

DOSSIER DE VALIDATION

Ingénieur en science des données

spécialisé en infrastructure data

pour le Titre RNCP 39586

Bloc 1: Collecter, transformer et sécuriser des données

| Nom Prénom | CORIN Gaëtan |
| --- | --- |
| Nom Prénom du tuteur | MOULIN Alexis |
| Niveau visé | RNCP 7 |
| Date de la soutenance | Septembre 2025 |
| Lieu de la soutenance | Toulouse |

Table des matières

[**Introduction. 3**](#_lkdhe4y1ytyp)

[**C1.1.2 : Mettre en œuvre des techniques de collecte de données en exploitant les API externes et les bases de données disponibles, des techniques de web crawling et de web scraping afin de recueillir les données ciblées. 4**](#_b3ft6usja70a)

[**C1.1.3 : Automatiser la collecte de données en mettant en place des tâches planifiées et/ou des flux temps réel, en utilisant des logiciels d’automatisation afin de garantir l’actualisation des données. 4**](#_xu4v09r5mxto)

[**C1.2.1 : Élaborer la stratégie de stockage des données avec un modèle de données adéquat en intégrant les différents types de données, l’utilisation envisagée (analyse, stockage, disponibilité, accessibilité) et le volume à stocker afin d’organiser le stockage des données. 4**](#_xedo97hswxa2)

[**C1.2.2 : Construire une base de données en sélectionnant la technologie (SQL, NoSQL), un système de gestion de base de données (SGBD) et/ou une solution de stockage BIG DATA, en assurant le paramétrage et l’implémentation afin de mettre en œuvre le modèle de données qui garantit la disponibilité et l’intégrité des données. 4**](#_llo8lkahs3yt)

[**C1.3.1 : Sélectionner les technologies et les outils de traitement de données en identifiant les solutions existantes et en comparant leurs avantages et leurs inconvénients afin de traiter efficacement les données collectées. 4**](#_ueljvy826n95)

[**C1.3.2 : Transformer les données à l’aide de langage de programmation ou en utilisant des outils dédiés (Talend, Spark…) afin d’obtenir des données nettoyées exploitables. 4**](#_hrv3uxlrcrci)

[**C1.3.3 : Développer un processus ETL en identifiant les bénéfices des technologies ETL (ex : facilité de développement), en exploitant la technologie ETL préalablement sélectionnée afin d’automatiser l’extraction, la transformation et le chargement de données. 5**](#_iouoshj55hut)

[**C1.4.1 : Définir la politique de sécurisation des données en évaluant les risques, en qualifiant leur niveau de sensibilité, en identifiant les droits d’accès selon les rôles des différentes parties prenantes et en respectant les exigences légales (ex : RGPD ) afin de garantir la bonne utilisation et l’intégrité des données. 5**](#_wu78rnjl1i7c)

[**C1.4.2 : Concevoir une architecture sécurisée et robuste, en intégrant des mesures de sécurité multicouches et des contrôles d'accès stricts, en mettant en place des solutions de protection des données (ex : chiffrement) pour sécuriser les données en transit et au repos, en veillant à l’anonymisation des données personnelles afin de répondre aux normes de protection de la vie privée et de sécurité des données. 5**](#_wbnmobkfxy4j)

[**Conclusion. 5**](#_mdwlf2lh30f9)

# 

# Introduction.

Je m’appelle Gaëtan Corin, j’ai 30 ans, et je suis en reconversion professionnelle (précédemment boulanger patissier). Je travaille actuellement en alternance chez Menaps, une startup sur Toulouse, ou je travaille principalement sur des extractions et transformations de données, ainsi que des créations de pipelines permettant de formater les données pour des KPIs métiers. Malgré ma volonté très tôt de vouloir devenir Data-Engineer, mes années de licence 2 et 3 ont été principalement orientées vers un profil Fullstack, ce qui me permet aujourd’hui d’avoir une vision complète des projets.

Le projet que je vais vous présenter dans ce mémoire a été intégralement développé sur mon temps libre. Il est disponible sur mon github:

<https://github.com/gaetancorin/Datapipeline_comparaison_official_vs_gas_stations_reporting>

Ce mémoire a été intégralement écrit à la main sans générateur de texte ou LLM.

