DOSSIER DE VALIDATION

Ingénieur en science des données
spécialisé en infrastructure data
pour le Titre RNCP 39586

Bloc 3: Élaborer et piloter un projet Data

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

Je m'appelle Gaëtan Corin, j'ai 30 ans, et je suis en reconversion professionnelle depuis 4 ans. J'ai réalisé un CAP boulanger et un CAP pâtissier où j'ai exercé ces métiers pour une durée totale de 10 ans, puis j'ai décidé de me reconvertir.

Mon parcours de reconversion a commencé au centre de formation de l'Adrar Pôle Numérique ou j'ai réalisé une année de formation en présentiel me donnant le titre RNCP 5 Développeur d'application Web avec option Devops.

J'ai ensuite réalisé une année en alternance avec l'école supérieure IPI Blagnac, ainsi que l'entreprise CELAD. Durant cette alternance, j'ai travaillé dans une équipe sur un projet interne en tant que Développeur Fullstack sur les langages Python et Angular, me donnant une première expérience professionnelle sur un projet ambitieux. J'ai eu l'opportunité d'avoir un chef de projet et un Scrum Master qui maîtrisaient parfaitement la méthode Scrum, ce qui a été riche en enseignements et qui m'a inspiré pour ce mémoire ainsi que dans ma méthode de travail au quotidien.
En plus de cette mission, j'ai aussi eu l'opportunité de partir d'un projet de zéro pour le client Renault, où j'ai réalisé l'intégralité de l'application de moi-même en 3 mois avec le support et les conseils de mon responsable d'alternance qui m'a aider pour la conception, et qui faisait l'intégralité de mes revues de code.
Ce projet a pu me servir de sujet de mémoire pour passer mon titre RNCP 6 Concepteur développeur d'application numérique, où j'ai obtenu les félicitations du jury.

Je continue ainsi mon parcours avec l'école supérieure Ynov, ou je me spécialise en Data-Engineer. C'est un métier qui m'intéresse depuis la fin de ma première année d'informatique et où j'ai eu la chance de concrétiser cette ambition pour mon mastère. Mon alternance est réalisée avec l'entreprise Menaps, une startup toulousaine. J'ai pu être encadré par un Data-Engineer sénior durant la première année qui m'a appris les fondements de ce métier sur un grand projet orienté Data pour le client Stellantis. Par la suite, le Data-Engineer sénior est parti à la fin de ma première année d'alternance, et j'ai pris la relève de son poste, étant le dernier Data-Engineer de l'entreprise. J'ai donc été promu au rang de Data-Engineer référent sur le projet aux yeux du client (ayant un Data-scientist en référent technique si besoin). J'ai réalisé des réunions avec le client une fois par semaine, où je devais lui expliquer l'état d'avancement du projet, prendre en compte les éventuelles remarques et demandes du client, et continuer les améliorations et réalisations en cours.
Le projet étant désormais terminé, je travaille avec mon chef de projet et l'équipe afin de rendre notre projet simple et fonctionnel pour un futur potentiel client.

J'ai énormément appris durant ces 4 années et j'ai comme ambition de continuer dans le métier de Data-Engineer ou je me sens prêt à relever les défis qui me seront proposés.

Je tiens a préciser que l'ensemble de ce mémoire a été écrit entièrement à la main, sans aide de rédaction de LLM, et que le projet présenté dans ce mémoire a été entièrement réalisé par mes soins durant mon temps libre, disponible sur mon Github:
https://github.com/gaetancorin/Datapipeline_comparaison_official_vs_gas_stations_reporting

C3.1.1 : Définir les objectifs à atteindre et le périmètre du projet, en analysant les contraintes techniques et réglementaires, en étudiant le contexte et les enjeux afin de dimensionner le projet en termes de délai et budget.

Problématique:
Le pôle data du gouvernement français possède un jeu de données officiel sur l'évolution du prix des différentes essences vendues 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 celles-ci ont été calculées.

Ils se posent donc la question de la véracité et de la fiabilité des transformations des données anciennement récoltés, et souhaiteraient qu'un recalcul soit fait sur d'autres données journalières existantes afin de vérifier la crédibilité des transformations passées.
Ils aimeraient aussi une vision plus fine des cours des différentes essences vendues en France, afin d'en tirer des conclusions et avoir une meilleure visibilité à l'échelle journalière.
Cet aperçu devra être mis à jour régulièrement.

Objectif:
Réalisation d'un audit sur la qualité des données officielles hebdomadaires en comparaison à un autre jeu de données journalier fourni par le gouvernement.
Recherche de décalages de prix sur les données historiques qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles.
Analyses sur une granulométrie temporelle plus fine qu'actuellement afin d'en tirer des conclusions.
Le rendu sera fait sous format de visualisations claires qui pourront être automatiquement rafraîchies lors de nouvelles données.
La base de données ainsi que les visualisations doivent pouvoir être sauvegardées et restaurées dans un environnement externe, afin d'assurer la facilité de déploiement du système dans d'autres environnements.

