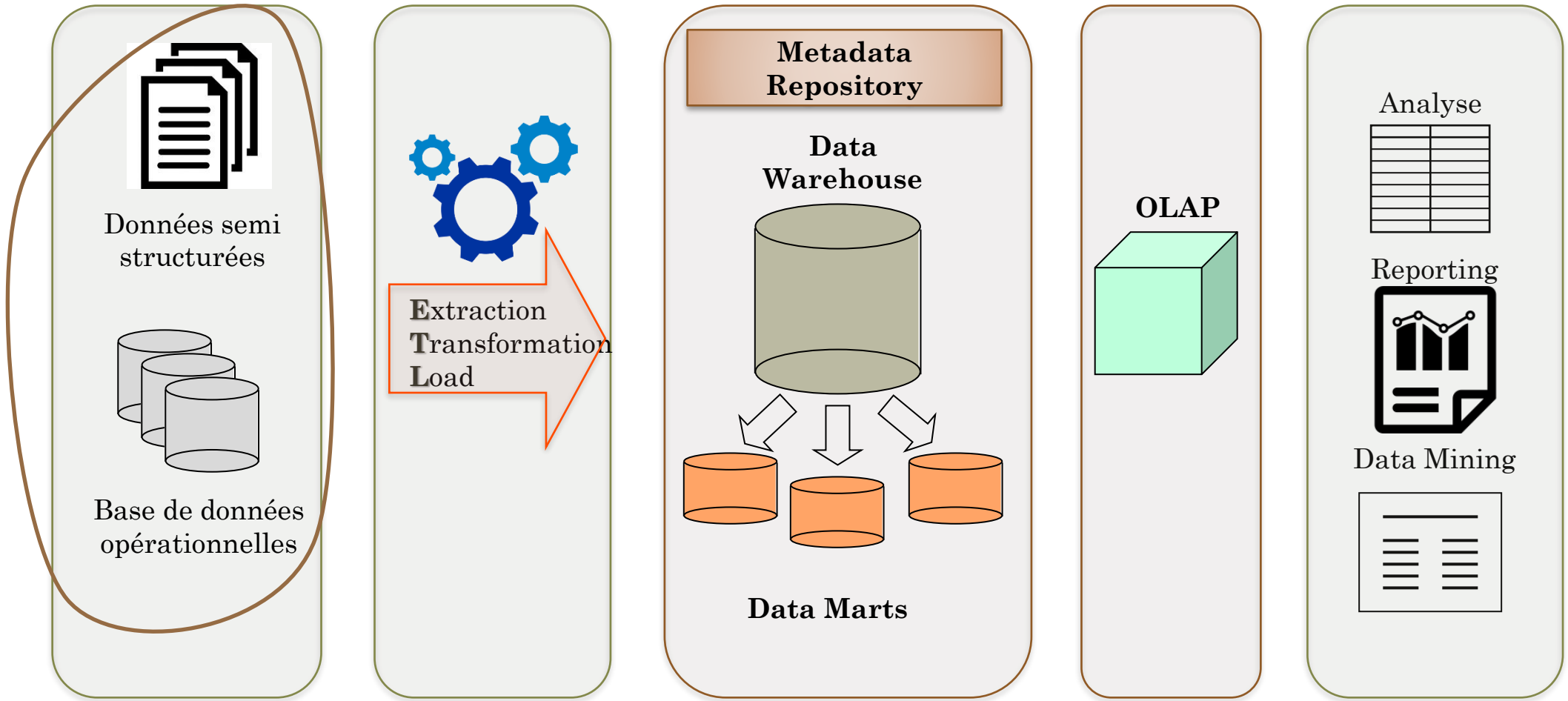
- L'**analyse exploratoire ou prédictive** exploite un ensemble d'événements observés et historisés afin de prévoir l'évolution d'une activité : **Datamining**

Chapitre 1: Architecture d'un Datawarehouse

ARCHITECTURE



ARCHITECTURE



Les sources de données

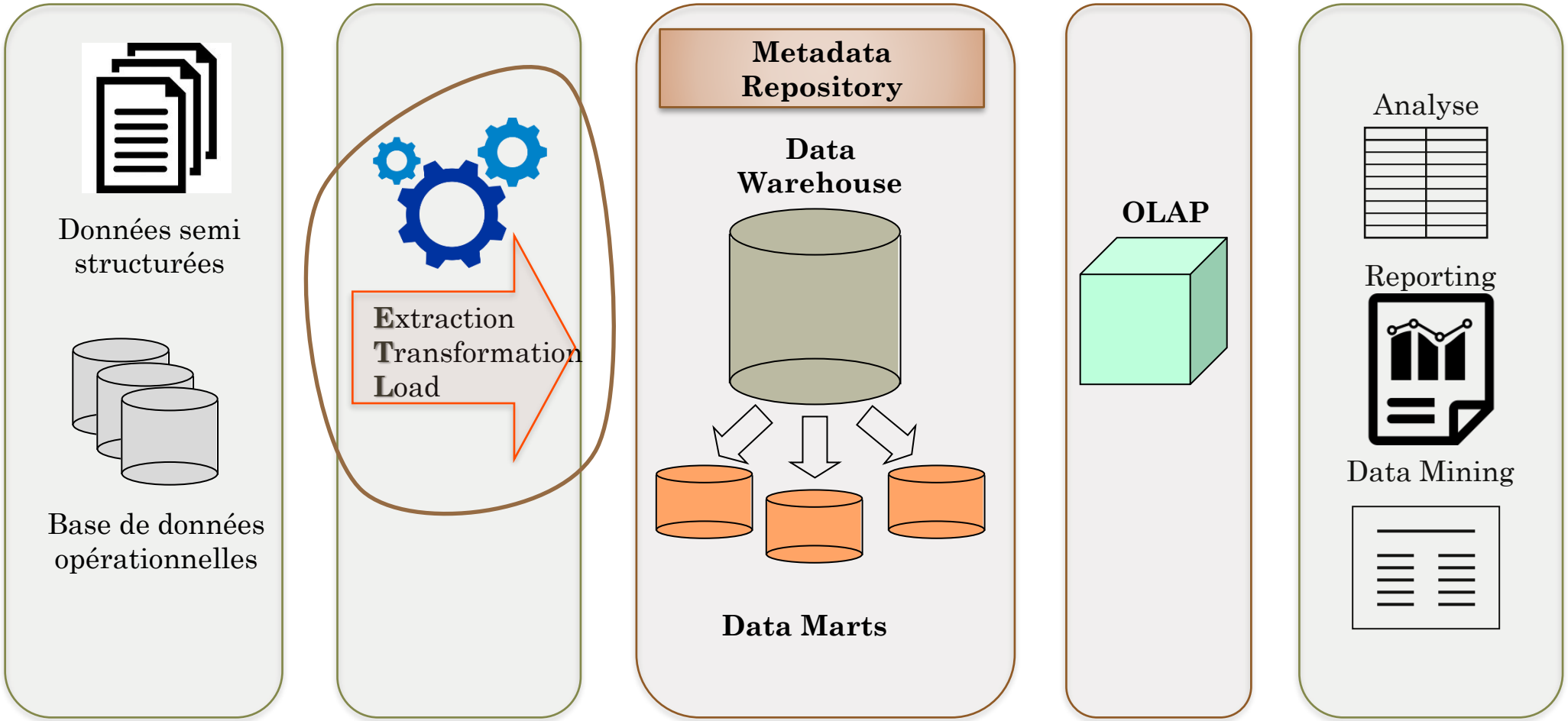
Sources internes

- Bases de production de l'entreprise
- Bases créées par les utilisateurs
- Fichiers semi structurés

Sources externes à l'entreprise

- Données achetées à des fournisseurs de données
- Données récupérées sur Internet
- Données de veille

ARCHITECTURE



ETL

- Extraction (**E**xtraction)
- Transformation (**T**ransformation)
- Chargement (**L**oad)

Un **ETL** est un outil permettant d'automatiser les chargements des données dans le Datawarehouse. Il transpose le modèle entité-relation des bases de données de production ainsi que les autres modèles utilisés dans les opérations de l'entreprise, en modèle à base de **dimensions** et de **faits**.

Un ETL permet de :

- découvrir, analyser et extraire les données à partir des ressources hétérogènes
- nettoyer et standardiser les données
- charger les données dans un Datawarehouse

ETL

Extraction

- Extraction possible à partir de plusieurs plateformes
- Chargement incrémental ou complet

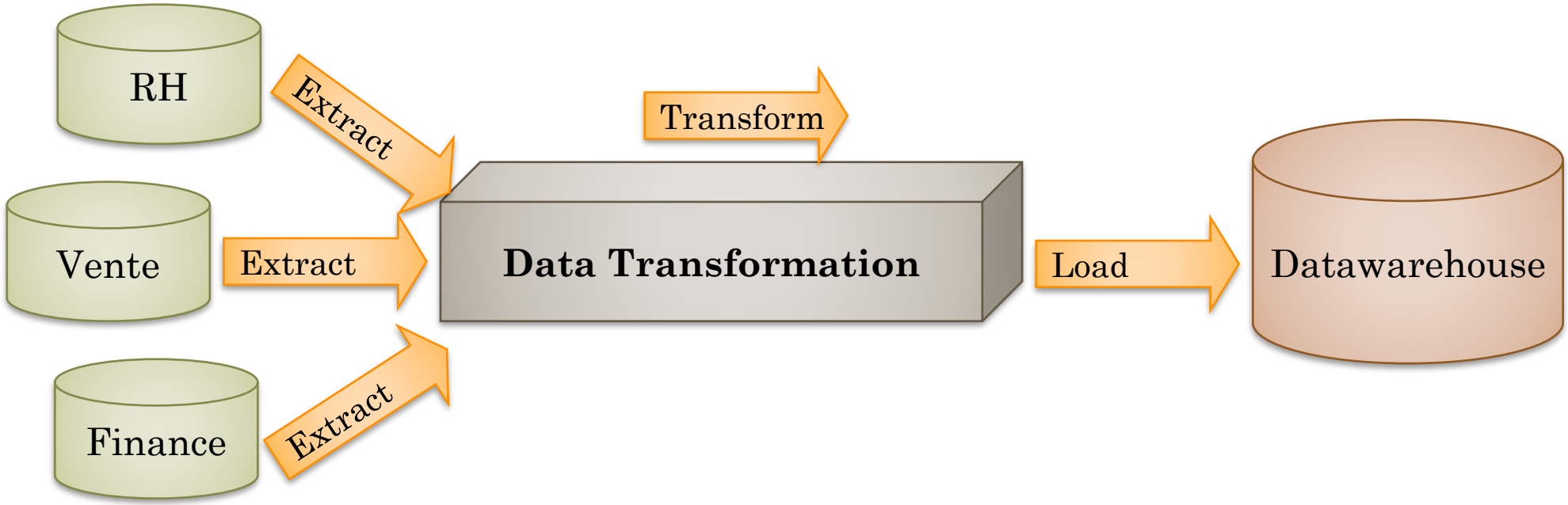
Transformation

- Uniformiser l'information
- Gérer les différents codes
- Majuscule / minuscule
- Orthographe

