Somaire

Contexte et objectifs

Le processus ETL mis en place a pour objectif d'importer, transformer, et insérer des données relatives à deux maladies, le Coronavirus et la variole du singe, dans une base de données MySQL. Ce système permet de centraliser les données provenant de différentes sources dans un format cohérent, en vue d'une analyse ultérieure.

Les principales étapes du processus sont :

- Extraction des données depuis des fichiers CSV.
- Transformation des données pour assurer leur cohérence.
- Chargement des données dans une base de données relationnelle.

Architecture du processus

Extraction des données

Les données sont extraites depuis trois fichiers CSV :

- **worldometer_coronavirus_daily_data.csv**
- **countries_and_continents.csv**
- **owid-monkeypox-data.csv**

Après extraction, les valeurs manquantes sont remplacées par 0 pour éviter des erreurs lors du traitement ultérieur.

Transformation des données

Nettoyage et harmonisation

1. **Noms des pays** : Les noms des pays sont harmonisés via une fonction de renommage pour assurer la compatibilité entre les différentes sources.
    - Exemple : "United States" est transformé en "USA".
2. **Filtrage des données** : Les entrées contenant des codes ISO spécifiques (à l'aide de "OWID") dans le fichier monkeypox sont supprimées.

Structure des données

Trois tables sont créées pour structurer les données :

1. **Disease** : Contient les différentes maladies présentes dans les csv.
2. **Localization** : Contiens les pays et continents de jeux de données.
3. **ReportCase** : Décrit les cas signalés pour chaque maladie, par localisation et par date avec des chiffres.

Les fichiers sont enrichis grâce à des jointures avec les données de localisation (pays et continents) pour créer une base uniforme.

Justification des choix techniques

1. **Python et Pandas** :
- Python est choisi pour sa simplicité et sa richesse en bibliothèques de manipulation de données.
- Pandas facilite le traitement des CSV et offre des opérations performantes pour les jointures, filtrages et transformations.
2. **Nettoyage des Données** :
- La fonction de renommage assure la cohérence entre les différents fichiers, évitant des erreurs lors des jointures.
3. **Base de Données MySQL** :
- MySQL est un système de gestion de bases de données relationnel performant et largement utilisé. Son utilisation garantit la persistance et la structuration des données.
4. **Approche Modulaire** :
- Les tables (Disease, Localization, ReportCase) sont conçues pour être extensibles, permettant d'ajouter d'autres maladies ou localisations à l'avenir.