Cadre réglementaire:
La "doctrine cloud au centre" de l'État incite les applications gouvernementales à être déployées sur des environnements cloud de confiance.
Il sera donc nécessaire de construire l'application de manière à être entièrement déployé sur des clouds réputés.

Environnement:
L'environnement est laissé libre durant la réalisation du projet. Le projet devra ensuite être entièrement dockerisé afin d'être déployé sur les services cloud.
Les données pourront aussi être stockées dans des clouds, car il ne s'agit pas de données sensibles. En effet, nous ne travaillons qu' avec des données OpenData.

Contraintes:
Il est obligatoire de ne croiser que les données gouvernementales issues des différents sites internet de l'Etat, afin d'assurer une viabilité des données reconnue par celui-ci.

RSE:
Une attention particulière devra être faite sur le chargement de données.
En effet, il faut éviter le surchargement inutile sur l'ensemble de l'historique des données chaque jour, dans une optique de préservation écologique.
Les problèmes de confidentialité des données suivant la loi RGPD sont faiblement impactants sur ce projet, car l'intégralité des données est déjà disponible en Open Data et anonymisée. Les données sources sont donc déjà considérées comme respectant la loi RGPD.

C3.1.2 : Dimensionner le projet en évaluant la charge de travail et les ressources nécessaires (humaines, matérielles) au regard des exigences attendues et des contraintes préalablement définies afin d'estimer le temps et le budget nécessaires à la faisabilité du projet.

Le projet sera réalisé par 2 Data-Engineers de niveau intermédiaire à plein temps ayant chacun un ordinateur suffisamment performant pour faire fonctionner des programmes avec des grands jeux de données.

La charge financière de la base de données dans le cloud sera financée par le client lors du déploiement, et pourra être hébergée en local lors du développement.

La charge financière des sauvegardes de bases de données et des graphiques de visualisations stockées dans le cloud sera financée par le client lors du déploiement.
En prenant en compte la taille des deux jeux de données traités, l'estimation du prix des sauvegardes de base de données lors de la période de développement sur un stockage S3 est estimée à un volume de 50Go par mois (surestimé) pour un montant moyen de 1 euro par mois. Il s'agit donc d'une charge négligeable.

Quant à la charge financière de l'outil de visualisation, celui-ci pourra être un outil opensource.

La charge de travail globale est estimé comme ceci:

Attentes	Temps estimés
Collecte du besoin et des attentes	2 jours
Planification de la méthode de travail (Agile)	2 jours
Collecte et analyse initiale des données brutes journalières	2 jours
Nettoyage et transformation des données brutes journalières	5 jours
Chargement des données brutes journalières en base de données	3 jours
Collecte et analyse initiale des données officielles hebdomadaire	2 jours
Nettoyage et transformation des données officielles hebdomadaire	2 jours
Chargement des données officielles hebdomadaire en base de données	2 jours
Création d'un jeu de données dénormalisé pour faciliter la visualisation	2 jours
Analyses et création de graphiques de visualisation	5 jours
Implémentation d'outils de sauvegarde et de restauration des bases de données	2 jours
Implémentation d'outils de sauvegarde et de restauration des graphiques de visualisations	2 jours
Rédaction de documentations techniques permettant la maintenance de l'application	3 jours
TOTAL	**34 jours**

La charge financière de travail estimée est donc de 34 jours travaillés.

En prenant en compte le fait que les 2 Data-Engineers travailleront sur le projet à plein temps a 500 euros brut par jour en simultané, la durée totale du projet est estimée à environ **3 semaines et demi**, pour un montant total de **17 500 euros brut**. Il faut cependant que le client reste conscient que des imprévus peuvent arriver durant la réalisation du projet, et qu'une marge de flexibilité peut être demandée suivant les complexités et les imprévus rencontrés.

En termes de **faisabilité**, celle-ci semble optimiste car les différentes contraintes, cadre réglementaire, environnement et RSE ne semblent pas être des points bloquants quant à la capacité de conception et réalisation du projet.

C3.1.3 : Rédiger la documentation projet, en identifiant les parties prenantes concernées, en prenant en compte l'ensemble des caractéristiques du projet, afin de clarifier et formaliser les attendus.

Lors de la réalisation de ce projet, un référent métier sera désigné au sein du pôle data du gouvernement afin d'être un point de contact durant tout le long de la réalisation.

Un des Data-Engineer sera élu chef de projet et sera le responsable de la réalisation. L'autre Data-Engineer sera un exécutant tout en étant un acteur de la méthode Agile.

