Partie 1

Data Warehouse

Entrepôts de données

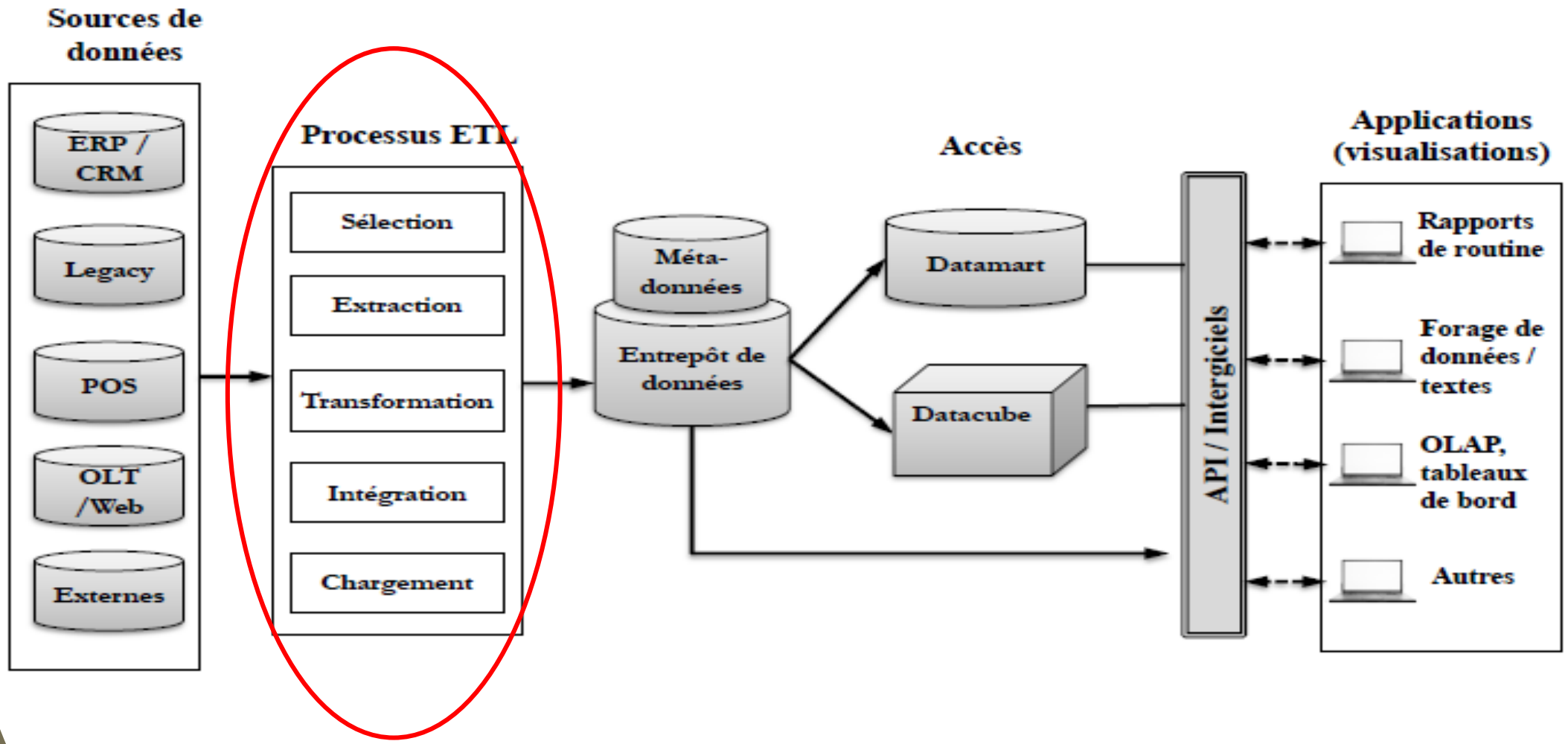
Chapitre 4 : Le processus ETL

Dr H. EL BOUHISSI Epse BRAHAMI

Plan

- Introduction
- Le processus ETL

Introduction



Introduction

Un entrepôt de données doit être chargé régulièrement afin qu'il puisse servir à faciliter l'analyse. Pour ce faire, les données d'un ou plusieurs systèmes opérationnels doivent être extraites et copiées dans l'entrepôt de données. Le défi dans les environnements d'entrepôt de données consiste à intégrer, réorganiser et consolider de grands volumes de données sur de nombreux systèmes.

