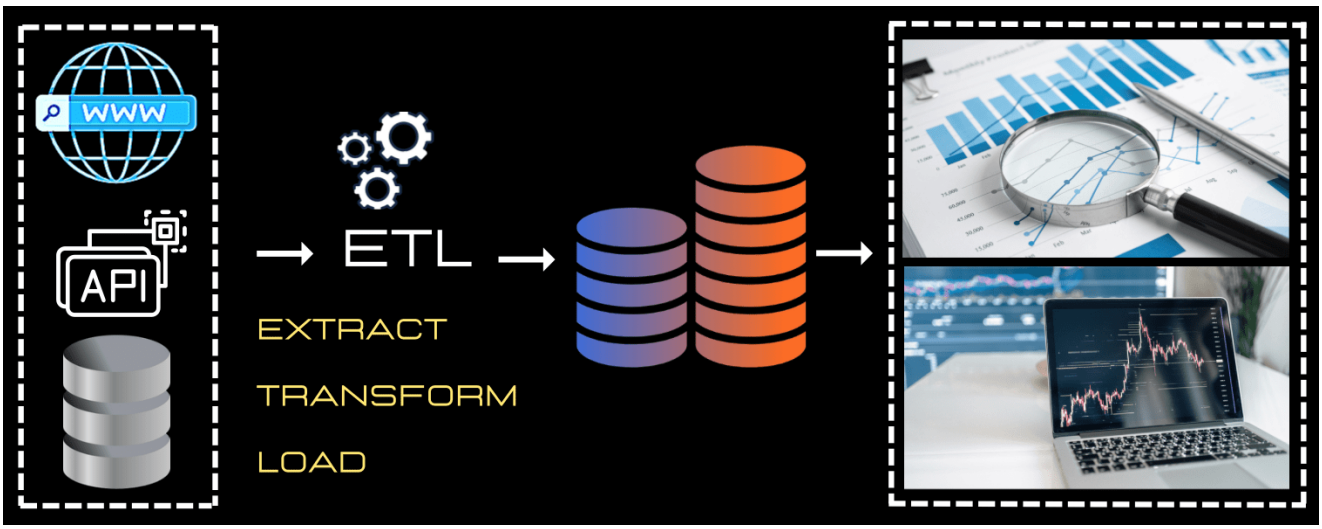
DIT Master I - IA - 2023 / 2024

Introduction à la mise en place de pipelines ETL (Extract Transform Load)

- Fiche de TP - Création de PIPELINES ETL, prétraitement et visualisation de données



---

1. Récupérez les 12 fichiers à partir de l'URL suivante: https://github.com/ousmanhamit/sales_datasets.git

2. Création d'un environnement virtuel et installation des dépendances en utilisant les commandes ci-dessous:

- python3.12 -m venv .venv && source ~/.venv/bin/activate
- pip install -r requirements.txt

TRANSFORMATION DE DONNEES

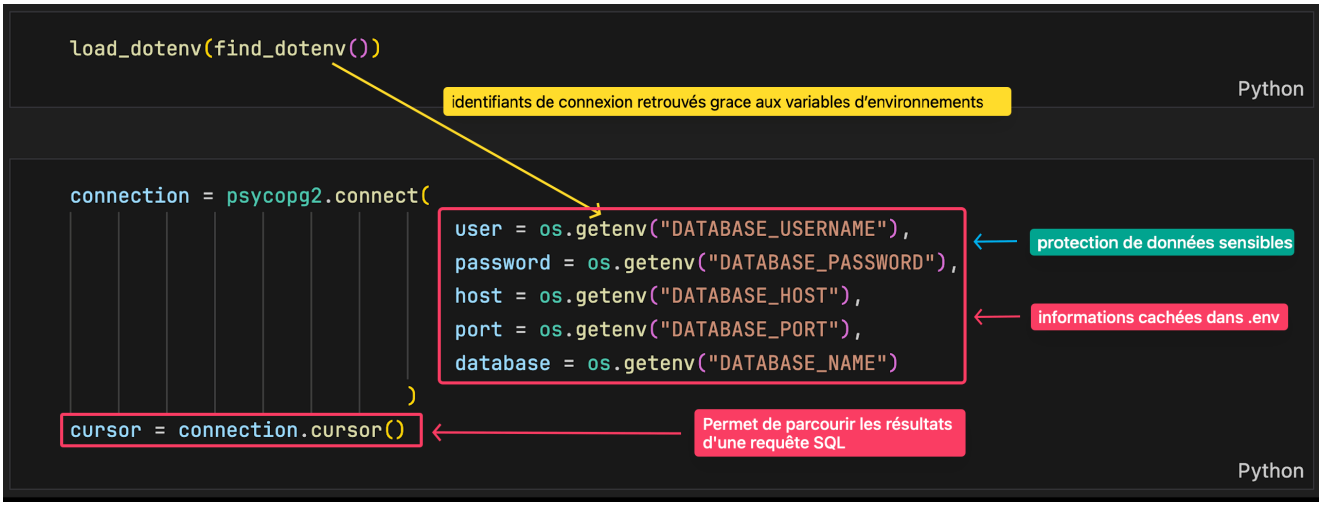
- Écrire une fonction qui permet de fusionner les 12 fichiers en un seul document
- Renommer toutes les colonnes qui comportent un espace dans leur libellé (par exemple, la colonne 'Quantity Ordered' serait renommée 'QuantityOrdered', etc.).
- Transformer en entier (int) ou en nombre décimal (float) les colonnes numériques actuellement exprimées sous forme de chaînes de caractères.
- Convertir la colonne 'Order Date' to datetime

- Ajouter une colonne "Hour" en se basant sur la colonne "OrderDate" et l'insérer immédiatement à la suite de cette dernière.
- Identifier et supprimer les données dupliquées
- Identifier et traiter les valeurs manquantes, c'est-à-dire les remplacer, les retrouver ou les supprimer.
- Identifier et traiter les valeurs aberrantes, c'est-à-dire les remplacer, les retrouver ou les supprimer.
- Afficher la liste des 10 produits les plus vendus
- Afficher la liste des 10 produits ayant enregistré le moins de ventes
- Afficher les produits les plus vendus pendant le mois de décembre
- Determiner la distribution moyenne de prix selon chaque categorie de produit
- Visualiser le total de ventes par mois
- Combinez toutes les étapes du prétraitement dans un pipeline (scikit-learn) afin de simplifier le processus de modélisation et d'obtenir des résultats plus cohérents

---

LOAD DATA

- Se connecter à une base de données Postgres (il faut l'avoir installée au préalable sur sa machine, créer une base de données nommée "mydb" et un utilisateur "myuser" avec ses identifiants de connexion), utilisez des variables d'environnement pour protéger vos informations d'identification, comme illustré ci-dessous



- Écrire une fonction qui permet de charger (load) les données que vous venez de nettoyer à la base de données mydb dans une table nommée "dit_etl_pipeline"