Spécifications data réseau vélo

Contenu des fichiers

Dossier Cyclistes

cycliste_2.csv à cycliste_51.csv

Liste les mouvements de chaque cycliste à chaque minute pendant 1 mois.

Colonnes:

- id : numéro d'identification du cycliste
- timestamp : date et heure de l'enregistrement
- sur_velo : le cycliste est-il actuellement sur un vélo
- velo : numéro d'identification du vélo
- vitesse : vitesse du cycliste (à vélo si sur_velo = True, à pied si sur_velo = False)
- destination_finale : coordonnées de la destination du cycliste (travail ou home)

Dossier Prestataires

Prestataire_1.csv

Donne des informations sur la taille du réseau du prestataire. Pas très utile pour l'analyse.

Colonnes:

- id = numéro d'identification du prestataire
- largeur = largeur du réseau (en km par exemple)
- hauteur = hauteur du réseau (en km par exemple)

Dossier Reparateurs

reparateur_1.csv à reparateur_3.csv

Liste des mouvements de chaque réparateur du réseau à chaque minute pendant 1 mois.

Colonnes:

- timestamp : date et heure de l'enregistrement
- id : numéro d'identification du réparateur
- trajet : coordonnées point de départ => coordonnées point d'arrivée
- vitesse : vitesse du réparateur
- position : coordonnées de la position au moment de l'enregistrement
- etait_sur_station : le réparateur est-il en train de réparer une station?
- station_id : numéro d'identification de la station (si etait_sur_station = True)

Dossier Stations

log_stations_1.csv à log_stations_25.csv

Liste les logs sur chaque station des cyclistes qui prennent ou rendent un vélo.

Colonnes:

- timestamp : Date et heure de l'enregistrement
- station_id : numéro d'identification de la station
- cycliste_id : numéro d'identification du cycliste
- velo_id : numéro d'identification du vélo
- velo_performance : état du vélo (valeur entre 0 et 1, 0 = mauvais état, 1 = bon état)
- action : prendre ou déposer un vélo
- position : colonne fausse, À SUPPRIMER ENTIÈREMENT
- position_station : coordonnées de la station

Dossier Villes

ville_1.csv

Informations sur les cyclistes de la ville

Colonnes:

- id : numéro d'identification du cycliste (entre 2 et 51)
- vitesse_a_pied : vitesse moyenne du cycliste à pied
- vitesse_a_velo : vitesse moyenne du cycliste à vélo
- home : coordonnées du domicile du cycliste
- travail : coordonnées du lieu de travail du cycliste
- sportif : le cycliste est-il un sportif?
- casseur : le cycliste est-il un casseur?
- statut : métier du cycliste
- salaire : salaire annuel du cycliste
- sexe
- age
- sportivité : score de sportivité
- velo_perf_minimale : vitesse minimale enregistrée à vélo