Load

- Les tables
- Les agrégats

ETL


RH
Vente
Finance
Extract
Extract
Extract
Transform
Data Transformation
Load
Datawarehouse

ETL

Problèmes de normalisation

- Différents encodages, langues ..
- Différentes abréviations
- Equivalence sémantique
- Différentes normes & unités de mesures

Problèmes de données

- Champs manquants (âge, adresse, …)
- Valeurs incorrectes
- Redondance sémantique

Incohérences

- Incohérence des codes
- Incompatibilité référentielle

Transformation

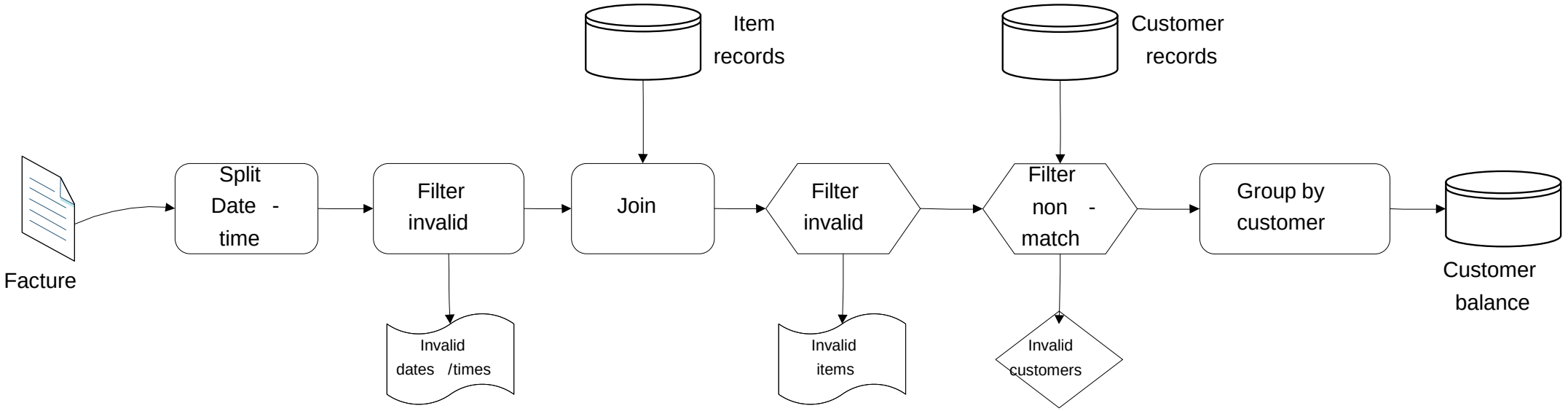
- Révisions de format
- Traitement des valeurs NULL
- Valeurs calculées & dérivées
- Fusion des données
- Fractionnement des champs
- Conversion des unités de mesure
- Conversion des dates
- Déduplication

ETL – Qualité de données

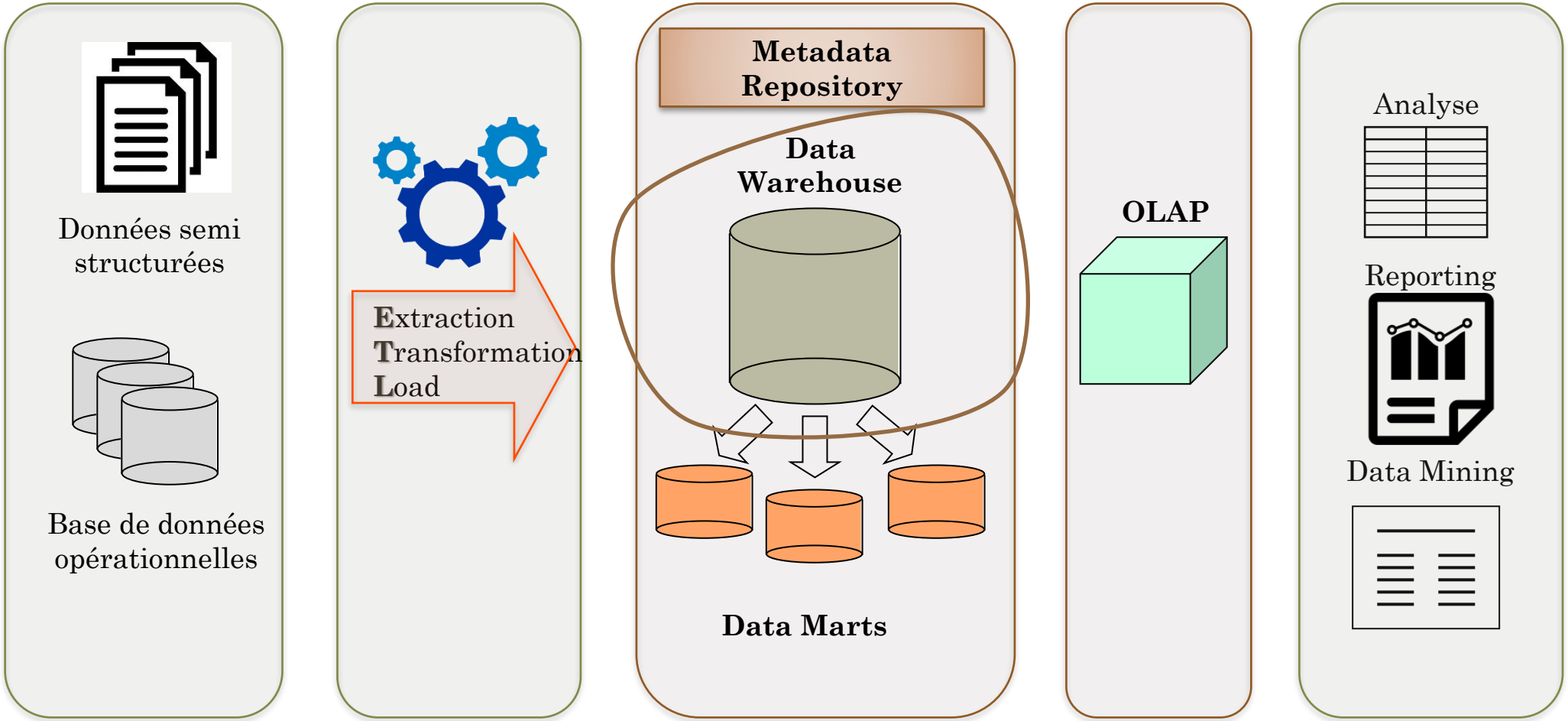
Les données dans un DW doivent être:

- **Précises**: Les données doivent correspondre à un sujet bien défini
- **Complètes**: Le DW doit comporter toutes les données pertinentes
- **Cohérentes**: Le DW ne doit pas comporter de données contradictoires: les agrégats correspondent aux données détaillées
- **Uniques**: Les mêmes choses s'appellent les mêmes et ont la même clé
- **Rapides**: Le DW doit assurer la mise à jour fréquente des données

ETL EXEMPLE


Item
records
Customer
records
Facture
Split
Date - time
Filter
invalid
Join
Filter
invalid
Filter
non - match
Group by
customer
Customer
balance
Invalid
dates /times
Invalid
items
Invalid
customers

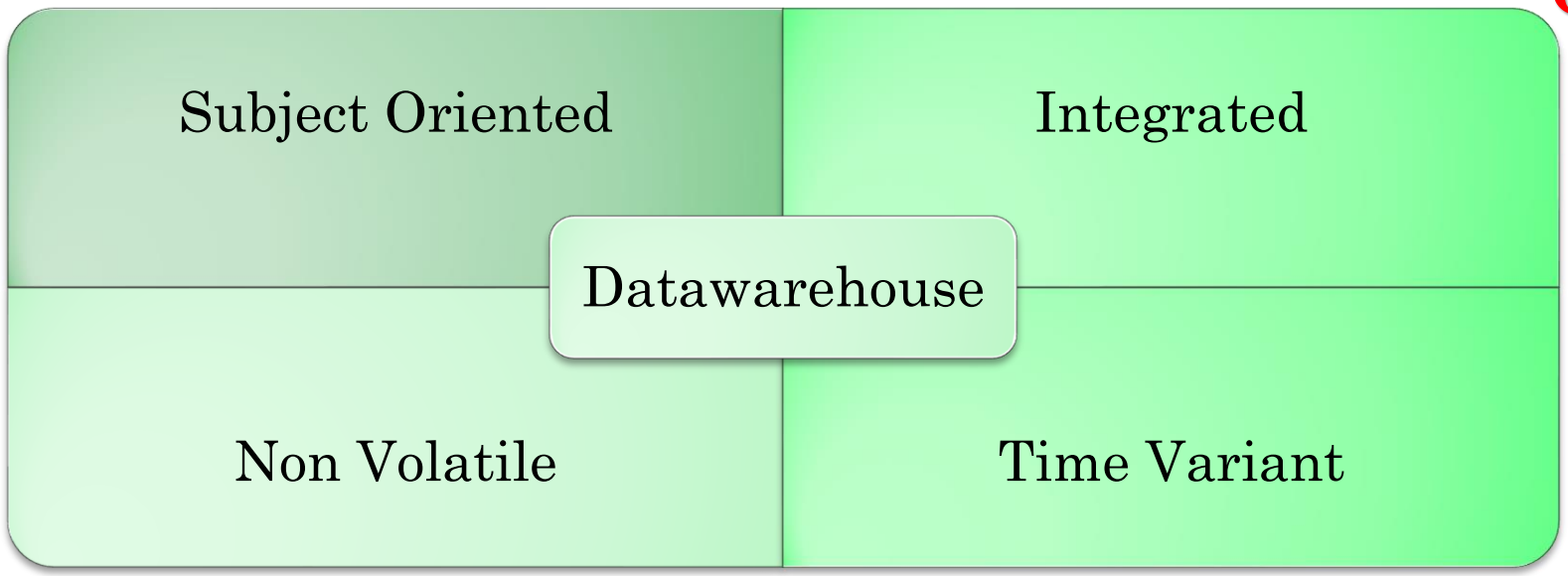
ARCHITECTURE



Propriétés d'un Datawarehouse

Un datawarehouse est une collection de données **intégrées**, **orientées sujet**, **historisées** et **non variables**, susceptibles d'appuyer le processus décisionnel dans une entreprise ou organisation

