

# **Documentation technique**

## [Datapipeline comparaison officiel vs gas stations reporting](#)

Réalisé par Gaëtan Corin

### **Introduction :**

Cette documentation technique est à destination des Data-Engineers et Data-analyste ayant des capacités sur le fonctionnement d'un ETL.  
Il n'est pas nécessaire d'être un expert dans le domaine des flux pétroliers pour comprendre le fonctionnement et la réalisation du projet "Datapipeline\_comparaison\_officiel\_vs\_gas\_stations\_reporting" .

### **Résumé du projet :**

L'objectif est de récupérer les données officielles gouvernementales pour les 6 plus grands types d'essences français [Gazole](#), [SP95](#), [SP95-E10](#), [SP98](#), [E85](#), [GPLc](#).  
Puis de récupérer les données rapportées par les stations essences françaises, travailler ces données pour ce rapprocher le plus possible des données officielles gouvernementales, puis de les comparer.

### **Les données :**

Les données sources proviennent:

- pour les données officielles gouvernementales, il s'agit du site <https://www.ecologie.gouv.fr/politiques-publiques/prix-produits-petroliers>  
Il s'agit de données Open Data disponibles pour tous les utilisateurs.  
Le formulaire d'extraction étant un formulaire dynamique ajax, et le l'url final d'extraction des données contenant un UUID réinitialisable tous les 15 jours, il est nécessaire de scrapper le site pour extraire le lien d'import des données.  
La donnée est mise à jour de manière hebdomadaire, et la quantité de données est raisonnable a 2 100 enregistrements entre 1985 et 2025
- pour les relevés des stations essences, il s'agit du site <https://www.prix-carburants.gouv.fr/rubrique/opendata/>  
qui fournit les informations en Open Data pour tous les utilisateurs.  
Il s'agit d'un autre site gouvernemental, mais il s'agit cette fois ci de données brut de stations essences, non nettoyées, ou chaque modifications de prix réalisé par une station essence au cours de la journée en france crée un nouvel enregistrement. Nous arrivons donc pour les années disponible de 2007 à 2025 a une quantité de 56 600 000 enregistrements pour 13 600 stations essences existantes ou ayant existé au total en format xml (3.86 go de données)  
Les données sont accessibles en ayant un lien url pour chaque années.

Les données sont extraites par url, transformées, puis chargées sur une base MongoDB.

Voici les différentes étapes de transformations de données:

**Pour les données de stations essences:**

(extract) Pour chaque année:

- Récupération des données zippés par l'url
- dezip des données
- transformation en dataframe pandas
- nettoyage des retour à la ligne et espaces non désirables pour les colonnes concernés
- formatage des dates en un seul format (3 formats différents sur les 18 années)
- formatage du prix en un seul format (2 formats différents sur les 18 années)
- sauvegarde en csv en local dans le serveur flask

(transform) Pour chaque année:

- Récupération du fichier csv de l'année concernée
- Définition de l'écart type et réalisation de z-score pour nettoyer les données aberrantes pour chaque jour et type d'essences
- Réduction des données en sauvegardant uniquement la dernière valeur cohérente de la journée par station et par type d'essences
- Sauvegarde en csv en local dans le serveur flask

(load) Pour chaque année:

- Récupération du fichier csv de l'année concernée
- Division de la partie des données de station essence
- Retrait des duplications en gardant uniquement les stations essences avec la dernière date connue en filtrant sur la date et l'id unique de la station essence (pour avoir les données sur les stations essences les plus à jours)
- Mise à jour ou chargement sur MongoDB des stations essences en privilégiant toujours l'information la plus récente (bdd datalake collection gas\_stations\_infos)
- Division de la partie des données des prix d'essences des stations
- Chargement des données de prix d'essence des stations en ayant la date et le type d'essence en index sur MongoDB (bdd datalake collection gas\_stations\_price\_logs\_eur) (passage de 56 millions à 39 millions de données cohérentes)

(denormalization) Pour chaque année:

- Récupération des données des prix d'essences sur MongoDB

- Rassemblement des prix d'essences pour avoir uniquement un enregistrement par jour contenant la moyenne de tous les prix des stations essences pas type d'essence
- Chargement des données sur MongoDB (bdd denormalization collection denorm\_station\_prices)(6 749 enregistrements)

#### **Pour les données des prix d'essences officiel hebdomadaire:**

- Fonctionnement du bot python qui va récupérer une url valide
- Chargement du csv en utilisant l'url, contenant toutes les données de 1985 à 2025
- Renommer des colonnes
- Chargement des données sur MongoDB (bdd datalake collection official\_oils\_prices)

#### **Pour le merge des deux types de données ensemble:**

- Récupération des données sur la collection denorm\_station\_prices
- Récupération des données sur la collection official\_oils\_prices
- Réalisation d'un merge sur la date
- Création de colonnes stratégiques pour aider à la visualisation (day\_of\_week, Month, Year, DayMonth). En effet, certains traitements ne sont pas disponibles sur l'outil de visualisation Metabase avec des données issues de MongoDB.
- Chargement des données sur MongoDB (bdd denormalization collection denorm\_station\_vs\_official\_prices)(7 897 enregistrements)

Les données sont ensuite lues en temps réel par le Docker Métabase en version gratuite.

Il récupère les 2 bases de données "datalake" et "denormalization", et met à jour de manière automatique les graphiques a chaque minute.

## La Dataviz :

Voici les noms des différents graphiques présentés sur la dashboard principal:  
“Presentation comparaison donnees officiels et releves stations essence”

### *page Données comparés:*

Il s'agit du compte d'enregistrement par type d'essence pour le côté officiel et pour le côté station essence.