Le processus d'extraction des données des systèmes sources et de leur transfert dans l'entrepôt de données est communément appelé ETL.

Le processus ETL est un processus itératif qui se répète dés que de nouvelles données sont ajoutées à l'entrepôt. Ce processus garantit que les données de l'entrepôt de données sont exactes, complètes et à jour.

Processus ETL : Extract – Transforme - Load

L'ETL (extraction, transformation et chargement) est le processus qui consiste à combiner les données provenant de plusieurs sources dans un grand référentiel central appelé entrepôt des données.

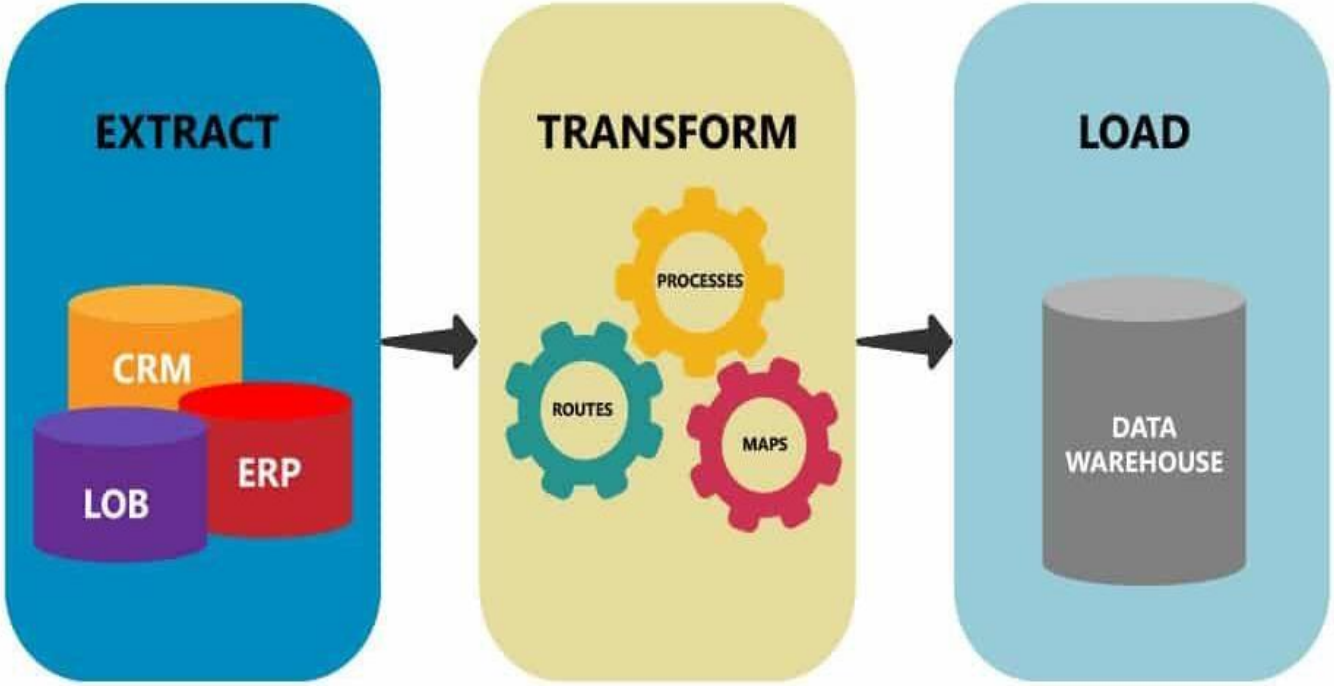
L'ETL utilise un ensemble de règles opérationnelles pour nettoyer et organiser les données brutes et les préparer pour le stockage, l'analytique des données et le machine Learning (ML).

L'analytique des données permet de répondre à des besoins spécifiques en matière d'informatique décisionnelle (par exemple, prévoir le résultat de décisions commerciales, générer des rapports et des tableaux de bord, réduire l'inefficacité opérationnelle, etc.).

Processus ETL : Extract – Transforme - Load

- Sélection des données sources.
- Extraction des données.
- Nettoyage et Transformation.
- Chargement

Outils : Talend, Oracle Warehouse Builder, etc.