Oracle



Orientée Sujet

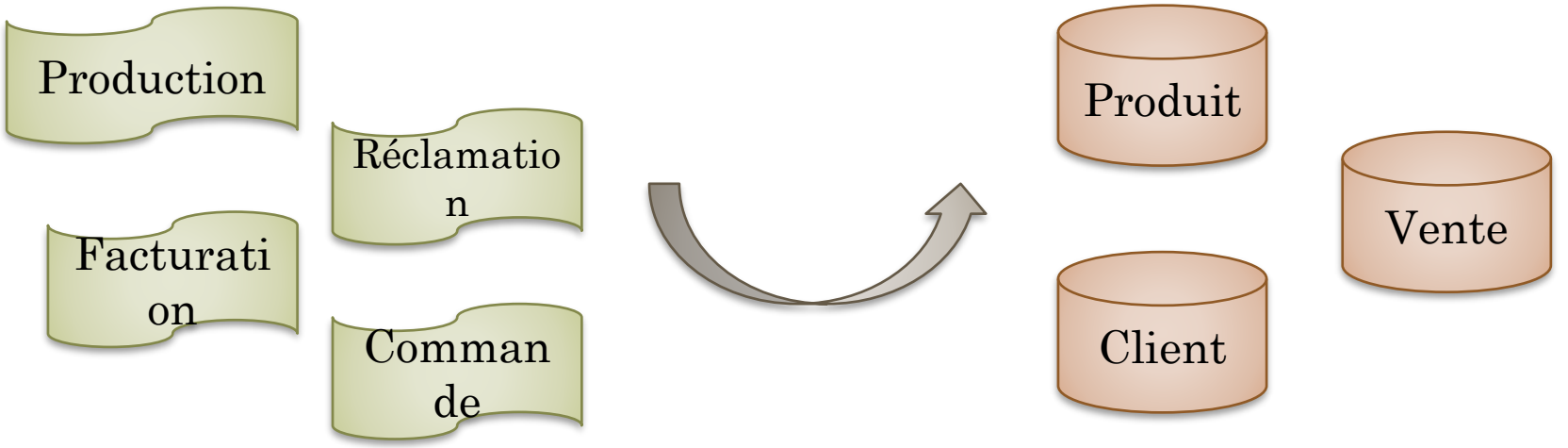
Les entrepôts de données sont conçus pour aider à analyser les données.

L'information est présentée selon un sujet ou un domaine d'intérêt spécifique, pas simplement comme des enregistrements informatiques. Les données fournissent ainsi des informations sur un sujet particulier.

Ventes

- Qui était notre meilleur client pour ce produit l'année dernière?

ORIENTÉE SUJET



Base de données opérationnelle

Datawarehouse

Intégrée

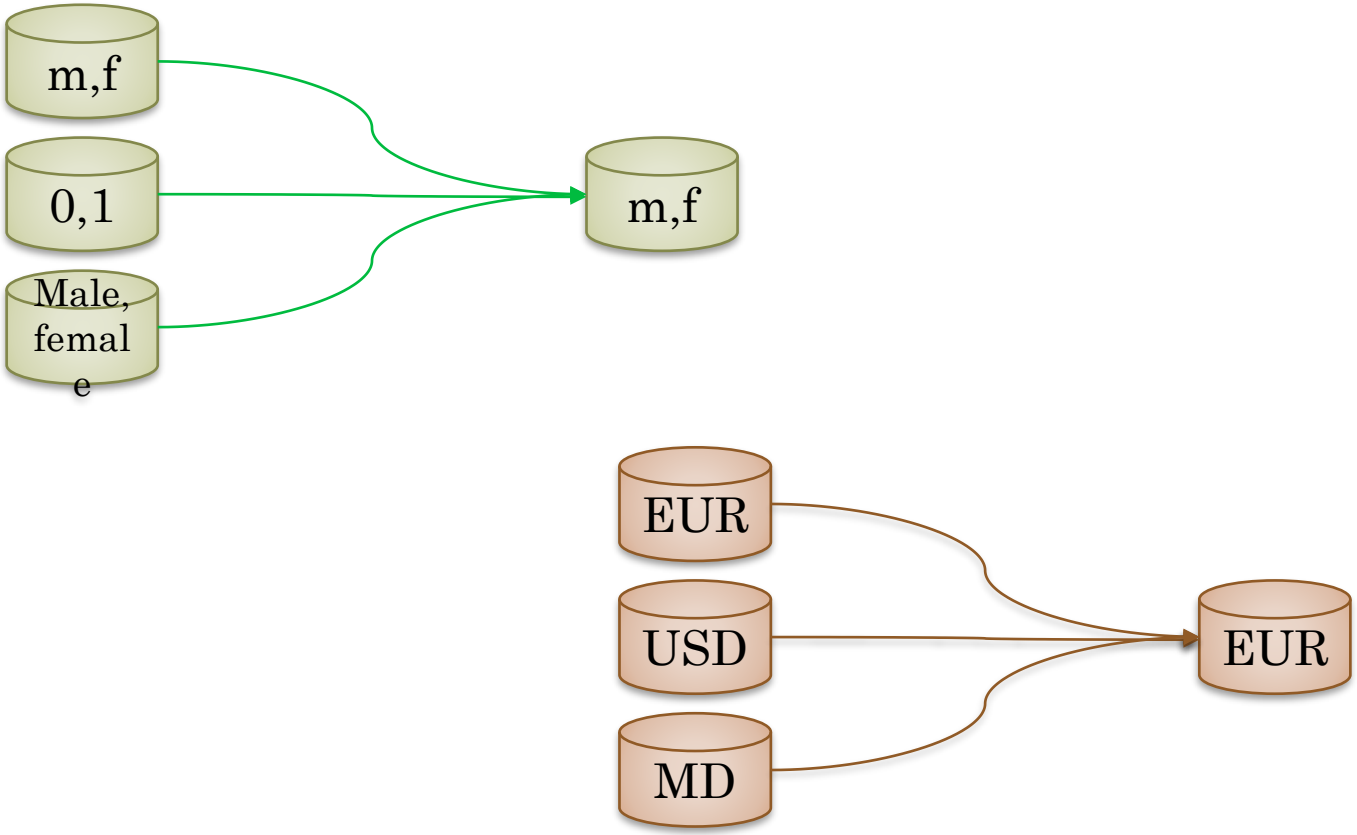
Intégration de données provenant de multiples sources **hétérogènes** dans un **format cohérent**

- Base de données relationnelle
- Fichier
- Enregistrement de log

L'intégration doit résoudre les problèmes suivants

- Conflits de noms
- Incohérences entre les unités de mesure.

Intégrée



Non volatile

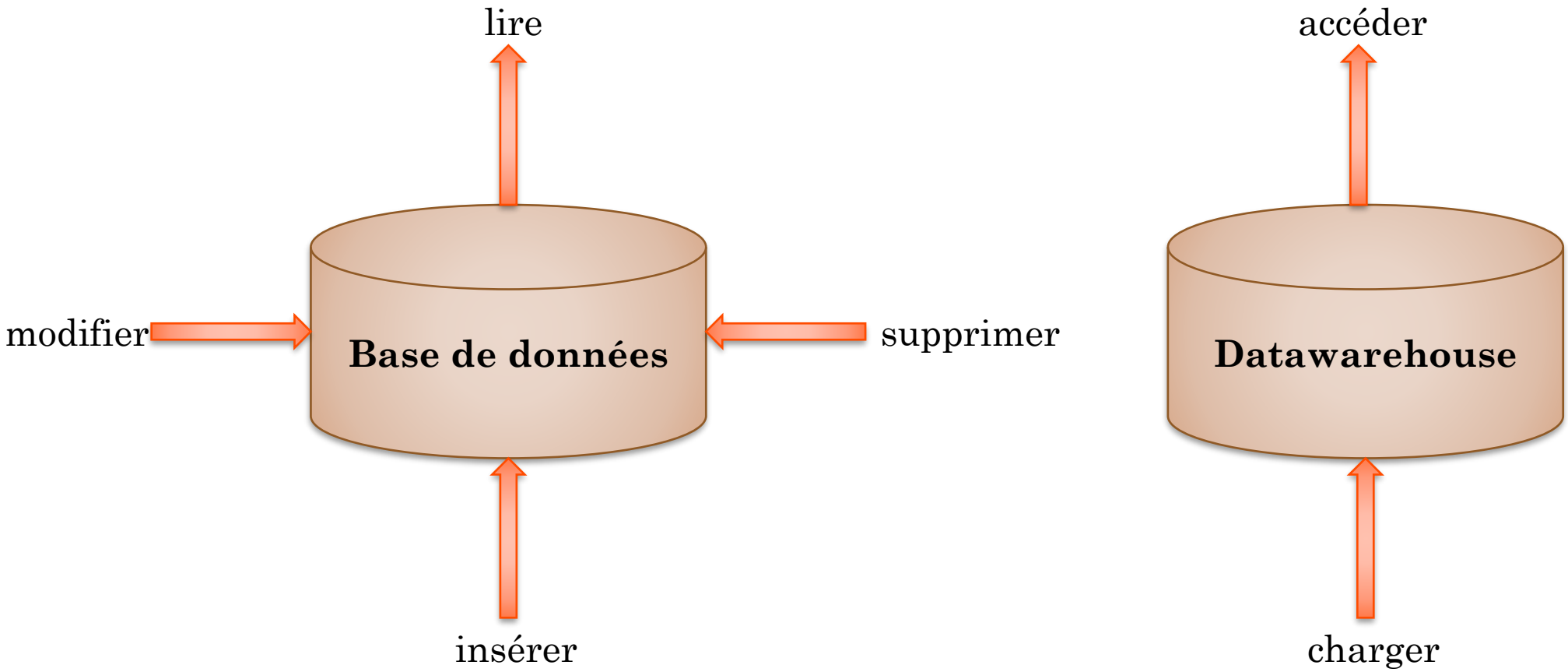
Les données, une fois intégrées dans le datawarehouse, ne devraient pas changer. Le but d'un datawarehouse est de permettre d'analyser ce qui s'est passé.

Le datawarehouse ne permet que deux opérations : le **chargement initial** des données et **l'accès** aux données.

Les données non volatiles sont donc

- Stable & non modifiables
- Accessibles en lecture seule
- Non supprimables

NON VOLATILE



VARIANTE DANS LE TEMPS

Les données **historiques** sont conservés dans un datawarehouse. Il est possible de récupérer des données à partir de 3 mois, 6 mois, 12 mois, ou des données encore plus anciennes. Contrairement aux bases de données traditionnelles où souvent seuls les données les plus récentes sont conservées.

