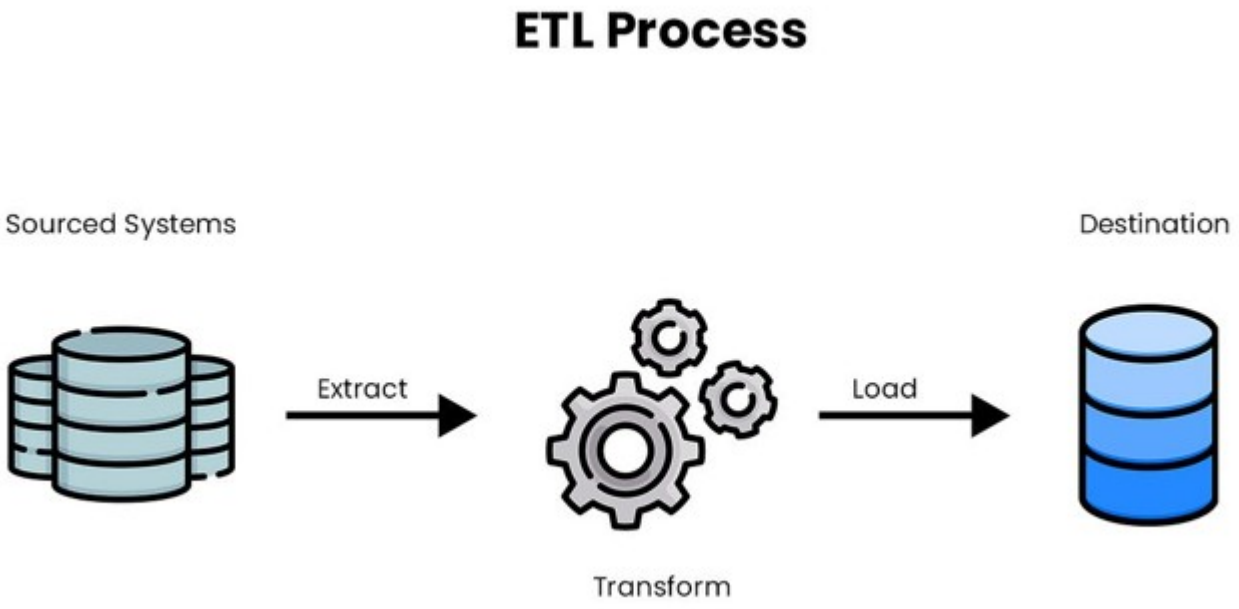
Automatisatoin d’un Pipeline ETL

Pipeline ETL



Web scraping

- Collecte automatique de données
- Suivi de l’actualité
- Automatiser la navigation web

Critères de comparaison entre les bibliothèques

Simplicité du code
Performance : capacité à traiter des gros volumes de données
Support Javascript
Gestion des erreurs

Processus de web scraping

1) Identifier l’emplacement exacte de la source de données
2) Envoyer une requete pour collecter le contenu brut
3) Stocker les données extraites

BeautifulSoup

Extraction des données statiques HTML et XML
Nettoyage de contenu HTML : conservation du texte propre pour la préparation des données à l’analyse du texte