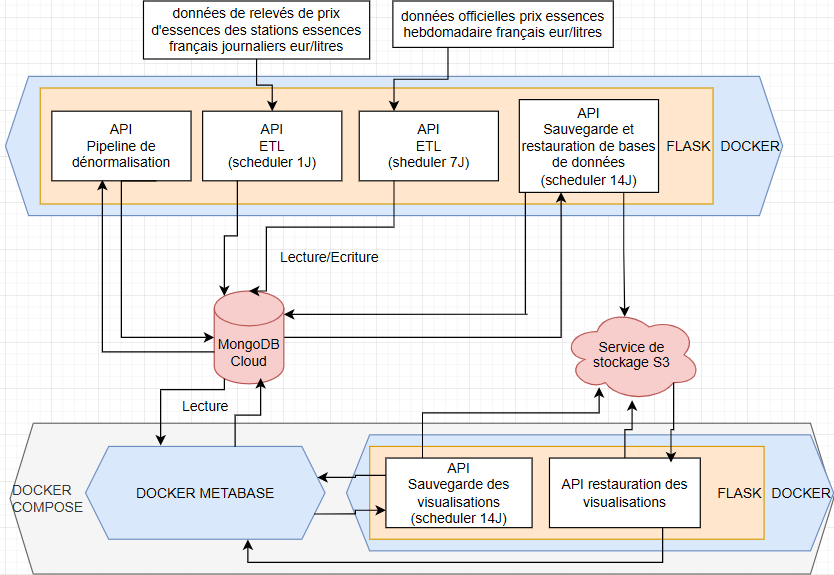
**La problématique** de ce projet vient d’une situation théorique, où le pôle data du gouvernement français possède un jeu de données officiel sur l'évolution du prix des différentes essences vendu en France. Ces données sont rassemblées en moyenne hebdomadaire. Malheureusement, ils ne connaissent pas la méthodologie qui a permis de calculer ces moyennes hebdomadaires à l'époque où celle-ci ont été calculées.

Ils se posent donc la question de la véracité et fiabilité des transformations des données anciennement récoltés, et souhaiteraient qu'un recalcul soit fait sur d'autres données journalières existantes issus des stations essences, afin de vérifier la crédibilité des transformations passées.

Ils aimeraient aussi une vision plus fine des cours des différentes essences vendus par les stations essences française, afin d'en tirer des conclusions et avoir une meilleure visibilité en terme journalière. Cet aperçu devra être mis à jour régulièrement.

Le cahier des charges a préalablement été réalisé dans le bloc 3 de ce titre RNCP.

Voici l’architecture du projet qui a était retenu:



C1.1.1 : Élaborer une stratégie de collecte de données en définissant les données utiles et nécessaires pour répondre à une problématique, et en identifiant les sources des données afin de cadrer le travail à réaliser pour collecter les données ciblées.

Au vu de la problématique expliquée plus haut, la stratégie déployée pour la collecte des données consiste à extraire les données de relevés de prix des stations essences journalières, puis de les traiter et les stocker en base de données.

Il faudra ensuite extraire les données officielles des prix des essences hebdomadaires, les traiter et les stocker en base de données.

Enfin, il va falloir créer une table de dénormalisation contenant les précédentes données transformées, afin de pouvoir facilement en faire des comparaisons et analyses lors des visualisations.

**Pour les données de relevés des stations essences journalière:**

Elles sont accessible sur le site suivant: <https://www.prix-carburants.gouv.fr/rubrique/opendata/>

Il s’agit d’un site gouvernemental qui met à disposition ces données dans un objectif de transparence, de régulation du marché et d’information du consommateur.

Le formatage de ces données est sous format “zip”, avec une url disponible pour chaque année de données. Il faudra donc récupérer l’ensemble des données en faisant des appels API vers chacune des urls, afin de récupérer l’intégralité des données. Il faudra ensuite dézipper chaque fichier récupéré, réaliser les transformations nécessaires, puis les stocker en base de données. L’ensemble des données brut est au nombre de 56 600 000 enregistrements, car chaque modification de prix au cours de la journée est répertoriée pour chaque station essence de 2007 à aujourd'hui.

**Pour les données officielles des prix des essences hebdomadaire:**

Elles sont accessible sur le site suivant: <https://www.ecologie.gouv.fr/politiques-publiques/prix-produits-petroliers>

Il s’agit d’un autre site gouvernemental qui met à disposition ces données, cette fois-ci dans un but écologique et de compréhension des différentes tranches de composition de prix des essences.