Chaque donnée collectée se voit affecter une **date** ou un **numéro de version**, afin de suivre son évolution au cours du temps et de conserver l'historique.

DW vs SGBD

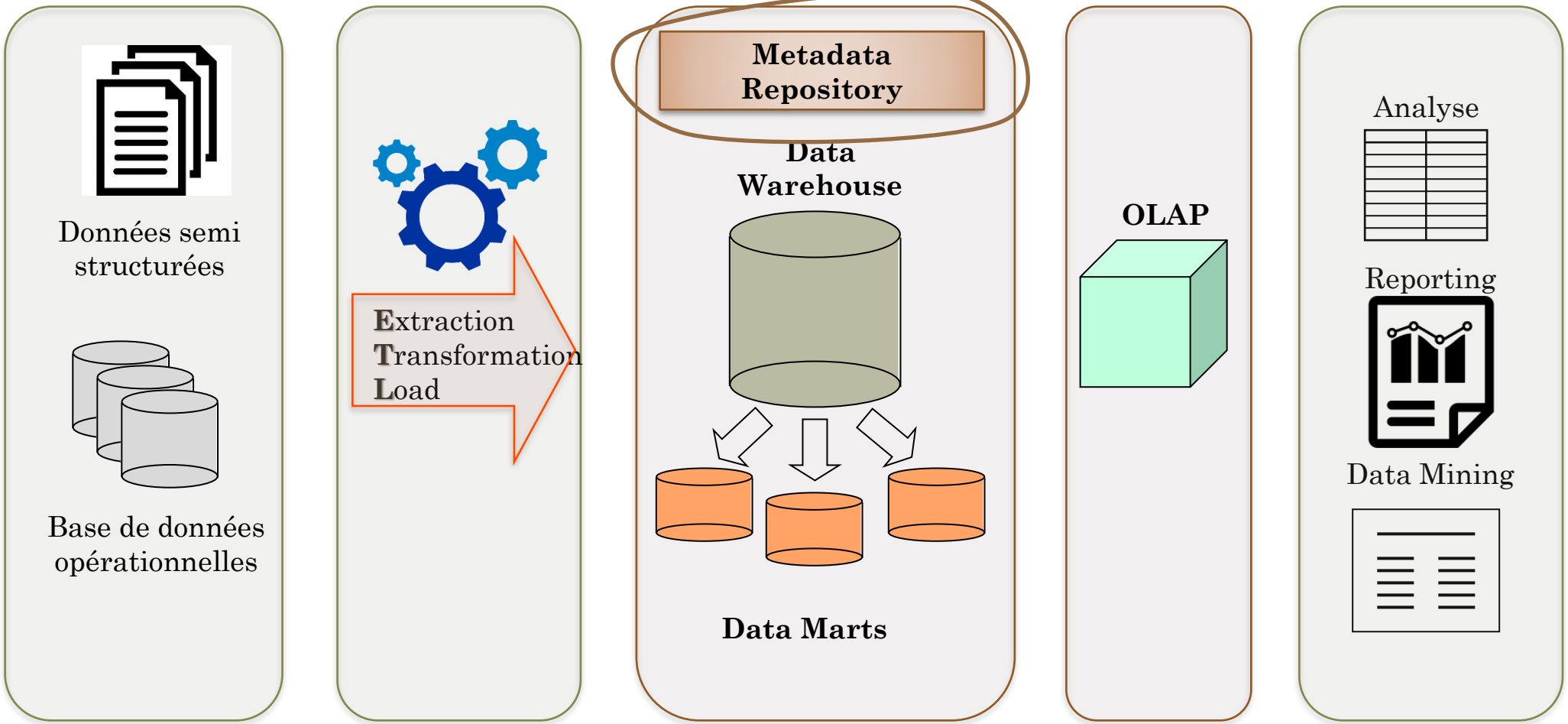
Différente performance

- **SGBD (OLTP)**: méthodes d'accès, indexation, contrôle de concurrence, récupération
- **DW (OLAP)** : requêtes complexes, vue multidimensionnelle, consolidation

Différentes données

- **données manquantes**: l'aide à la décision nécessite des données historiques que les bases de données opérationnelles ne conservent pas généralement
- **données consolidées**: l'aide à la décision nécessite la consolidation (agrégation, synthèse) des données provenant de sources hétérogènes
- **qualité des données**: différentes sources utilisent généralement des représentations de données, des codes et des formats incompatibles qui doivent être uniformisés

ARCHITECTURE



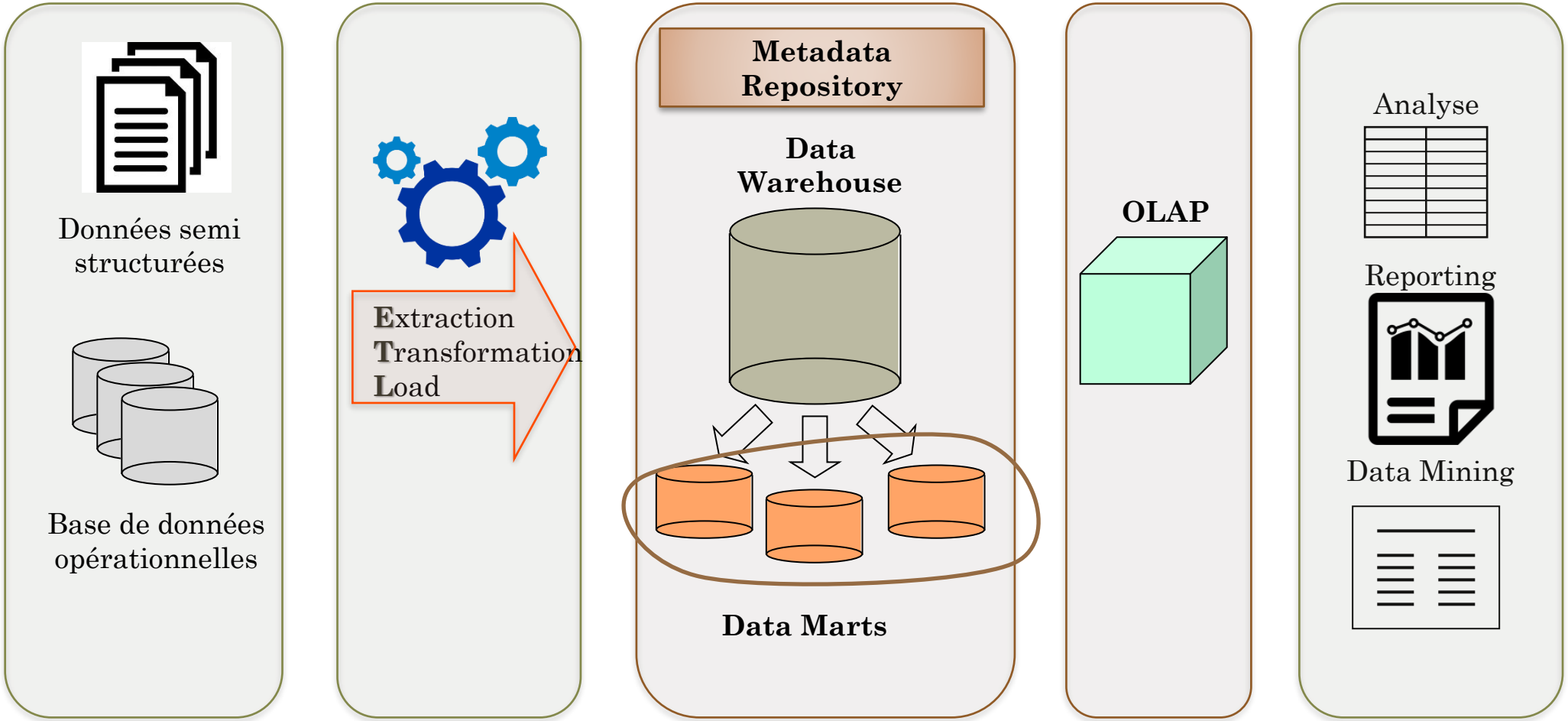
Metadata

Les métadonnées sont les informations relatives à la structure des données, les méthodes d'agrégation et le lien entre les données opérationnelles et celles du Datawarehouse.

Les métadonnées doivent renseigner sur :

- Le modèle de données
- La structure des données telle qu'elle est vue par les développeurs
- La structure des données telle qu'elle est vue par les utilisateurs
- Les sources des données
- Les transformations nécessaires
- Suivi des alimentations