En prenant en compte les caractéristiques du projet, les attendus sont divisés en plusieurs parties:

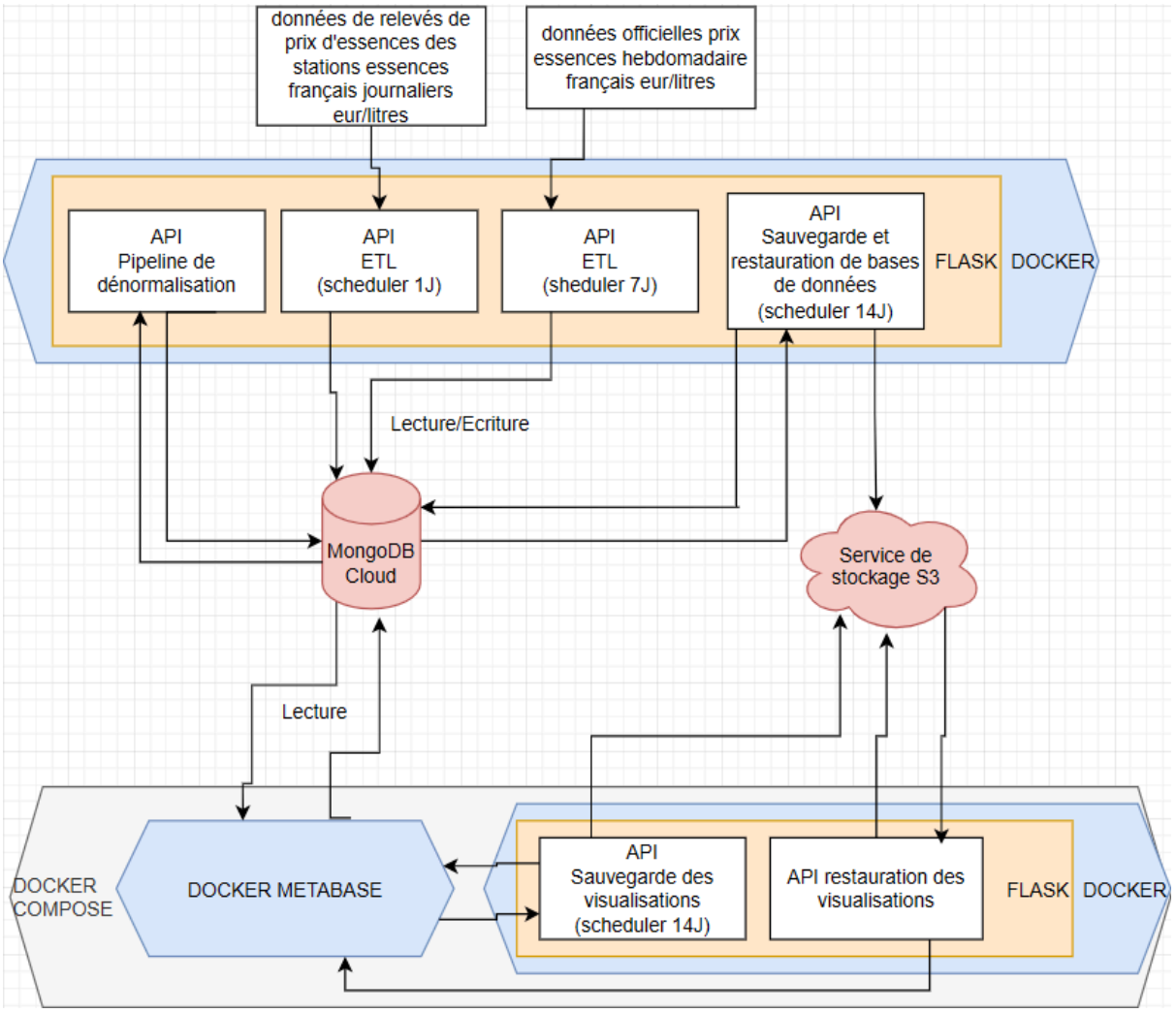
- Le docker contenant la partie d'extraction, de transformation et de chargement des deux flux de données en base de données cloud. Il doit aussi permettre d'extraire les données traitées de la base de données vers un espace de stockage externe cloud, ainsi que les restaurer vers la base de données. Tout doit être fonctionnel via des APIs et des planificateurs de tâches.
- Les dockers contenant l'outil de visualisation. Il doit pouvoir se connecter en temps réel à la base de données cloud, et mettre automatiquement à jour ces visualisations. Nous devons avoir la possibilité d'extraire les visualisations vers un espace de stockage externe cloud, ainsi que de les restaurer. Tout doit être fonctionnel via des URLs, APIs et des planificateurs de tâches.
- Des documentations détaillées sur le fonctionnement des 2 dockers, sur le fonctionnement de l'outil de visualisation, sur les données sources ainsi que sur l'architecture des données stockées en base de données.

