

Documentation : Collecte et Nettoyage des Données

Collecte des Données

Nous avons collecté des jeux de données portant sur :

- **Le COVID-19** : données épidémiologiques et statistiques associées.
- **La variole du singe** : données épidémiologiques et statistiques associées.
- **Les pays et fédérations du monde** : noms et continents.

Ces données ont été obtenues depuis le site **Kaggle**, qui offre une variété de ressources fiables et de qualité pour l'analyse de données.

Nettoyage des Données

Le processus de nettoyage des données a été entièrement automatisé à l'aide d'un **ETL** (Extract, Transform, Load) que nous avons développé en Python. Les principales étapes du nettoyage sont :

1. **Chargement des données brutes** via la bibliothèque **Pandas**.
2. **Transformation des données** :
 - Suppression des valeurs manquantes et des doublons.
 - Normalisation des formats (dates, chaînes de caractères, etc.).
 - Harmonisation des unités de mesure.
 - Gestion des anomalies et des valeurs aberrantes.
3. **Validation des données** pour garantir leur cohérence et leur intégrité.

Une fois nettoyées, les données sont intégrées dans une **base de données SQL**, facilitant ainsi leur exploitation pour l'api.