ARCHITECTURE



DATA MART

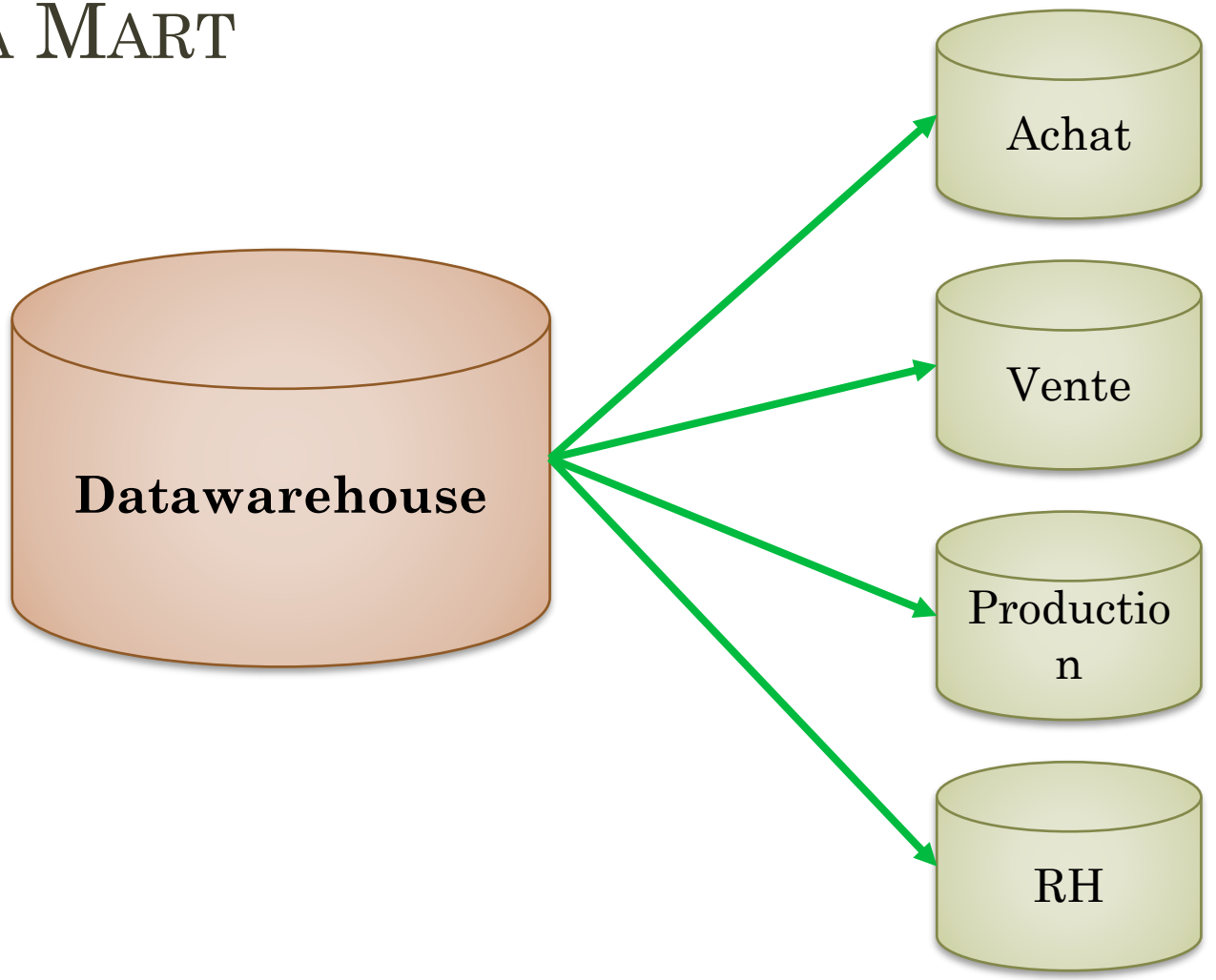
Data Mart ou Magasin de données

Les data marts sont un **sous-ensemble** du datawarehouse ou se produit la plupart des activités d'analyse de l'environnement BI.

Les données de chaque data mart sont généralement adaptées pour une capacité ou une **fonction particulière** (l'analyse de la rentabilité des produits, l'analyse démographique de la clientèle, ...)

Chaque data mart spécifique n'est pas nécessairement valable pour d'autres usages.

Data Mart



Data mart

DATAWAREHOUSE VS DATAMART

Datawarehouse	Data mart
Est défini à l'échelle de l'entreprise	Est défini à l'échelle départemental
Contient plusieurs domaines	Contient souvent un seul domaine
Contient des informations très détaillées	Peut contenir des données plus résumées
Intègre toutes les sources de données	Intègre les informations à partir d'un sujet donné ou d'un ensemble de systèmes sources

Chapitre 2 : Conception d'un Datawarehouse

Différence de Conception

Schéma Relationnel – Entité/Relation

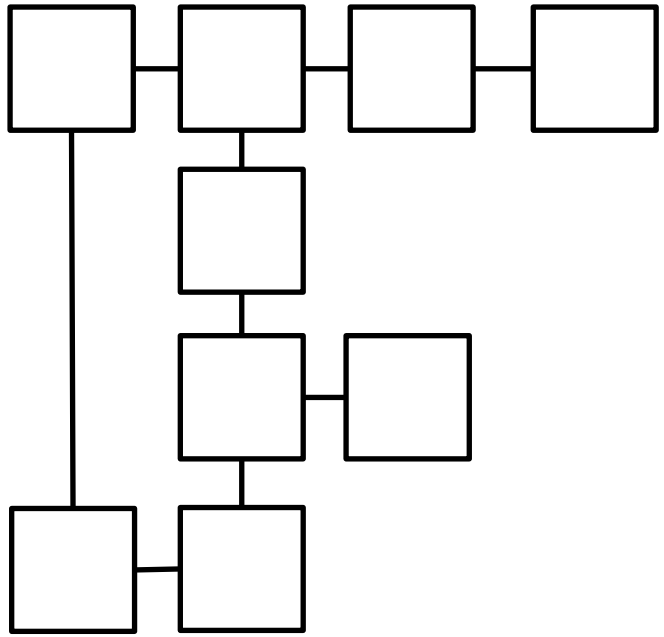
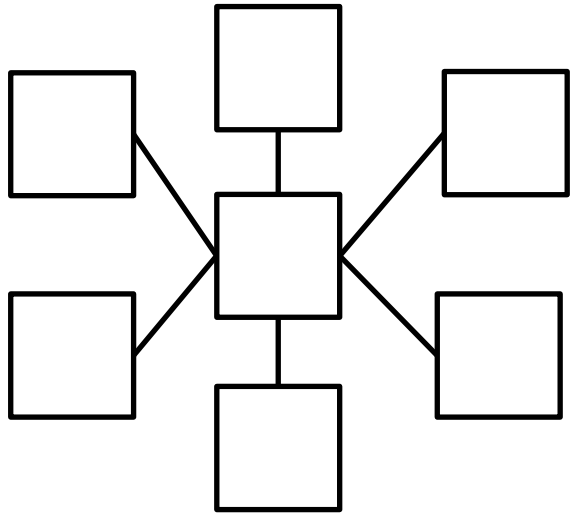


Schéma en Etoile



Modélisation

Schéma en étoile

Une table de faits au milieu du schéma est connectée à un ensemble de tables de dimensions

Schéma flocon de neige (snowflake)

Un raffinement du schéma en étoile où des tables de dimensions sont décomposées

Constellation de faits

Plusieurs tables de faits partagent des tables de dimension (constellation d'étoiles)

Modélisation multidimensionnel

Table de fait

La table de faits constitue une table de référence centrale permettant d'accéder aux événements ou activités archivés et inhérents à un processus déterminé.

Dimension

Les tables de dimension sont des compagnons intégrés à une table de fait.

Les tables de dimension contiennent le contexte textuel associé à un événement de mesure des processus métiers. Ils décrivent le «qui, quoi, où, quand, comment, et pourquoi" associé à l'événement.

Attribut

Les attributs qualifient les dimensions. Généralement, les attributs sont textuels et discrets (par opposition aux faits).

TABLE DE FAITS

Les tables de faits sont des collections de **mesures** associées à un processus métier spécifique. Les mesures sont stockés dans les colonnes.

Les tables de faits contiennent généralement un grand nombre de lignes, parfois dans les centaines de millions d'enregistrements lorsqu'elles contiennent une ou plusieurs années d'historique pour un grand organisme.

Table de faits agrégée

L'agrégation est le processus de calcul des données de synthèse à partir de données plus détaillées. Il permet de réduire la taille des tables de faits par l'agrégation des données dans les comptes rendus analytiques.

Table de faits - Granularité

Le grain détermine le niveau de détail de la mesure de table de faits:

Par exemple

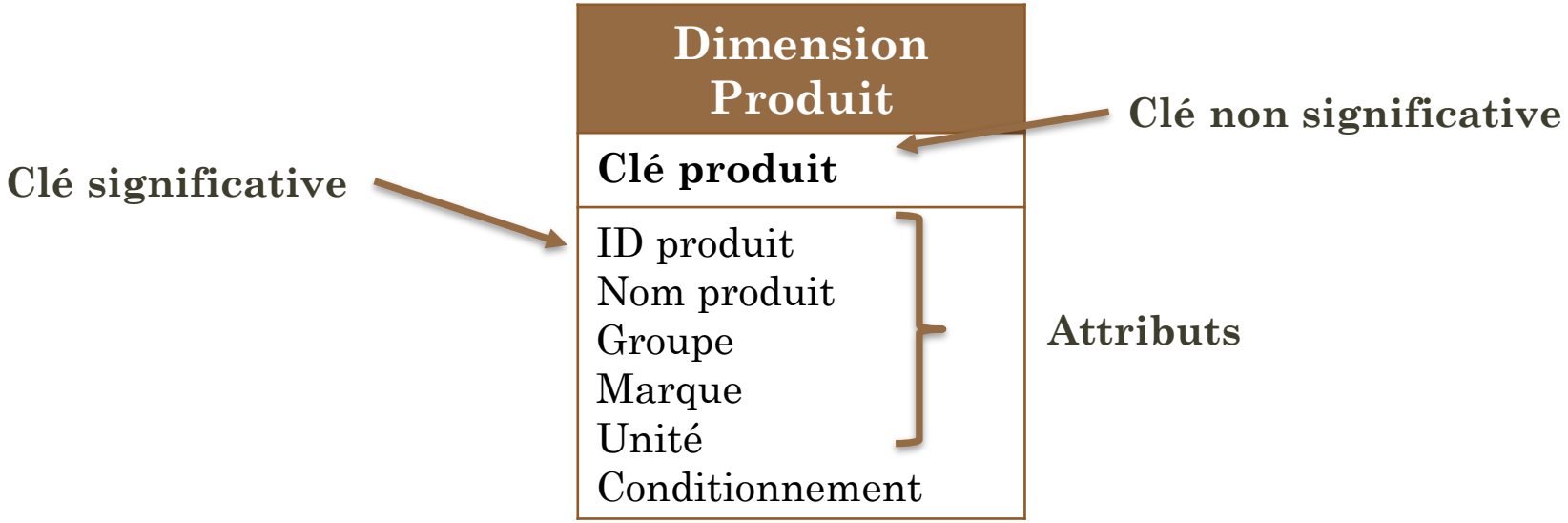
- Transactions individuelles
- Transactions Instantanés (points dans le temps)
- Éléments de ligne sur un document

Le plus petit niveau de granularité est meilleur pour l'analyse (mais pas pour le stockage)

Dimension

Une table de dimension se compose de:

- Une clé non significative établissant un lien avec les lignes de la table de faits
- Une clé significative reprise d'une source de données opérationnelle ou externe
- Un nombre d'attributs permettant de caractériser la dimension



Clé non significative

Aussi appelé clé de substitution (Surrogate Key)

Les clés de table de dimension sont non intelligentes et non liées à l'entreprise car:

- Les clés métier (clé significative) peuvent changer avec le temps
- Les clés de substitution gardent une trace des valeurs des attributs non clés pour une clé significative donnée
- Les clés de substitution sont plus simples et plus courtes
- Les clés de substitution peuvent avoir la même longueur et le même format pour toutes les clés

Slowly Changing Dimension (SCD)

