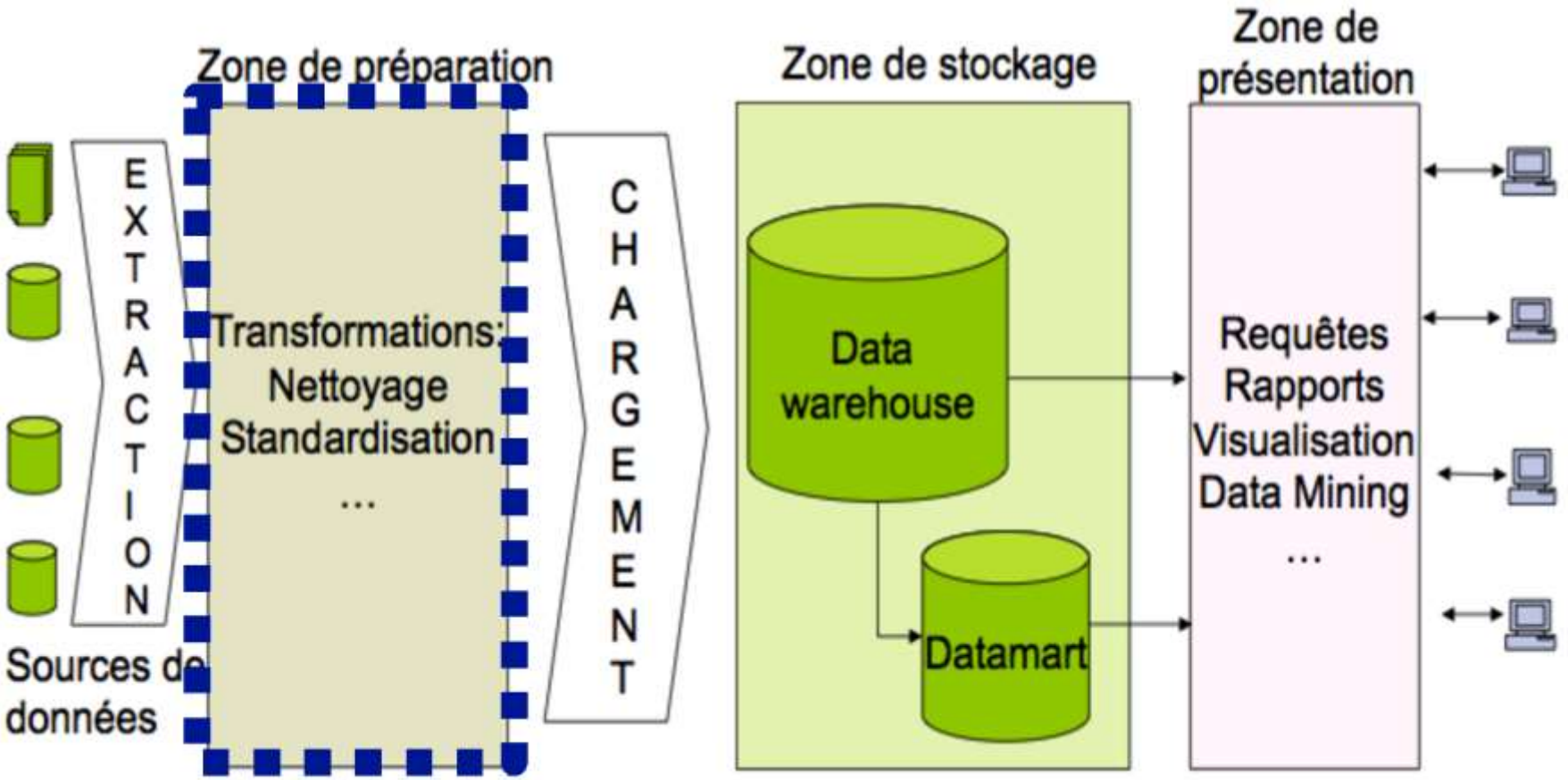
Chapitre 3 : Intégration de données et processus ETL

Intégration de données

- Place de l'intégration de données dans le processus d'entreposage



Intégration de données

- Principe d'intégration de données :
  - Sources hétérogènes → homogénéiser
  - Anomalies, Erreurs, Valeurs Manquantes → Corriger, Compléter
  - Processus continu :
    - Chargement initial
    - Rafraichissement périodique
  - Intégration logique ou physique :
    - Physique : architecture réelle
    - Logique : architecture virtuelle

- Méta données
  - Dans les BD : données sur les données (structure des tables, informations sur les colonnes, etc)
  - Dans les ED : données sur les composants d'ED, sur les sources, sur le processus, etc.