La création de base de données en cloud ainsi que la création d'espace de stockage externe cloud ne sont pas considérées comme un rendu, car ceux-ci seront implémentés par l'équipe du pôle data du gouvernement. Cependant, il est nécessaire d'avoir une base de données fonctionnelle et un espace de stockage externe fonctionnel lors du développement afin de tester la solution et que l'application soit prête à l'emploi lors de la livraison finale.

Le code sera réalisé en anglais, mais la documentation technique devra être en français.

En prenant en compte le fait que le projet est à destination de personnes travaillant dans le domaine du traitement de données, les documentations et cahiers des charges pourront contenir des mots techniques.

Voici l'architecture complète attendue pour ce projet:



Nous pouvons voir dans cette architecture la première partie d'ETL réalisée en serveur Flask contenu au sein d'un Docker. Différents points d'API serviront à réaliser les cas d'usages nécessaires à l'application, tels que la collecte et transformation des données récupérées en Open Data sur les sites gouvernementaux, mais une pipeline lancée par un planificateur de tâches permettra de réaliser les tâches quotidiennes de manière automatique. Cette partie d'ETL est connectée à une base de données MongoDB Cloud avec comme connexion un acteur en droit de lecture et d'écriture. Elle possède aussi une connexion à un espace de stockage S3 pour réaliser les sauvegardes et les restaurations automatiques des données stockées en base de données.

La seconde partie de cette architecture est la partie de visualisation avec Métabase. C'est un outil qui offre un déploiement par docker de manière natif et gratuit. Il suffit de simplement utiliser l'image officielle de Métabase pour construire son Docker, et de paramétrer sa connexion vers la base de données MongoDB cloud en utilisant un acteur en droit de lecture

uniquement. Un serveur flask dockerisé s'occupera de la partie de sauvegarde et de restauration de la base de données des visualisations de Métabase en planificateur de tâches vers un espace de stockage externe S3. Ces deux dockers sont ensuite rassemblés dans un docker-compose afin de faciliter le déploiement et la communication entre eux.

C3.2.1 : Planifier l'exécution du projet en organisant la répartition et l'ordonnancement des activités, le planning prévisionnel de réalisation et les ressources nécessaires à son exécution, en prenant en considération les personnes en situation de handicap afin de suivre les différentes phases du projet.

La méthodologie choisie pour la réalisation de ce projet est la Méthode Scrum, en utilisant des outils de planification et de visualisation. La réalisation de cette méthode se fera par l'élection d'un Data-Engineer qui tiendra le rôle de référent vers le client et de Scrum Master en plus de son travail de Data-Engineer.

Le Data-Engineer référent devra diviser le travail en tickets, animer un Poker Planning avec l'autre Data-Engineer afin d'évaluer la complexité de chaque ticket, puis réaliser un diagramme de Gantt permettant d'organiser le planning théorique et la répartition des tâches sur chaque sprint en prenant en compte la matrice RACI.

Cette méthode de travail en Scrum sera réalisée avec 6 sprints de 3 jours travaillés, avec un Daily tous les matins, ainsi qu'un rétro planning à la fin de chaque sprint pour faire un point sur l'avancement du projet. Des présentations client seront faites à la fin de chaque sprint afin de leur fournir une visualisation de l'avancement du projet. Cette méthodologie permet d'impliquer le client et de travailler en agilité avec lui tout en renforçant la dynamique collective et l'implication de chacun.

Cette méthode agile telle que décrite permet de présenter régulièrement au client des versions potentiellement testables ou utilisables du projet. Elle permet une meilleure estimation des tâches, un meilleur suivi de l'avancement et une forte réactivité face aux imprévus. Le client restera joignable durant toute l'exécution du projet afin de répondre rapidement à de potentiels points bloquants.

Avant la répartition des tâches, la matrice RACI est réalisée pour bien comprendre les compétences et les rôles de chacun:

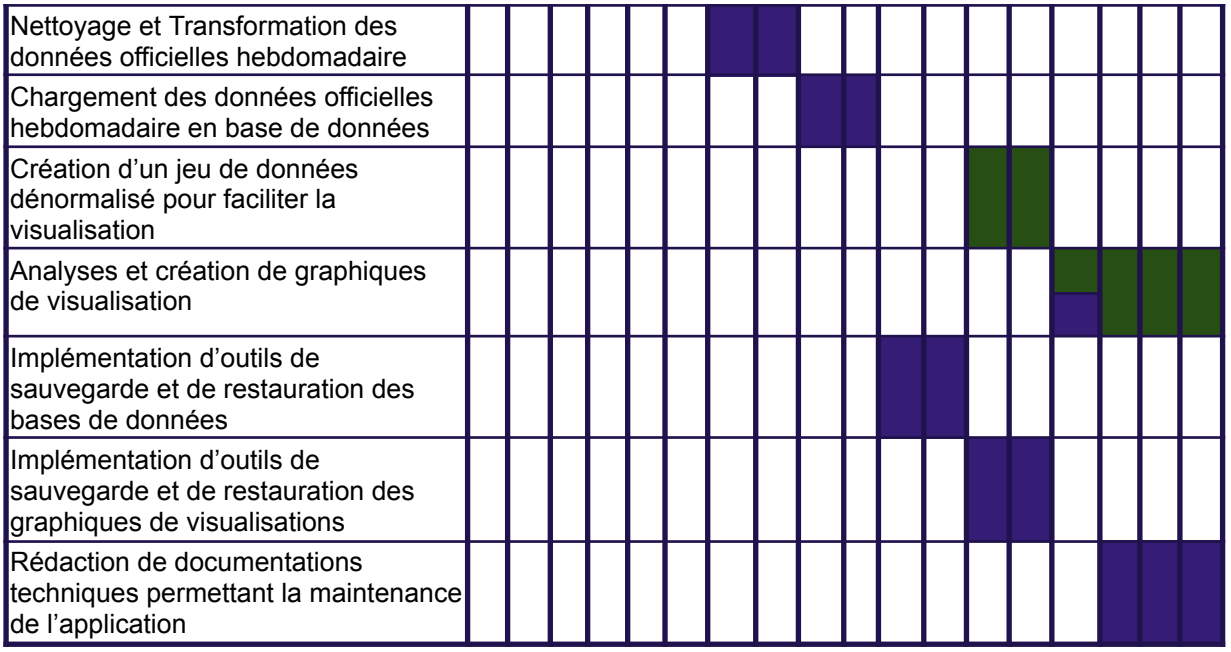
R	Responsable	Celui qui réalise concrètement la tâche, l'exécutant.
A	Approbateur	Celui qui est redevable du résultat, valide ou prend la décision finale.
C	Consulté	Celui qui est sollicité pour donner un avis ou un conseil.
I	Informé	Celui qui est tenu au courant de l'avancement ou du résultat.

Tâches	Data Engineer referent	Data Engineer 2	Référent métier du pôle data	Autres personnes du pôle data
Collecte du besoin et des attentes	R	C	A	C
Planification de la méthode Agile	R	A	A	I
Collecte et analyse initiale des données brutes journalières	A	R	C	I
Nettoyage et transformation des données brutes journalières	A	R	A	I
Chargement des données brutes journalières en base de données	A	R	C	I
Collecte et analyse initiale des données officielles hebdomadaire	R	A	C	I
Nettoyage et transformation des données officielles hebdomadaire	R	A	C	I
Chargement des données officielles hebdomadaire en base de données	R	A	C	I
Création d'un jeu de données dénormalisé pour faciliter la visualisation	A	R	C	I
Analyses et création de graphiques de visualisation	A	R	C	I
Implémentation d'outils de sauvegarde et de restauration des bases de données	R	A	A	I
Implémentation d'outils de sauvegarde et de restauration des graphiques de visualisations	R	A	A	I
Rédaction de documentations techniques permettant la maintenance de l'application	R	A	A	I

Le diagramme de Gantt est ensuite réalisé en prenant en compte la matrice RACI et les notes de complexité donné a chaque ticket lors du poker planning.

La couleur du **Data Engineer référent** sera représenté en Bleu, tandis que le **second Data Engineer sera représenté en Vert** sur le diagramme de Gantt.

Tâches	Sprint 1 (3 jours)	Sprint 2 (3 jours)	Sprint 3 (3 jours)	Sprint 4 (3jours)	Sprint 5 (3 jours)	Sprint 6 (3 jours)
Collecte du besoin et des attentes						
Planification de la méthode Agile						
Collecte et analyse initiale des données brutes journalières						
Nettoyage et Transformation des données brutes journalières						
Chargement des données brutes journalières en base de données						
Collecte et analyse initiale des données officielles hebdomadaire						



Si un collaborateur était handicapé, je ferais en sorte que l'équipe se rapproche du Référent handicap afin de mieux comprendre le handicap et recevoir des conseils et démarches nécessaires au bon accueil de la personne handicapée.

Je demanderais aussi des conseils directement à la personnes handicapée, puis je mettrais en place les recommandations exprimées afin d'aménager au mieux l'espace de travail ainsi que la méthodologie de travail au handicap de la personne.