Dimension à variation lente

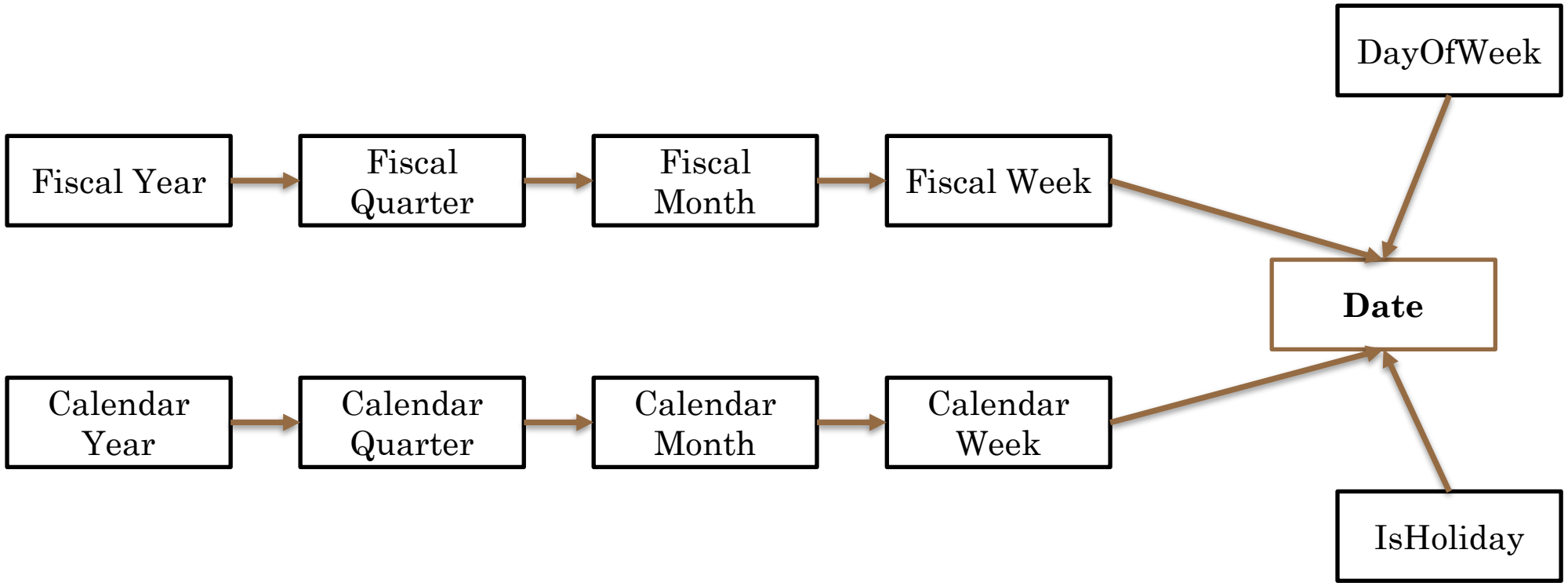
- **Type 1**: remplacer les anciennes données par les nouvelles (perte des données historiques)
- **Type 2**: créer une nouvelle ligne de table de dimension chaque fois que l'objet de dimension change, avec toutes les caractéristiques de la dimension au moment du changement.
  - **Approche la plus courante**
- **Type 3**: pour chaque attribut changeant, créer un champ de valeur courant et plusieurs champs d'ancienne valeur (plusieurs valeurs)

Dimension Temporelle

Dimension Date
Clé date
date complète jour mois année semestre Trimestre ……

- Type particulier de dimension
- Dimension cruciale pour l'analyse
- Dupliquer les attributs même s'ils peuvent être déduits

Dimension Temporelle



Fait vs Dimension

Les observations numériques continuellement estimées

→ **Faits**

Les observations numériques discrètes tirées d'une petite liste

→ **Dimensions**

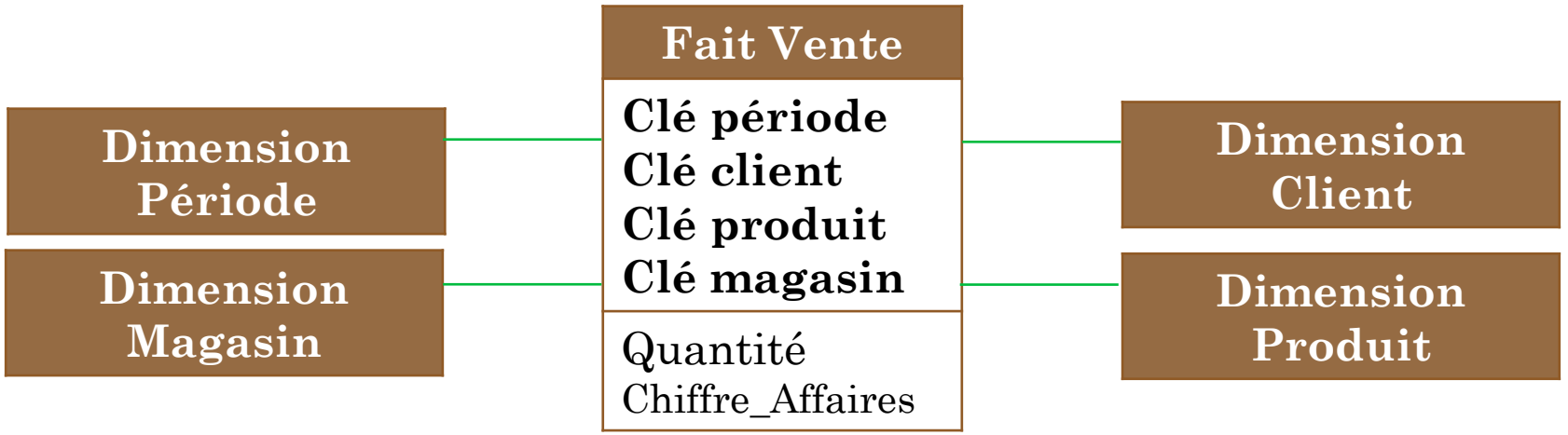
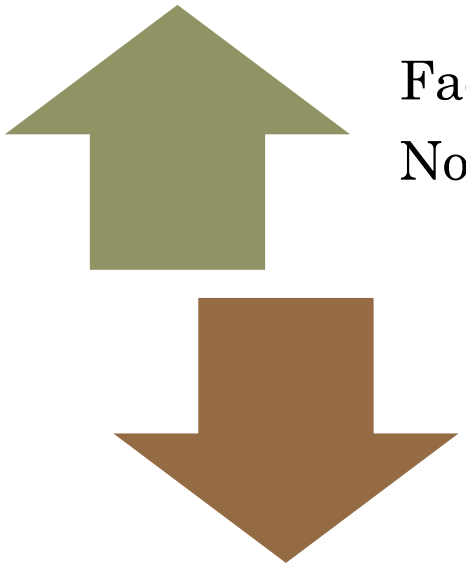


Schéma en étoile

- Une (ou plusieurs) table(s) de faits.
- Plusieurs tables de dimension **dénormalisées**
- Les tables de dimension ne sont pas reliées entre elles



Facilité de navigation

Nombre de jointures limité

Redondance dans les dimensions

Toutes les dimensions ne concernent pas les mesures

Alimentation complexe

Rappel - Normalisation

Définition

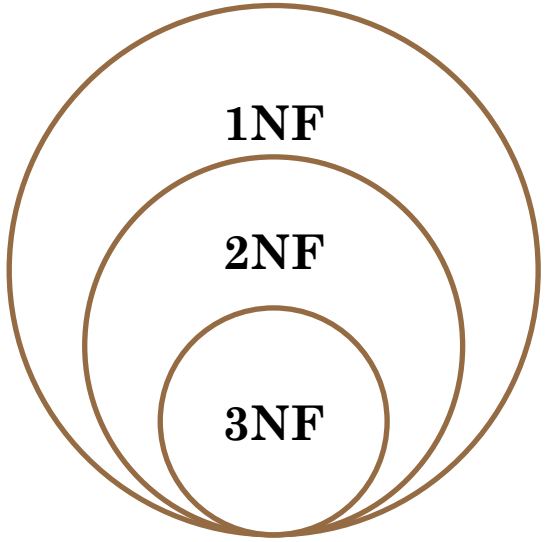
- La normalisation est le processus qui permet d'éviter les données redondantes dans les bases de données.
- Cela implique de restructurer les tables pour atteindre successivement des formes plus élevées de normalisation.

- Une base de données correctement normalisée respecte les caractéristiques suivantes:
  - Les valeurs scalaires (atomiques) dans chaque champs
  - Absence de redondance.
  - Utilisation minimale des valeurs nulles.
  - Perte minimale d'informations.

Rappel - Normalisation

Il existe une séquence aux formes normales:

- 1NF est considéré comme le plus faible,
- 2NF est plus fort que 1NF
- 3NF est plus fort que 2NF



Chaque niveau supérieur est un sous-ensemble du niveau inférieur

Rappel - Formes normales

Première forme normale

Une relation est en première forme normale si et seulement si tout attribut contient une valeur atomique.

Livre(<u>ID_livre</u>, titre, auteurs)

→ *Livre(<u>ID_livre</u>, titre, auteur1, auteur2, auteur3)*

ou mieux encore

Livre(<u>ID_livre</u>, titre, auteurs)

→ *Livre(<u>ID_livre</u>, ID_auteur, titre)*

→ *Auteur(ID_auteur, nomAuteur)*

Rappel - Formes normales

Deuxième forme normale

Les attributs non clé dépendent de toute la clé et non d'une partie de la clé

Commande(ID_fournisseur, refArticle, raisonSocialeFournisseur, adresseFournisseur, quantité, prix)

→ *Fournisseur(ID_fournisseur, raisonSociale, adresse)*

Commande(ID_commande, ID_fournisseur, refArticle, quantité, prix)

Pour qu'une table soit en 2NF, il faut que:

- **La table soit en première forme normale**
- **Tous les attributs non clés de la table doivent être fonctionnellement dépendants de la clé primaire entière**

Rappel - Formes normales

Troisième forme normale

Chaque attribut de la relation ne dépend que de la clé et pas d'un autre attribut de la relation

Employe(ID_Employe, nomEmploye, posteEmploye, ID_service, nomService)

→ *Employe(ID_Employe, nomEmploye, posteEmploye, ID_service)*

→ *Service (ID_service, nomService)*

Pour qu'une table soit en 3NF, il faut que:

- **La table soit en deuxième forme normale**
- **Aucun attribut ne dépend de manière transitoire de la clé primaire**

Schéma en étoile

Un schéma en étoile contient une seule table centrale, appelée une **table de faits**, entouré de plusieurs tables appelées **dimensions**.

Une schéma en étoile couvre un secteur d'activité. Dans ce cas, le schéma couvre les ventes d'une entreprise. Un datawarehouse couvre plusieurs domaines d'activité et se compose de plusieurs schémas étoiles et/ou flocon de neige.

