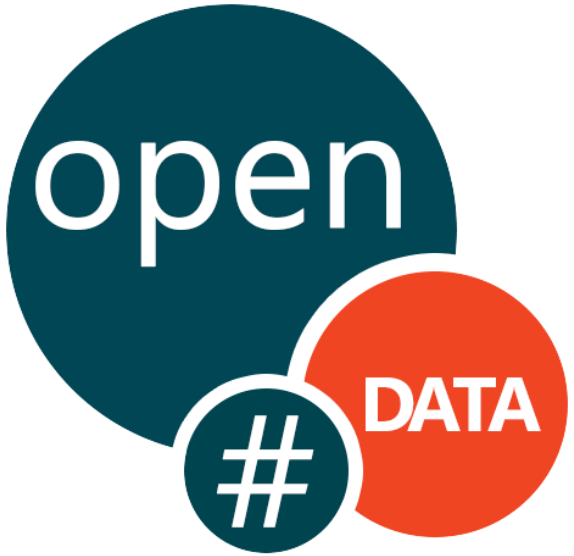
Projet open data

Exploitation du répertoire SIREN



Objectif :

Objectif :

- Valoriser les données issue de l’open data et gérer de grande quantité de données.
- Exploiter des données brut, le répertoir des entreprises, dans le but de produire des statistiques sur le tissu économique d’une ville, d’un département ou d’une région.

Sommaire

1. Prendre connaissance de la documentation et créer les tables dans la quel on va charger les données des établissements.
2. Connecter python à MySQL
3. Alimenter une base de données depuis un fichier csv hébergé sur le web.
4. Accélérer les requêtes sur cette base avec l'indexation.
5. Créer une table, à partir des cette base, contenant des informations pertinente sur la ville de Marseille et ces arrondissements.

La documentation

Nous allons travailler sur les fichiers « **Fichier StockEtablissement du 01 janvier 2020** » et « **Fichier StockEtablissement du 01 janvier 2020** » présent à l'adresse

> https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret

Le contenu de ces tables est détailler dans la documentation se trouvant à la même adresse.

En se basant sur cette dernière, créer les tables dans MySQL pouvant recevoir les données contenu dans ces fichiers.

Connecter python à MySQL

a) Installer mysqlclient et pymysql avec l'utilitaire d'installation d'anaconda « conda ».
b) Importer Sqlalchemy (déjà installé avec anaconda).
c) Utiliser create_engine pour établir une connection.
d) Tester la connection en important les données de la table jeux_video.

Alimenter une base de données

Créer une fonction python qui va importer des données depuis une source se trouvant dur le web dans une base de données.

- Utiliser read_csv de pandas pour lire les données
  - Astuce : on peut spécifier une url en entré
  - Astuce : on peut spécifier un format de compression si le fichier est compressé.
  - Le fichier étant volumineux utiliser l'instruction « chunksize » pour traiter le fichier par partie.
  - Penser à spécifier les champs ayant un format date.
- Utiler to_sql de pandas pour charger les données dans la base de données
- La fonction doit renvoyer le temps total d'exécution au format h:mm:ss.

Accélérer les requêtes

Faire une veille sur l’indexation.