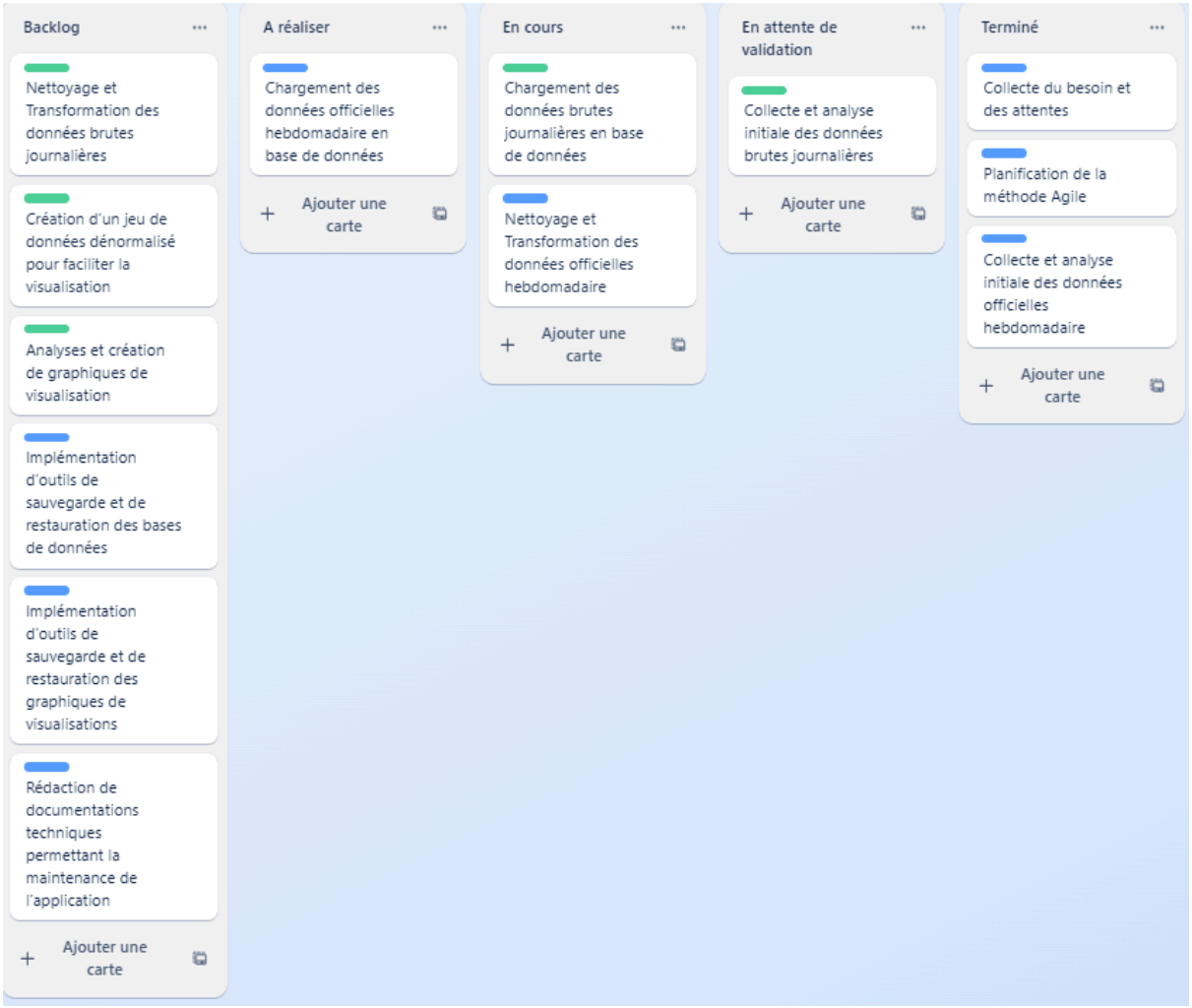
Par exemple, en cas de handicap de trouble de déficit de l'attention, je pourrais mettre un bureau à part supplémentaire afin que la personne puisse s'isoler lorsqu'elle en sent le besoin. Je lui laisserais aussi la possibilité de télétravailler lorsqu'elle le considère nécessaire, et je pourrais adapter le planning de charge à ses capacités.

C3.2.2 : Suivre l'avancement du projet en mettant en place un outil de suivi (logiciel de suivi, tableau de bord), en définissant les indicateurs (qualitatifs et/ou quantitatifs) pour chaque jalon défini dans le planning, en réalisant des reportings et des comptes rendus de réunion afin d'anticiper les aléas éventuels.

Les outils de suivi et de gestion de tickets les plus courants lors d'un développement de logiciel en méthode agile sont Azure, ou bien Jira.
Cependant, au vu de l'équipe réduite de 2 Data-Engineer, et de la dimension du projet a réaliser, le déploiement d'outils de suivi aussi lourds semble être démesuré par rapport au besoin de gestion et serait probablement contre-productif.