Schéma en étoile

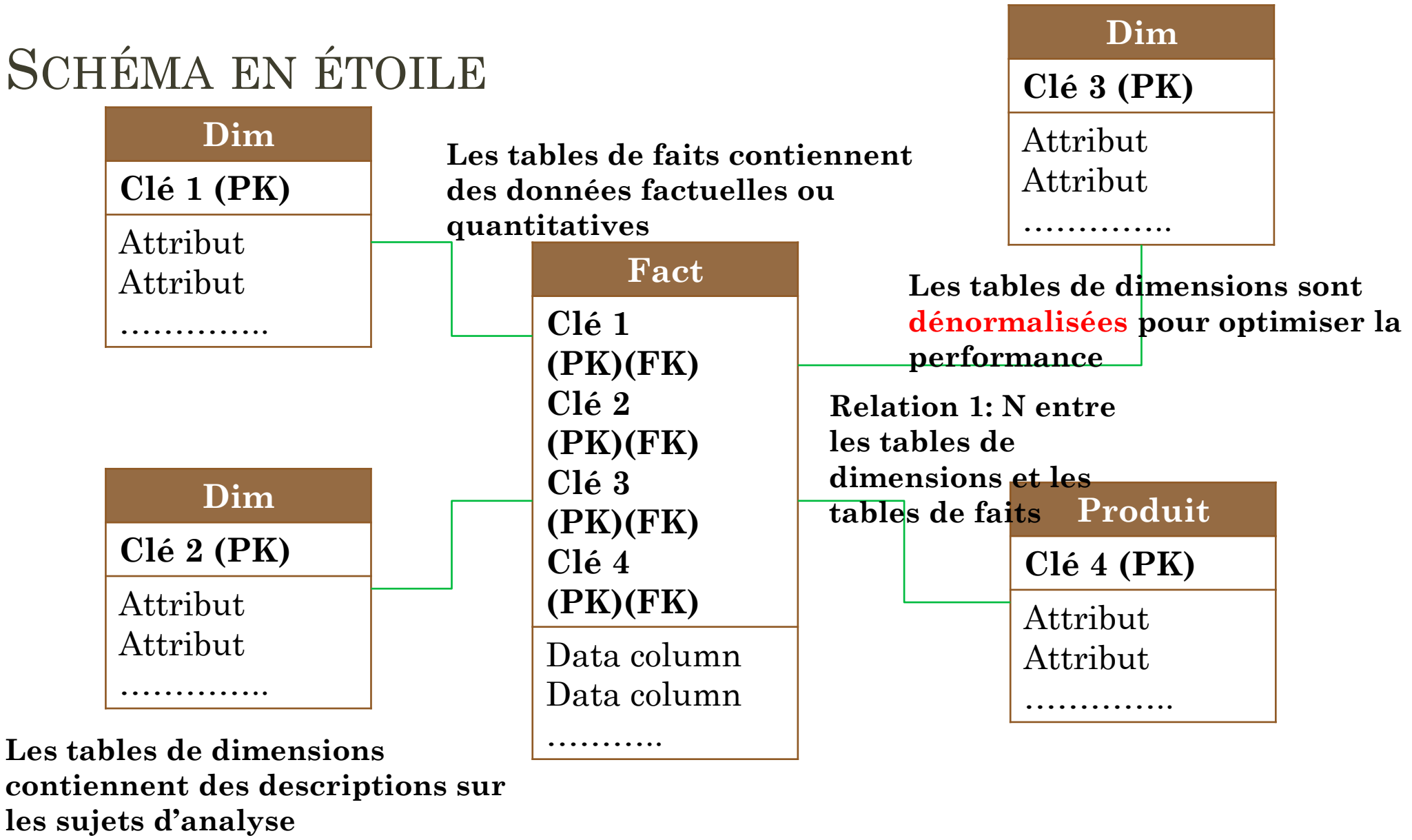


Schéma en étoile

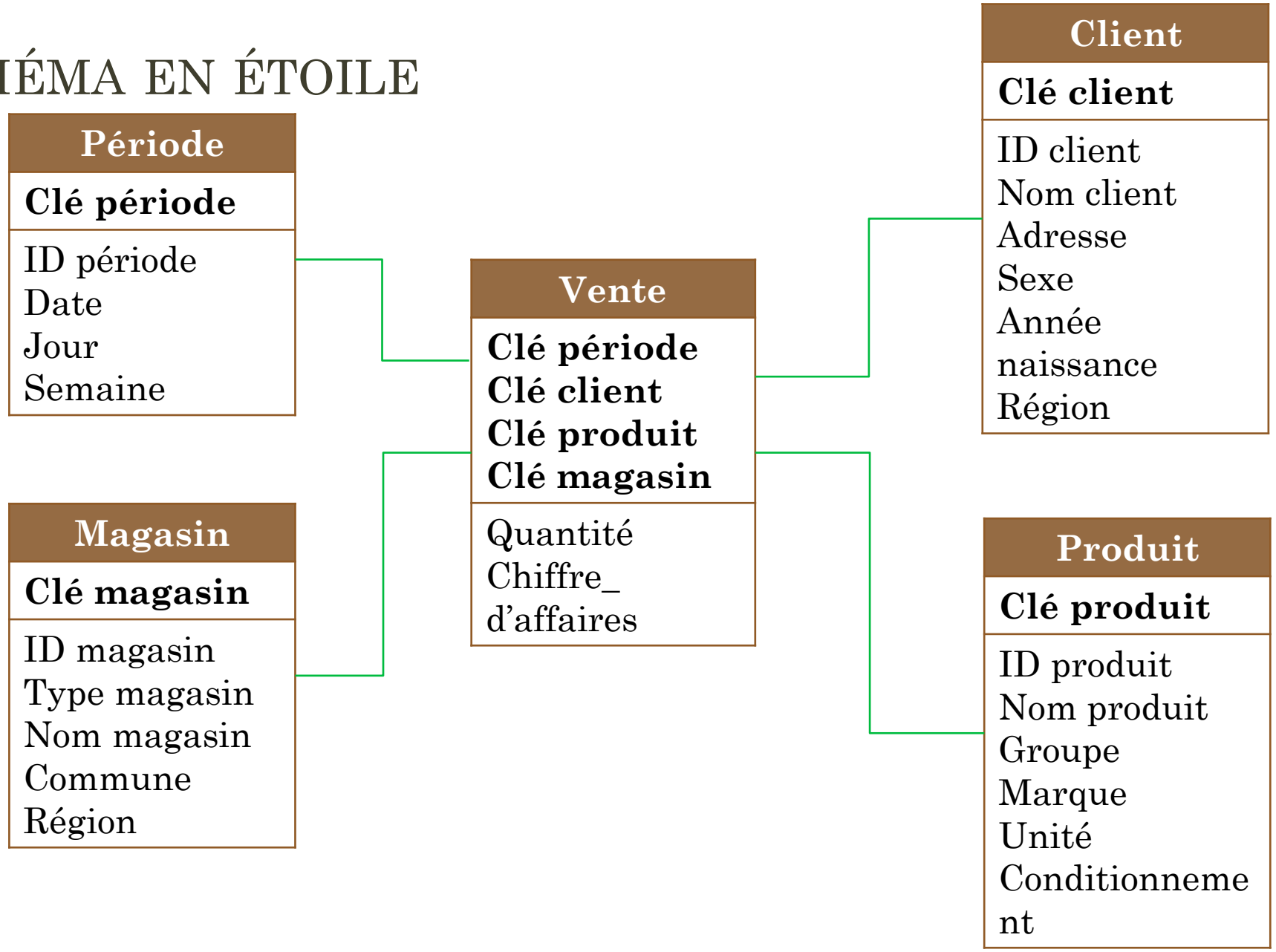


Schéma en étoile

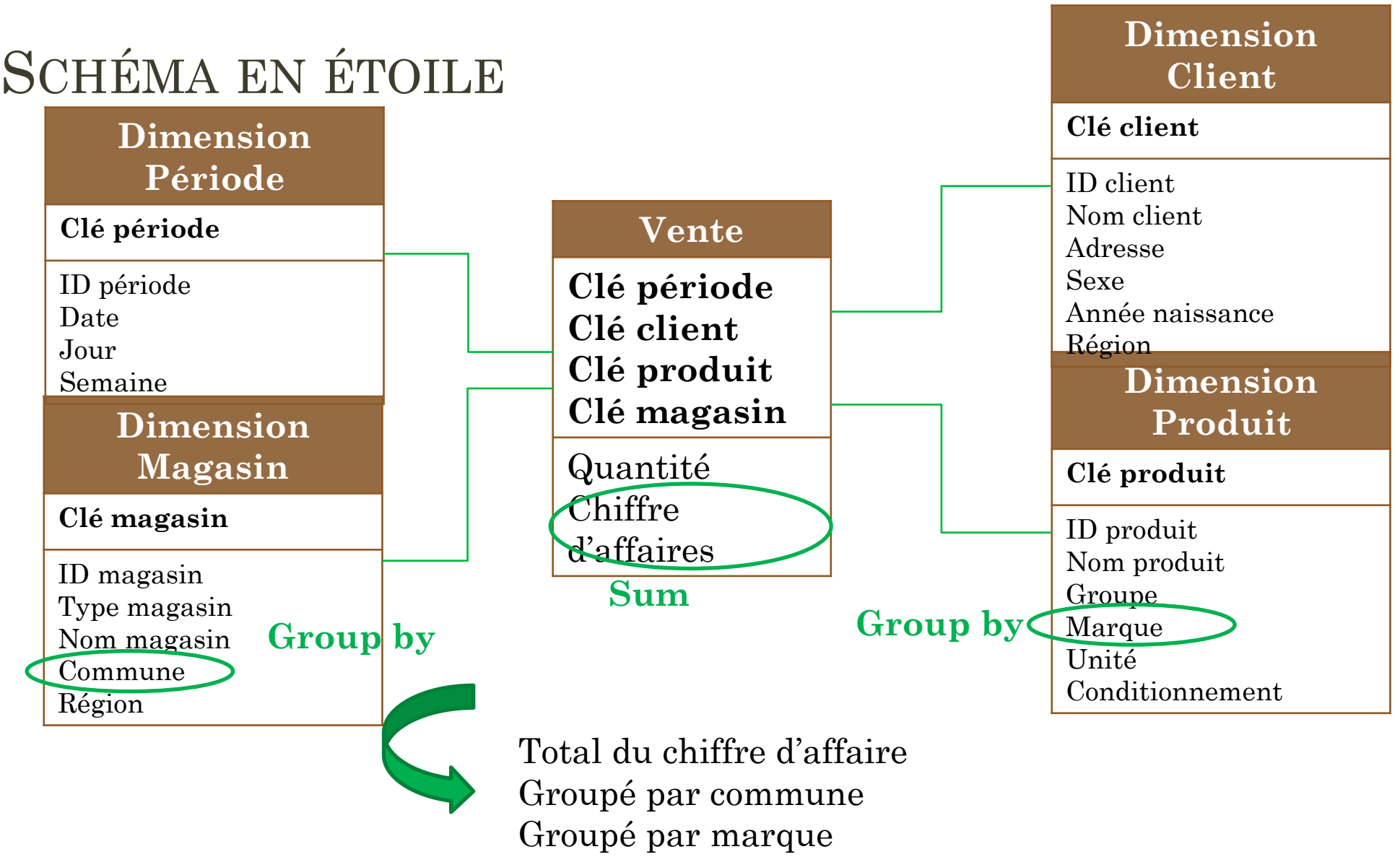
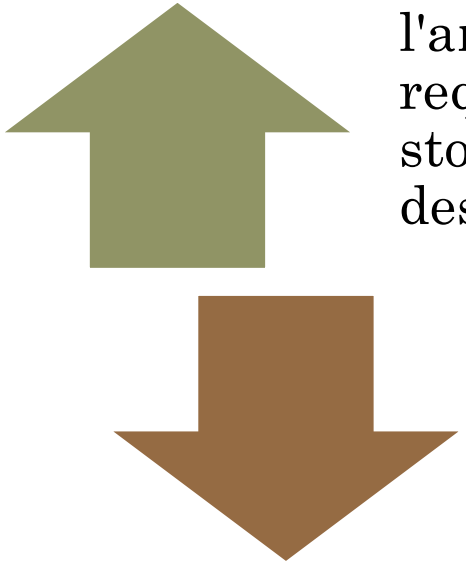