- Rôles des méta données
  - Permettre d'automatiser (certains) composants d'entreposage (dont l'ETL)
  - Assurer les liens entre sources et ED

- Type des Méta données
  - Sources de données : noms, liens, propriétés...
  - Modèle d'ED : serveurs, bases de données, tables
  - Mapping source-ED : liens, transformations,
  - Outils d'intégration (ETL) : nom, période de rafraichissement...
  - Architecture de l'ED (ED, Data Marts, ...)
  - Règles et stratégies : indicateurs de performances, formules de calcul
  - Règles de sécurité: qui accède à quoi?

Intégration de données

- Méta données
  - Exemples: méta données au niveau table

Table ED	Colonne	Table Source	Colonne source	Observation
Produit	Réference	produit1	Code produit	
		produit2	Num Produit	
	Nom produit	Produit1	Nom	
	Prix unitaire	Produit2	Prix	Avec décimales dans l'ED

Intégration de données

- Méta données
  - Exemples : méta donnée d'une colonne de table d'ED

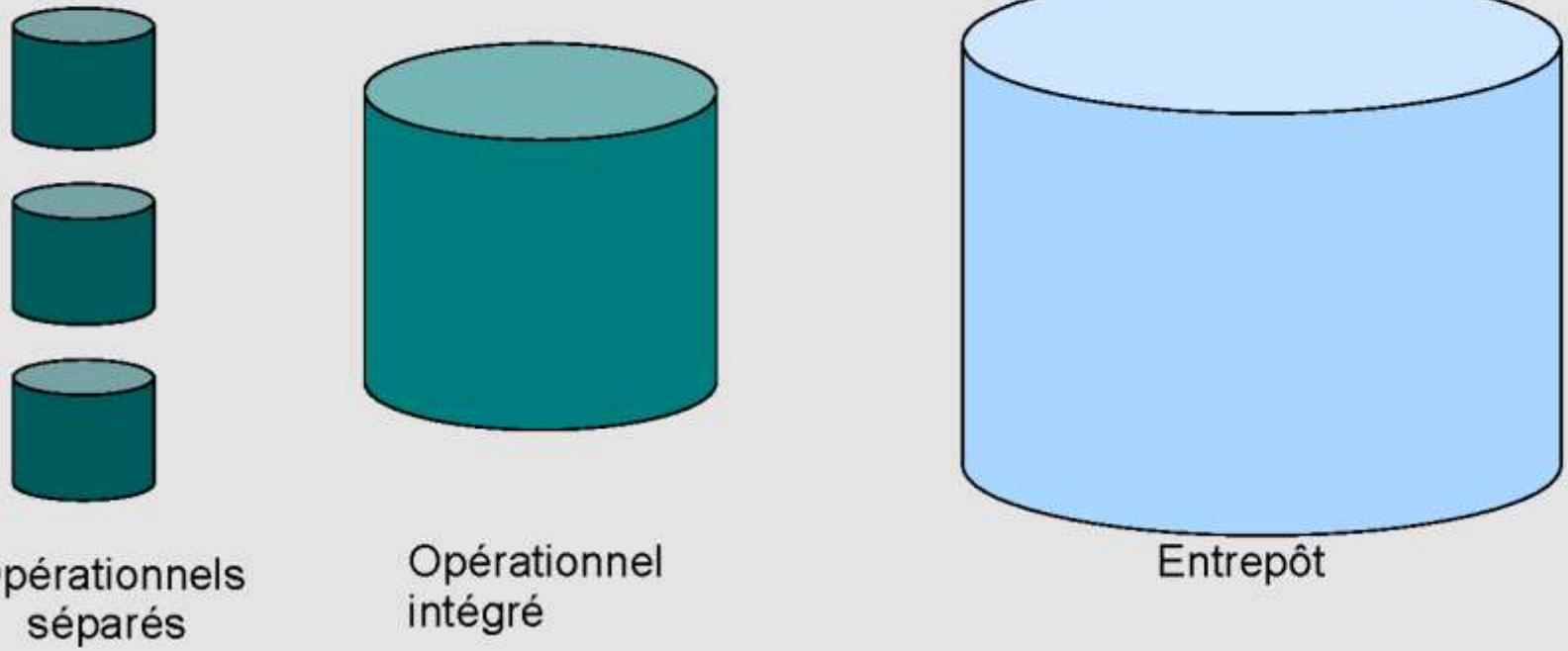
Nom de colonne	Réf_produit
Caption	Référence
Type de données	Number
Type d'index	B Tree
Clé ?	Oui
Format	NNNNNNN
Description	La référence d'un produit