L'outil de gestion utilisé pour la réalisation de ce projet sera donc Trello. Il permet d'être facilement accessible, gratuit, et offre l'essentiel des fonctionnalités nécessaires pour le suivi et la gestion des tickets tels que la création de tickets, la prise de notes dans les tickets,

l’assignation, et le déplacement à travers différents états. De plus, c’est un outil collaboratif où plusieurs personnes peuvent y avoir accès.



Chaque couleur sur les tickets(vert et bleu) représente l’assignation de la tâche à un Data-Engineer.

À chaque début de sprint, les tickets qui devront être réalisés passent de la colonne “Backlog” à la colonne “À réaliser”. Les tickets changent ensuite de colonne suivant l’avancé de chaque collaborateur. En fin de sprint, nous accordons une importance à chaque ticket en cours de réalisation ou terminé. Nous évaluons ensuite pourquoi certains tickets censés être terminés ne le sont pas encore, et évaluons l’état d’avancement par rapport au diagramme de Gantt. Ainsi, les délais et les coûts de développement sont maîtrisés.

Le client peut aussi avoir accès au Trello afin qu’il ait un accès direct à l'évolution de son projet, ce qui améliore grandement sa compréhension et sa visibilité de l'avancement. Cela lui permet d’avoir un reporting de l’état actuel de l’avancement, et simplifie sa compréhension lors des comptes rendus de réunion à chaque fin de sprint.

C3.3.1 : Évaluer les besoins en compétences de l'équipe projet, en collaborant avec le service Ressources Humaines, en établissant un plan de développement des compétences et en orientant les membres de l'équipe vers des formations adaptées, afin de renforcer l'équipe responsable de mener à bien le projet DATA.

Voici les compétences à mobiliser ainsi que la grille d'évaluation des différents intervenants sur chacune des compétences à mobiliser sur ce projet:

Compétences	**Data-Engineer Référent**	**Data-Engineer 2**
Connaissance du travail en agile	Avancé	Débutant
Conception et architecture d'une solution	Avancé	Intermédiaire
Communication client	Avancé	Intermédiaire
Capacité à créer des ETLs	Avancé	Avancé
Capacité à créer des APIs	Avancé	Avancé
Capacité à créer des planificateurs de tâches	Intermédiaire	Avancé
Capacité à créer et gérer les droits de gestion de bases de données	Avancé	Avancé
Capacité à créer et gérer les droits de gestion d'un espace de stockage cloud	Avancé	Intermédiaire
Capacités d'analyses, de présentations de visualisation et de synthèses métier	Débutant	Avancé
Capacité à réaliser et déployer des architectures dockerisables	Avancé	Débutant
Capacité à faire des documentations	Avancé	Avancé

Nous pouvons constater que l'équipe des deux Data-Engineers se complète bien sur l'ensemble des compétences à mobiliser. Cependant, pour que le projet gagne en efficacité et pour assurer l'autonomie de chacun des membres, un plan de développement des compétences sera mis en place sur les différentes compétences manquantes de chacun d'entre eux.

Le Data-Engineer référent devra monter en compétence sur la capacité d'analyse, de présentation de visualisation et de synthèse métier.
Il devra donc réaliser une auto-formation sur la datavisualisation, d'une durée d'un jour en amont du projet, dans l'optique de renforcer ses connaissances sur la datavisualisation, la création de graphiques métiers pertinents et la réalisation de synthèses exploitables par le

métier. Le second Data-Engineer sera disponible et pourra partager son expérience personnelle en le cadrant durant cette auto-formation.

Le second Data-Engineer devra monter en compétence sur le travail en méthode agile.
Il devra pour cela suivre des vidéos théoriques explicatives de la méthode, puis faire un atelier Agile avec le Data-Engineer référent, pour une durée d'une demi-journée.
Il devra aussi monter en compétence sur sa capacité à réaliser et déployer des architectures dockerisables. Il devra pour cela réaliser une auto-formation sur le fonctionnement de Docker d'une durée d'une demi-journée. La conception du premier Dockerfile sera réalisée en pair programming avec le Data-Engineer référent dans une optique de montée en compétence.

Le plan de développement des compétences sera adapté suivant la capacité et les éventuels handicaps de chacun. Il sera possible de réaliser ces formations en télétravail, ou bien de prendre un temps supplémentaire si cela est nécessaire. Le matériel sera aussi adapté en fonction des besoins personnels.

C3.3.2 : Piloter l'équipe projet en affectant les missions à réaliser, en prenant en compte les spécificités des membres de l'équipe, en intégrant les spécificités d'un contexte multiculturel, international, en utilisant les différentes techniques de communication et d'animation managériale pour favoriser le bon fonctionnement de l'équipe.