Schéma snowflake

- Le schéma snowflake (**en flocon**) est dérivé du schéma en étoile où les tables de dimensions sont **normalisées**
- La table des faits reste inchangée
- Les dimensions sont **décomposées** selon sa (ou ses) hiérarchie(s)



l'amélioration des performances de requête en raison de la réduction du stockage sur disque et de l'assemblage des tables

des efforts de maintenance supplémentaires nécessaires en raison de l'augmentation du nombre de tables de recherche.

Schéma snowflake

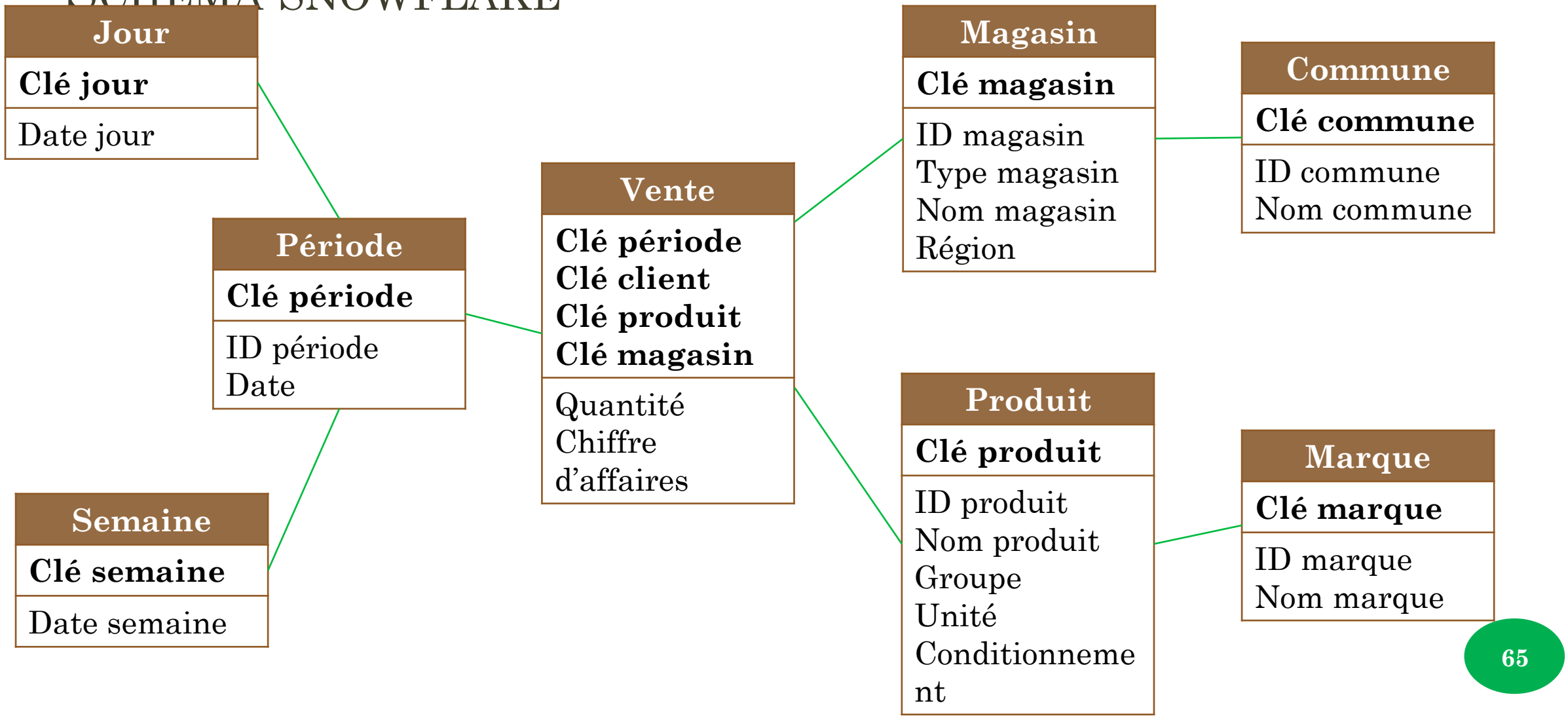


Schéma snowflake

Modèle en flocons de neige

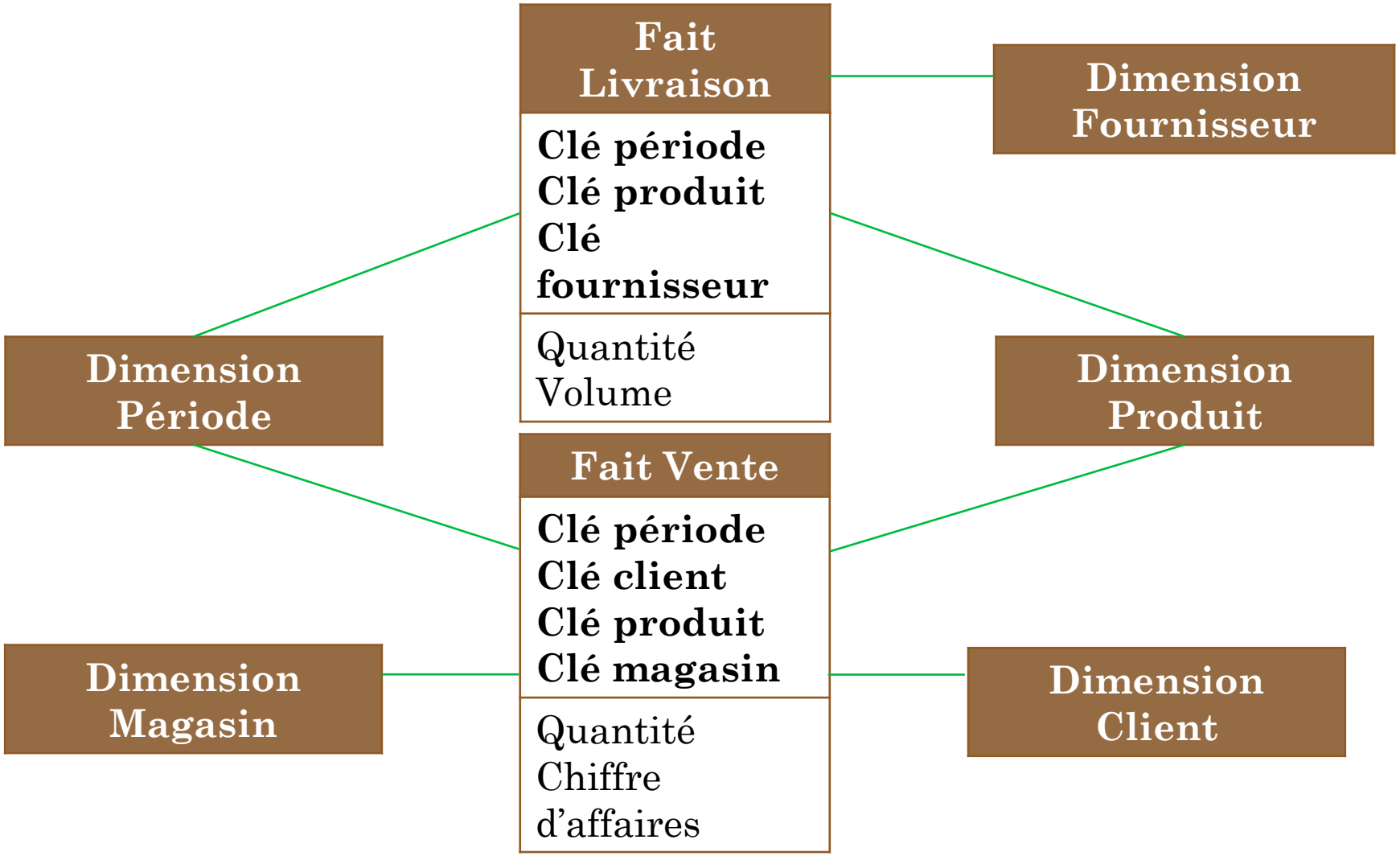
=

Modèle en étoile + normalisation des dimensions

Schéma en constellation

- fusionne plusieurs **modèles en étoile** qui utilisent des dimensions communes
- contient plusieurs **table de faits** et des dimensions communes ou pas

Schéma en constellation



EXERCICE D'APPLICATION :ETUDE DE CAS

Une entreprise emploie plus 50000 salariés répartis dans plusieurs départements chaque département contient plusieurs services. Pour Booster la productivité des salariés, la direction Ressources humaines veut octroyer une prime de fin d'année aux commerciaux qui ont pu générer la maximum du chiffre d'affaire ou qui ont pu réaliser le maximum de vente des produits. En général les responsables ont besoin de voir l'évolution des ventes par périodes et par villes.

Pour répondre à ce besoin la direction du système informatique doit construire un système décisionnel capable de générer des rapports ad-hoc ou préconfigurés, en se basant sur les données des factures qui sont stockées sur une base de données Oracle et mais aussi sur un ERP qui facilite la gestion de la relation client et des fichiers Excel contenant des détails des départements.

1. **Réaliser l'architecture du Système décision en détaillant le fonctionnement de chaque couche.**
2. **Réaliser le schéma conceptuel de l'Entrepôt de donnée permettant d'analyser la performance des employés selon les axes adéquats.**