Une complexité est présente pour extraire ces données. En effet, un formulaire en Ajax doit être rempli et validé pour avoir accès à un élément HTML dynamique contenant une URL constitué d’un UUID. C’est cette url qui devra être scrappée pour accéder aux données.

La plus grande complexité se situe dans le formulaire en Ajax, qui est un formulaire dynamique Javascript. Il faut donc un bot avec un simulateur de navigateur qui doit réaliser la connexion, remplir le formulaire et le valider afin de pouvoir accéder à l’URL a scrapper.

Il faudra par la suite appeler l’url pour récupérer le fichier de données, réaliser les transformations nécessaires, puis les stocker en base de données.

**Pour la table de dénormalisation:**

Il s’agit d’une table constituée des deux types de données précédentes. La colonne de la Date servira de jointure. Il faudra donc récupérer les deux types de données en bases de données, réaliser les éventuelles transformations dont la jointure, puis la stocker sous format de collection en base de données pour faciliter les visualisations et analyses futures.

# C1.1.2 : Mettre en œuvre des techniques de collecte de données en exploitant les API externes et les bases de données disponibles, des techniques de web crawling et de web scraping afin de recueillir les données ciblées.

Trois techniques de collecte de données sont utilisés dans ce projet

La première technique est réalisée sur les données officielles. Elle consiste à utiliser un bot réalisé avec la librairie python Selenium. Ce bot va simuler une navigation Chrome, naviguer sur le formulaire dynamique, remplir les champs les uns après les autres , puis valider le formulaire dans l’objectif d’accéder à l’élément HTML dynamique contenant l’URL constitué d’un UUID.

Il va ensuite scrapper l’HTML pour récupérer cette URL.

Une vérification de résultat à chaque étape du bot nous permet de définir son bon fonctionnement, jusqu'à l'existence de l’élément HTML dynamique ou nous pourrons scrapper l’URL voulu. Une vérification de formatage de l’URL scrapper est effectuée afin d’assurer son exactitude, et son futur appel sera contrôlé par le status code ainsi que les colonnes récupérées, afin de vérifier l’exhaustivité.

La seconde technique consiste à récupérer les données des relevés de stations essences en appelant les URLs des années concernés, puis les dezipper. Pour réaliser cela, les librairie “Requests” et “Zipfile” de python sont utilisés.  
Une vérification des statuts code est réalisée sur chacune des URLs, afin d’être certain de ne pas avoir d’année manquante sur la base de données finale et d’assurer l’exactitude de ces informations. Une vérification est aussi réalisée lors de la transformation du fichier zip en fichier xml, puis csv. De plus, une vérification des colonnes ciblés est contrôlée sur chaque csv représentant une année afin de s’assurer de l’exhaustivité finale des données.

La troisième technique est réalisée lors de la dénormalisation des deux types de données précédentes. Ces données étant désormais stockées en bases de données, il va falloir appeler les collections concernées directement sur la base de données en réalisant une requête pymongo, récupérer les données ce ces collections en vérifiant l’exactitude des données récupérées, les traiter en dénormalisant ces données, puis les stocker dans une nouvelle collection.

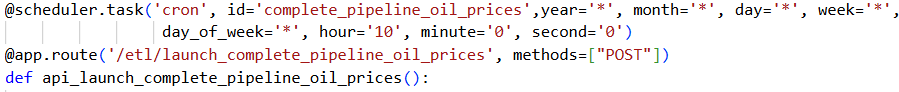
# C1.1.3 : Automatiser la collecte de données en mettant en place des tâches planifiées et/ou des flux temps réel, en utilisant des logiciels d’automatisation afin de garantir l’actualisation des données.

Au sein du serveur Flask contenant la logique métier des ETLs, il y a deux planificateurs de taches réalisés avec la librairie APScheduler de flask\_apscheduler.

Le premier planificateur de tâche consiste à réaliser des sauvegardes de base de données vers un espace de stockage cloud (Amazon S3). Ceci ajoute une sécurité supplémentaire aux données en cas de cyberattaque, ransomware, ou bien en cas d’erreur sur la base de données. Ce planificateur se lance toutes les 2 semaines de manière automatique, après le chargement journalier.

Le second planificateur de tâche est celui qui va exécuter les fonctions d’ETLs.

Pour chacune de ces fonctions ETLs, un programme est appelé juste avant l’extraction, qui va aller vérifier dans la futur collection quelles sont les données existantes déjà traitées. Cela permet ainsi de ne pas charger l’entièreté des données, mais uniquement les données manquantes, dans une optique de mise à jours et dans le respect du RSE.