- Méta données
  - Exemples : méta donnée d'une table d'ED

Nom de table	D_Produit
Caption	Produit
Description	La tables de produits
Option de clés	Artificielle, avec vesionnement
Source	
...	

Intégration de données

- Schéma d'intégration
  - Dans la pratique, l'intégration des données passe par un schéma d'intégration
  - La base de données d'intégration (Operational Data Source) est différente des sources et de l'entrepôt



- Schéma d'intégration
  - La base de données d'intégration permet de stocker les données d'intégration avant de les charger.
  - Exemple:
    - Source 1: vente_nord (num_vente, montant)
    - Source 2: vente_sud (numero_vente, volume)
    - Entrepot: ventes (no_ventes: montant, id_région)
      region (id_region, designation_region)

- Extraction

  - Extraire les données (nouvelles ou changées) à partir des sources.

  - Utilise les méta-données (liens entre les tables de l'entrepôt et les tables sources)

  - Deux phases d'extraction :

    - Découverte de données
    - Correction d'anomalies

- Extraction

  - Découverte de données

    - Documentation des systèmes sources si elle existe.

    - Découvrir les point origine des données (source dans laquelle la données est enregistrée la première fois).

- Extraction
  - Détection d'anomalies

- *Valeurs nulles* : surtout dans les clés étrangères
- *Mauvais types de données* : dates dans des champs non dates, numérique dans des champs non numériques, etc...
- *Incohérence* → différents types de données, différentes longueurs, différentes contraintes...

- Transformation
  - Correction des anomalies

  - Valeurs nulles : remplir, assimiler à des valeurs (ex : **NVL**), se baser sur les probabilités, ...
  - Mauvais types de données : choisir le type de données le plus adéquat...
  - Incohérence → unifier les types de données, les longueurs, les contraintes...

Intégration de données

- Chargement
  - Chargement initial
  - Chargement incrémental
- Chargement initial
  - Une seule fois
  - Consomme du temps (grand volume de données à charger)
  - Désactiver les contraintes d'intégrité pour
    - Accélérer l'alimentation
    - Paralléliser l'alimentation
  - Générer les clés artificielles

Intégration de données

- Chargement incrémental
  - Chaque fois que le changement se produit

ou

  - À des intervalles périodiques
  - Que les données ayant changé
- Types de changement
  - Ajout d'enregistrements, suppression d'enregistrement → détection facile par comparaison
  - Modification d'enregistrement (valeur d'attributs) → nécessite de détecter les changements

- Chargement incrémental
  - Détection des changements

  - Comparaison colonne par colonne → couteuse
  - Audit des colonnes : pour chaque colonnes, date de dernières modification
  - CRC : cyclic redundancy check de chaque enregistrement → moins coûteux

- Chargement incrémental
  - Détection des changements
    - Quand détecter les changements

- Méthode PUSH : à chaque fois que le changement se produit, on répercute les changement → utilisation de déclencheurs (triggers)

- Méthode PULL : on interroge à des intervalles périodiques les fichiers log des CRC pour détecter les changements et on les répercute dans l'ED (dans des périodes d'inactivité des sources et de l'ED).