Le calcul est nombre d'enregistrement ou type d'essence est égal au minimum à 0.1euros

### *page Différences de résultat sur le Gazole et page Différences de résultat sur le SP98:*

Il s'agit d'une visualisation temporelle permettant de comparer les prix côté officiel et côté station essence.

Le calcul est la moyenne de type d'essence côté officielle et moyenne de type d'essence côté station essence sur la date en journalier.

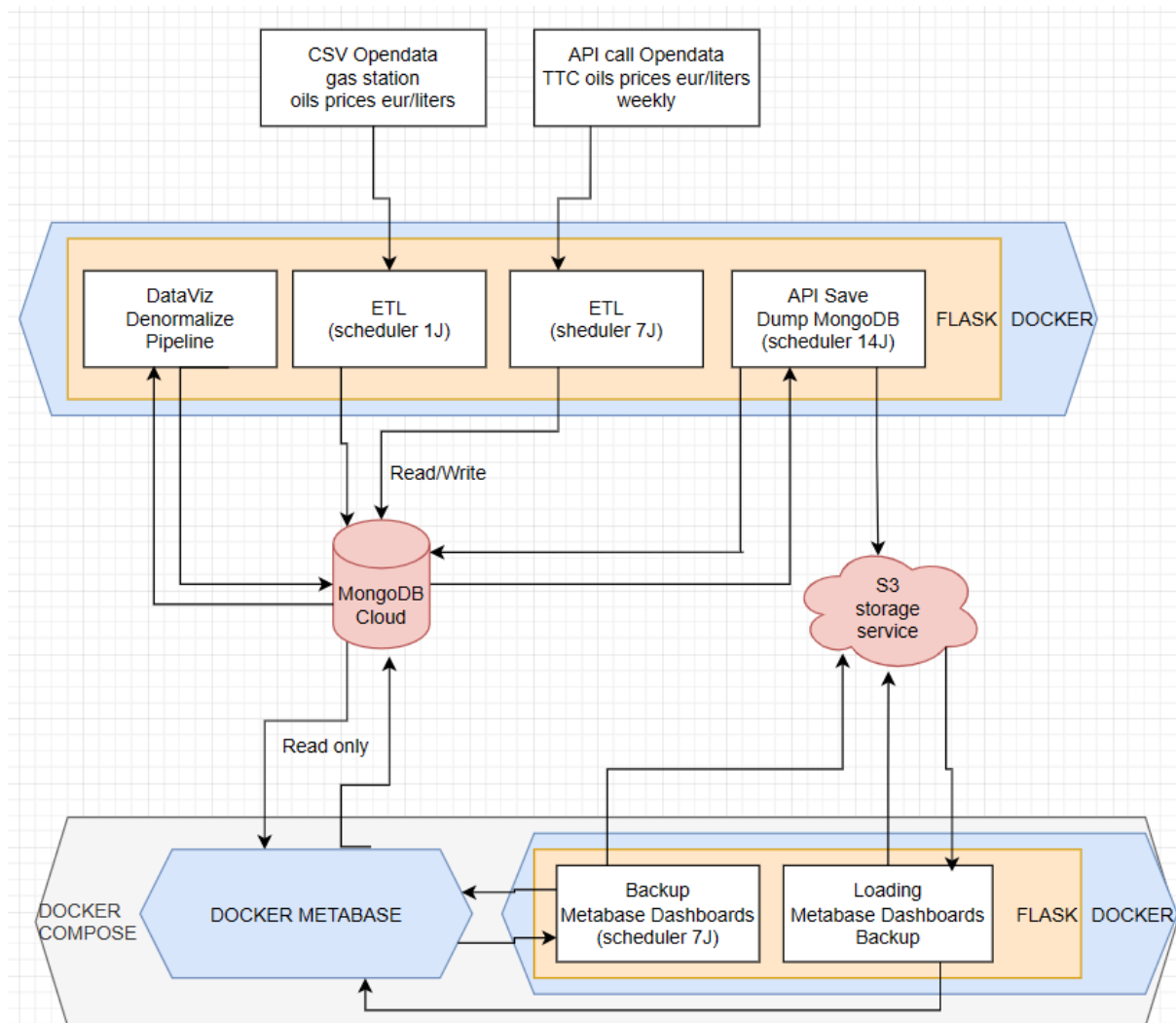
### *page Prix par jour de la semaine:*

Il s'agit d'une comparaison des prix des essences sur les données stations essences en divisant par jour de la semaine

Le calcul est la moyenne de type d'essence côté station essence par jour de la semaine ayant au moins une valeur supérieur à 0.1euros.

## Le schema d'architecture :

Voici le schéma d'architecture des ETLs et de la partie Datavisualisation.



Elle est composé de 3 grandes parties:

- La partie ETL situé dans un docker. C'est elle qui réalise l'ensemble des transformations de données. Elle a un droit de lecture et d'écriture sur la MongoDB et sur le S3 lors de dump ou de restauration de base de données MongoDB. Cette partie de dump et restauration est principalement utilisé lors du backup automatique des bases de la MongoDB.
- La partie MongoDB et S3 Amazon. Il s'agit d'une partie entièrement cloud lors de l'industrialisation. Les variables d'environnements sont données à la partie ETL et la partie Datavisualisation pour qu'elles puissent communiquer ensemble.

- La partie Datavisualisation est constituée de 2 dockers lancés en utilisant un docker compose.

Le premier docker est le docker natif de Metabase en version gratuite.

Le second docker est constitué d'un serveur flask permettant de réaliser des backup de la base de données de Metabase ainsi que des restauration de base de données.

Il est donc possible de relancer le docker Metabase et de récupérer un compte déjà existant avec des graphiques déjà construits sur des bases de données MongoDB déjà connectés.