# C1.2.1 : Élaborer la stratégie de stockage des données avec un modèle de données adéquat en intégrant les différents types de données, l’utilisation envisagée (analyse, stockage, disponibilité, accessibilité) et le volume à stocker afin d’organiser le stockage des données.

Il y a deux stratégies de stockage des données sur cette application car il y a deux types de besoins différents.

**La première stratégie de stockage** vient d’une volonté de **stocker les données brut** dans la mesure du possible sur un principe de **datalake**, afin de pouvoir être réutiliser dans des cas métiers n’ayant pas encore été défini. Cette volonté de stockage ne doit pas être dirigée par un cas de visualisation spécifique que l’on souhaite implémenter.

Pour cette première stratégie, les données officielles des prix des essences n’ont pratiquement subi aucune transformation. La colonne de date a été modifiée, les noms des colonnes ont été modifiés, puis les données ont directement été chargées en base de données. Ayant une quantité de données relativement faible de données(2000 enregistrements), aucune modification supplémentaire n’a été nécessaire.

{

**\_id**: 685ea7d9984bc1281b1967ad (type ObjectId)

**Date**: 2008-01-04T00:00:00.000+00:00 (type Date)

**official\_ttc\_GAZOLE\_eur\_liter:** 1.2226 (type float)

**official\_ttc\_SP95\_eur\_liter**: 1.3701 (type float)

**official\_ttc\_E10\_eur\_liter**: NaN (type float)

**official\_ttc\_SP98\_eur\_liter:** 1.4007 (type float)

**official\_ttc\_E85\_eur\_liter**: NaN (type float)

**official\_ttc\_GPLC\_eur\_liter**: 0.7472 (type float)

}

En revanche, sur les données relevées par les stations essences, le format d’origine en xml est formaté de manière à ce que chaque station essence possède un bloc de données correspondant à l’ensemble de ces relevés de prix pour chaque type d’essence. Il s’agit d’une quantité énorme de données par stations essence, car nous sommes à 56 millions d’enregistrements pour 13 600 stations sur la période de 2007 à 2025.

Afin d’alléger la quantité d’enregistrements et garantir l’intégrité des données, un nettoyage des données non pertinentes, ainsi que la conservation uniquement des derniers prix relevés par jour et par type d’essence est conservé.

Il y a ensuite une division de ce jeu de données en deux entités. La première représente les stations essences. Ces données changent rarement, et ne sont pas cumulées mais mises à jour par date de parution.

La deuxième entité représente les données des modifications de prix par jour et type d’essence sur une station essence. Chaque nouvel enregistrement est relié à la station essence par un id unique.

{

**\_id:** 685ea3f5984bc1281bbdba53 (type ObjectId)

**Id\_station\_essence:** 1000001 (type int)

**Adresse:** "596 AVENUE DE TREVOUX" (type string)

**Ville:** "SAINT-DENIS-LèS-BOURG" (type string)

**Cp:** 1000 (type int)

**Latitude:** 4620100 (type int)

**Longitude:** 519800 (type int)

**Derniere\_maj:** 2025-06-25T00:00:00.000+00:00 (type Date)

}

{

**\_id:** 685ea3fd984bc1281bbdd80c (type ObjectId)

**Date:** 2009-01-05T00:00:00.000+00:00 (type Date)

**Id\_station\_essence:** 1000001 (type int)

**Nom:** "Gazole" (type string)

**Valeur:** 0.909 (type float)

**Heuremin:** "08:53" (type float)

}

**La seconde stratégie de stockage** a pour objectif de réaliser des transformations sur les données en datalake afin de les formaté pour travailler dessus en analyses et en visualisation. Le travail sur ces données est donc réalisé sur un principe de **dénormalisation**.

La première dénormalisation a lieu sur la collection des prix des stations essences par type d’essence, jour et station essence. Ce jeu de données contient encore 39 millions d’enregistrements. L’objectif est donc de rassembler ces données en un seul enregistrement par jour, contenant l’ensemble des types d’essence ainsi que la moyenne des prix de chacun, créant un total de 6 700 enregistrements.