Etapes 1 et 2 : Jusqu'a 80 % du temps de développement d'un entrepôt

Processus ETL : Extract – Transforme - Load

Extract :

- Extraire les données des sources hétérogènes
- Identifier les données sources utiles.
- Déterminer les données qui ont changé.

Transform :

- Consolider les données/ Données redondantes, manquantes, incohérentes, etc.
- Découpage, fusion, conversion, agrégation, etc.
- Les conflits entre les modèles, les schémas et les données sont résolus durant cette phase.

Load :

- Charger les données intégrées dans l'entrepôt dans la BD cible. La BD cible est souvent implantée avec un SGBD relationnel-objet.
- Mode différé (batch) ou quasi temps-réel.

Sources 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.).

Systèmes legacy

Matériels et logiciels obsolètes mais difficilement remplaçables.

Point of sale (POS)

Matériels et logiciels utilisés dans les caisses de sorties d'un magasin.

Externes

Données concurrentielles achetées, données démographiques.

Sources de données

- Sources diverses et disparates (ex: BD, fichier texte, etc.).
- Sources sur différentes plateformes et OS.
- Applications legacy utilisant des technologies obsolètes.
- Historique de changement non-préservé dans les sources.
- Qualité de données douteuse et changeante dans le temps.
- Structure des systèmes sources changeante dans le temps.
- Incohérence entre les différentes sources.
- Données dans un format difficilement interprétable ou ambigu.

Extraction des données

Première étape du processus ETL consiste à extraire ou extraire les données de toutes les sources pertinentes et à les compiler. Cette étape comprendra la préparation nécessaire à la réalisation de l'intégration des données.

Les étapes :

- Compiler des données à partir de sources pertinentes.
- Organiser les données pour les rendre cohérentes.

Mode différé : Extrait tous les changements survenus durant une période donnée (ex: heure, jour, semaine, mois).

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

Transform : Transformation des données

- Convertir les données selon les besoins de l'entreprise.
- Reformatez les données converties dans un format standard pour la compatibilité.
- Nettoyer les données non pertinentes des ensembles de données.
  - Trier et filtrer les données
  - Effacer les informations en double
  - Traduire si nécessaire
- Révision de format: Changer le type ou la longueur de champs individuels, etc.
- Décodage de champs: ['homme', 'femme'] vs ['M', 'F'] vs [1,2] , etc.
- Pré-calcul des valeurs dérivées: profit calculé à partir de ventes et coûts , etc.
- Découpage de champs complexes: extraire les valeurs prénom, secondPrénom et nomFamille à partir d'une seule chaîne de caractères nomComplet , etc.
- Pré-calcul des agrégations: ventes par produit par semaine par région , etc.
- Déduplication : Plusieurs enregistrements pour un même client , etc.

Load : Chargement des données

La dernière étape du processus ETL consiste à charger les données transformées dans la cible le Data Warehouse.

Deux méthodes principales pour charger les données dans un entrepôt :

- Chargement complet : implique un déchargement complet des données qui a lieu la première fois que la source est chargée dans l'entrepôt.
- Chargement incrémentale : a lieu à intervalles réguliers. Ces intervalles peuvent être des incréments de flux (meilleurs pour de plus petits volumes de données) ou des incréments de lots (meilleurs pour de plus grands volumes de données).

Load : Chargement des données

Stockage des données nettoyées et préparées dans la BD opérationnelle.

Une opération risquant d'être assez longue.

Mais il est nécessaire de définir et mettre en place :

- Des stratégies pour assurer de bonnes conditions à sa réalisation :
- Une politique de rafraichissement.
- Faire les chargements en lot dans une période creuse (entrepôt de données non utilisé).
- Considérer la bande passante requise pour le chargement.
- Avoir un plan pour évaluer la qualité des données chargées dans l'entrepôt.
- Commencer par charger les données des tables de dimension.
- Désactiver les indexes et clés étrangères lors du chargement.

Quelques outils ETL

MarkLogic: https://www.marklogic.com/product/getting-started/

Oracle: https://www.oracle.com/index.html

Amazon RedShift: https://aws.amazon.com/redshift/?nc2=h_m1

D'autres outils à explorer ……

Fin

Dr. H. EL BOUHISSI BRAHAMI