Scrapy

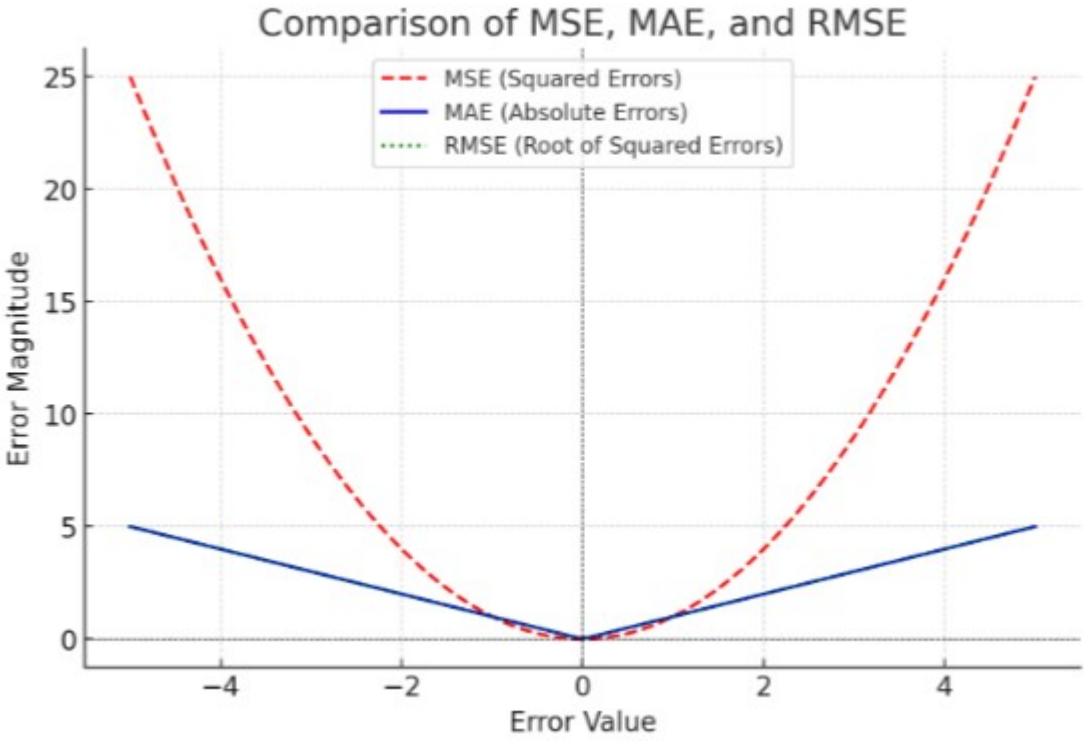
Framework Python open source qui permet :

- collecte de bases de données depuis des sites web
- automatisation de la navigation
- gestion de plusieurs pages

Selenium

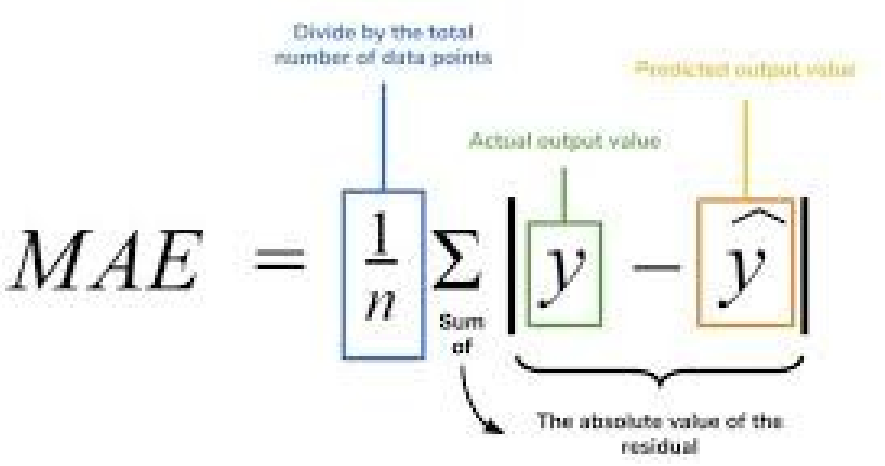
-Interaction avec des pages dynamiques automatisation complète
-Remplissage et soumission de formulaires
-Téléchargement de fichiers

Métriques mae / mse



MAE : Mean Absolute Error

La moyenne des distances absolues (distance de Manhattan) entre les valeurs réelles yi et les valeurs prédites ỹi



MSE : Mean squared error

La différence moyenne au carré entre les valeurs estimée ỹi et les valeur réelles yi
chaque erreur est élevée au carré

$$MSE = \frac{1}{n}\sum \underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

cron

- automatiser l’exécution régulière du pipeline ETL à une frequence défénie.