{

**\_id:** 685ea7df984bc1281b196b3d (type ObjectId)

**Date**: 2009-01-01T00:00:00.000+00:00 (type Date)

**station\_ttc\_E85\_eur\_liter**: 0.8279 (type float)

**station\_ttc\_GPLC\_eur\_liter:** NaN (type float)

**station\_ttc\_GAZOLE\_eur\_liter**: 1.10647 (type float)

**station\_ttc\_SP95\_eur\_liter:** 1.24179 (type float)

**station\_ttc\_E10\_eur\_liter:** 1.13348 (type float)

**station\_ttc\_SP98\_eur\_liter:** NaN (type float)

}

Enfin, la seconde dénormalisation consiste à rassembler ensemble les données des prix d’essences officiels avec les données des prix d’essences des stations essences. De nouvelles colonnes sont aussi ajoutées pour faciliter les analyses et la visualisation.  
{

**\_id:** 685eaa1a984bc1281b198bdd (type ObjectId)

**Date:** 2008-03-28T00:00:00.000+00:00 (type Date)

**official\_ttc\_GAZOLE\_eur\_liter:** 1.2587 (type float)

**official\_ttc\_SP95\_eur\_liter**: 1.3656 (type float)

**official\_ttc\_E10\_eur\_liter**: NaN (type float)

**official\_ttc\_SP98\_eur\_liter:** 1.3993 (type float)

**official\_ttc\_E85\_eur\_liter:** NaN (type float)

**official\_ttc\_GPLC\_eur\_liter:** 0.7494 (type float)

**station\_ttc\_GAZOLE\_eur\_liter:** 1.23422 (type float)

**station\_ttc\_SP95\_eur\_liter:** 1.36403 (type float)

**station\_ttc\_E10\_eur\_liter:** NaN (type float)

**station\_ttc\_SP98\_eur\_liter:** NaN (type float)

**station\_ttc\_E85\_eur\_liter:** 0.81043 (type float)

**station\_ttc\_GPLC\_eur\_liter:** 0.73473 (type float)

**Day\_of\_week:** "Friday" (type string)

**Month:** "March" (type string)

**Year**: "2008" (type string)

**DayMonth**: "28march" (type string)

}

Durant l’étape d’ETL, les données nouvellement stockées en bases de données sont uniquement les données manquantes, dans une optique de mise à jour, mais aussi pour assurer l’intégrité des données en évitant les valeurs doublons.

Concernant les moyens de manipulation de ces données, il est obligatoire de passer par les différents points d’APIs afin de réaliser des transformations. Enfin, en cas de désir de réinitialisation d’une base de données ou d’une collection, des APIs dédiés permettent de supprimer les données voulues de manière sécurisée sans toucher directement à la base de données.

Afin d’assurer l’intégrité des données, différents droits sont donnés suivant les besoins. Le docker d’ETL aura le droit de lecture et d’écriture sur l’ensemble des bases de données, tandis que le docker de visualisation n’aura que le droit que lecture, et cela uniquement sur les tables de dénormalisation destiné à la visualisation.

C1.2.2 : Construire une base de données en sélectionnant la technologie (SQL, NoSQL), un système de gestion de base de données (SGBD) et/ou une solution de stockage BIG DATA, en assurant le paramétrage et l’implémentation afin de mettre en œuvre le modèle de données qui garantit la disponibilité et l’intégrité des données.

Ninja !

# C1.3.1 : Sélectionner les technologies et les outils de traitement de données en identifiant les solutions existantes et en comparant leurs avantages et leurs inconvénients afin de traiter efficacement les données collectées.

# C1.3.2 : Transformer les données à l’aide de langage de programmation ou en utilisant des outils dédiés (Talend, Spark…) afin d’obtenir des données nettoyées exploitables.

# C1.3.3 : Développer un processus ETL en identifiant les bénéfices des technologies ETL (ex : facilité de développement), en exploitant la technologie ETL préalablement sélectionnée afin d’automatiser l’extraction, la transformation et le chargement de données.

# C1.4.1 : Définir la politique de sécurisation des données en évaluant les risques, en qualifiant leur niveau de sensibilité, en identifiant les droits d’accès selon les rôles des différentes parties prenantes et en respectant les exigences légales (ex : RGPD ) afin de garantir la bonne utilisation et l’intégrité des données.

# C1.4.2 : Concevoir une architecture sécurisée et robuste, en intégrant des mesures de sécurité multicouches et des contrôles d'accès stricts, en mettant en place des solutions de protection des données (ex : chiffrement) pour sécuriser les données en transit et au repos, en veillant à l’anonymisation des données personnelles afin de répondre aux normes de protection de la vie privée et de sécurité des données.

# Conclusion.

# 