Schéma des liaisons entre les tables

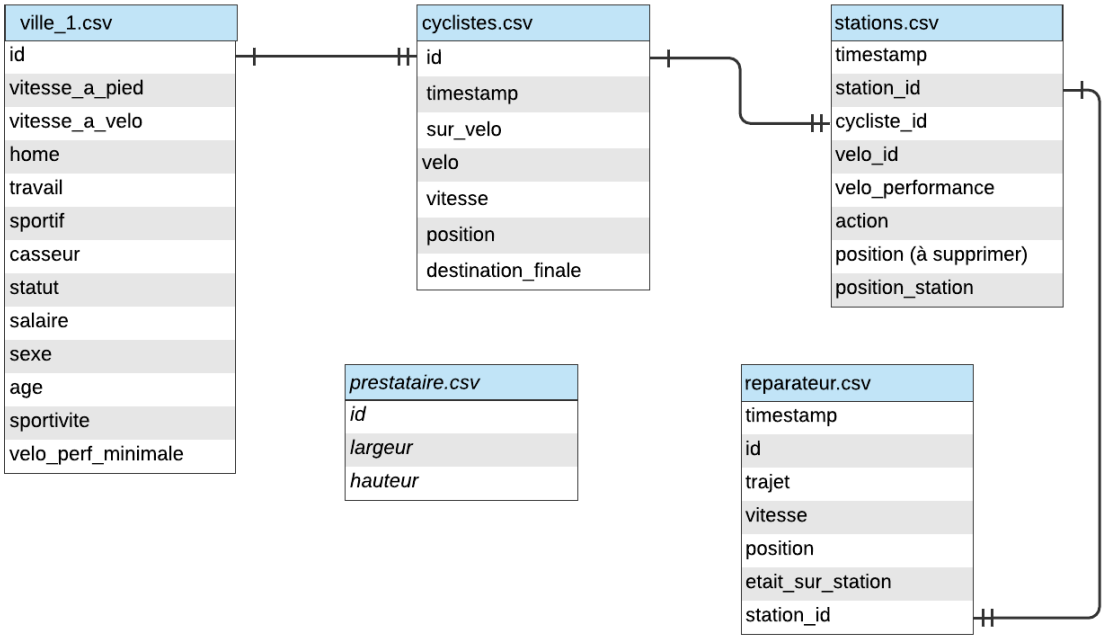
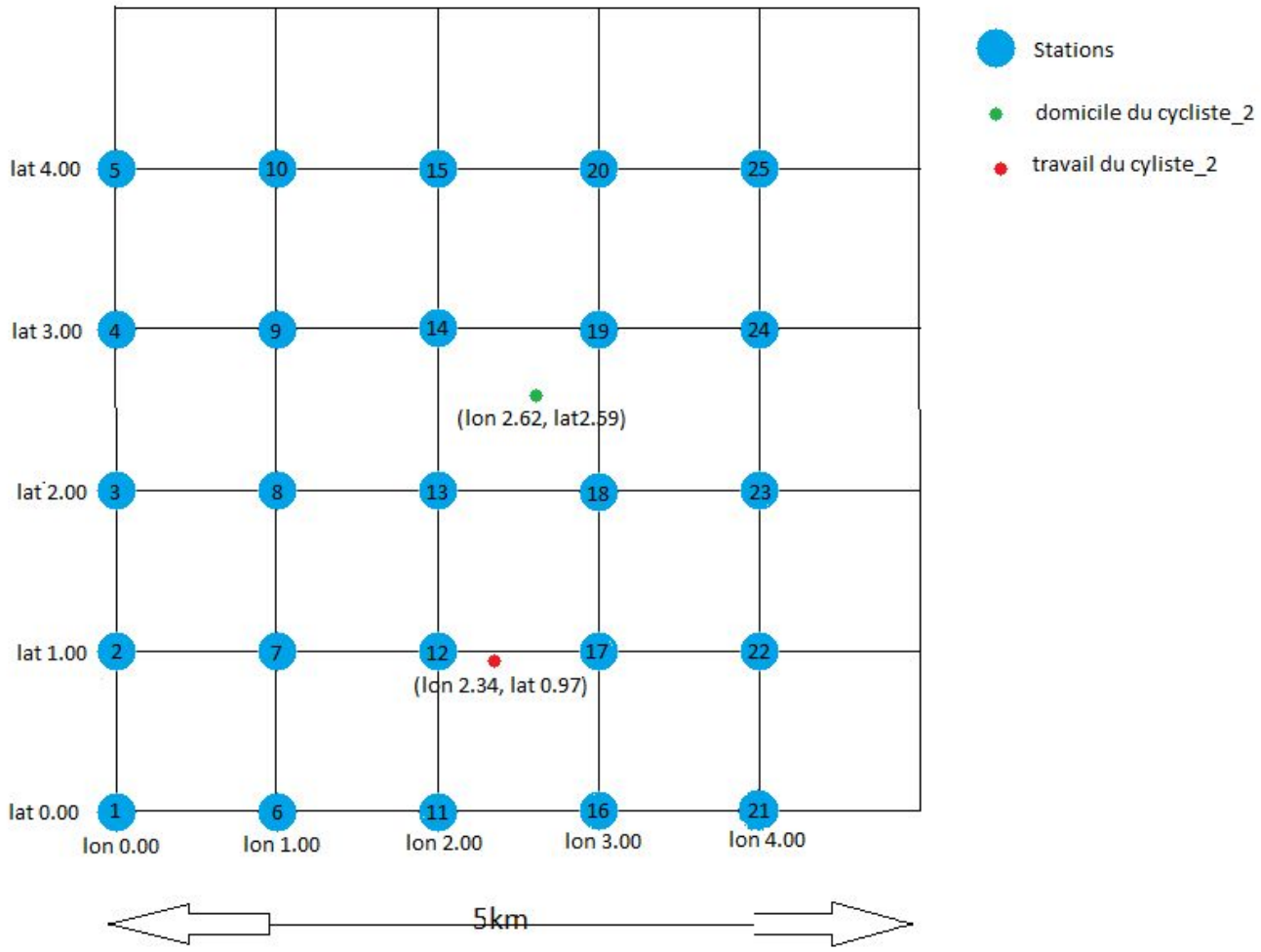


Schéma du réseau

Rappelons ici que cette donnée est une simulation d'un réseau.
On voit dans le fichier prestataire_1.csv que la largeur du réseau = 5 et que la longueur = 5 également.
Admettons que l'unité est le km (c'est probablement plus une unité américaine, le miles, mais pour simplifier les calculs on utilisera l'unité européenne), on a donc une map carrée de 5km sur 5km.
On voit aussi dans les fichiers log_stations_*.csv les coordonnées des stations, qui on le rappelle ne sont pas de vraies latitudes et longitudes mais les coordonnées x et y sur la map carrée de 5 sur 5.
Ces informations nous permettent de schématiser le réseau comme ceci:



Les différentes pipelines possibles

Batch :

Vous pouvez séparer les données cyclistes, stations et réparateurs par semaine par exemple, 1 mois de données cela donnera 4 batchs.
Utilisez les données du premier batch + fichier ville_1 pour construire et initialiser votre pipeline. Puis envoyez les 3 batch suivants à intervalle régulier.

Streaming :

Traiter le fichier ville_1 en batch, et utiliser à l'aide de kafka envoyer les lignes une par une à chaque minute pour simuler de la donnée en temps réel.