La méthode Scrum en agilité possède dans sa méthodologie la capacité de prendre très souvent en compte les avis et opinions des collaborateurs.

Le **Poker Planning** permet de prendre en compte la difficulté et le temps nécessaire de chaque ticket, qui pourra ensuite être réparti équitablement dans le Gantt ainsi que dans l'outil de ticketing de gestion de projet.

Le **Daily** permet de suivre l'avancement des tickets, de remonter l'ensemble des problèmes rencontrés la veille, et offre une opportunité pour demander de l'aide si nécessaire.

Le **Rétroplanning** permet de discuter des points de difficultés et de blocage rencontrés durant le dernier sprint, de justifier des retards ou des avancements, et cela en complémentarité du Daily.

Ces différents outils qu'offre la méthode Scrum permettent d'améliorer la collaboration des membres de l'équipe projet, offrant une visibilité complète à chacun d'entre eux, afin de minimiser au mieux les potentiels tensions ou incompréhensions entre les collaborateurs.

Elle permet aussi de discuter, détecter et prendre en compte les spécificités et difficultés de chacun, telles qu'elles peuvent être rencontrées chez des personnes handicapées. Par

exemple, une personne n'ayant pas un poste de travail adapté à son handicap pourra facilement le faire remonter lors d'un Daily, pour que cela soit rapidement pris en compte et adapté dans les plus brefs délais.

C3.3.3 : Procéder aux arbitrages et aux réajustements nécessaires à partir de l'analyse des écarts entre le prévisionnel et l'état du projet à date, en utilisant des outils d'aide à la décision (ex : logigramme) afin de garantir la performance du projet dans le respect des objectifs de qualité, coûts et délai.

La collecte des besoins et des attentes avec le client a mis plus de temps que prévu et à retarder les réalisations prévues du premier sprint.
Malgré une volonté à essayer de rattraper le retard sur le deuxième sprint, les équipes n'y sont pas parvenues.

Un risque potentiel survient alors dans le fait d'essayer de rattraper constamment un retard qui se transmet de sprint en sprint et cela durant toute la réalisation du projet. Cela créerait un risque de fatiguer les équipes et de dégrader la qualité du travail réalisé dans l'optique de gagner du temps.

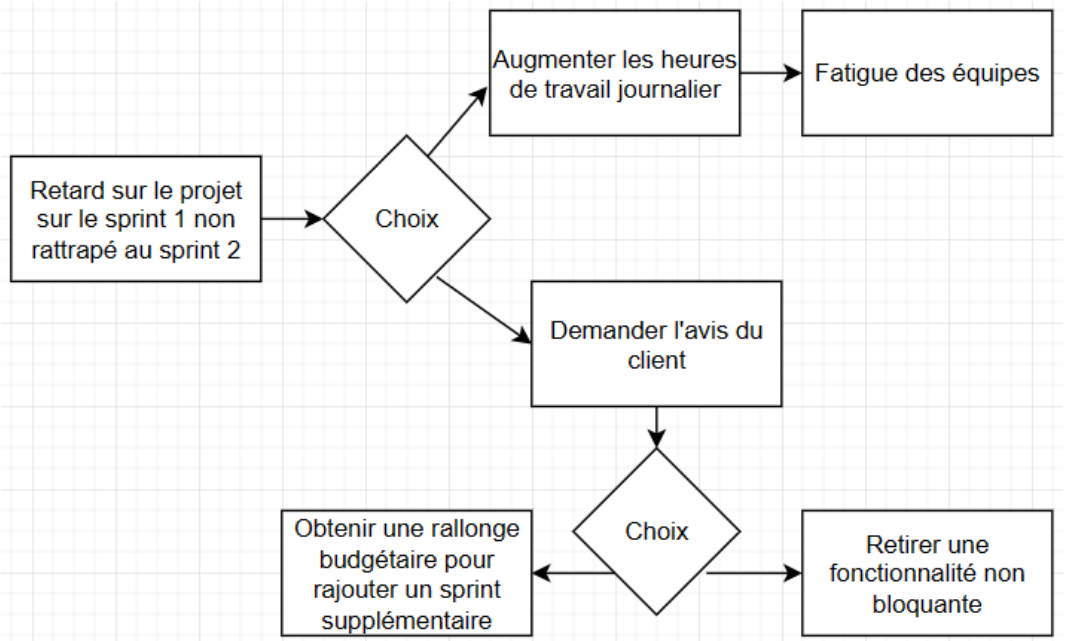
Afin de remédier à cela, plusieurs pistes sont envisagées:

- Augmenter les heures de travail journalier afin de rattraper le retard
- Demander une rallonge budgétaire au client afin d'obtenir un sprint supplémentaire de travail pour répondre à toutes les attentes du client.
- Retirer une fonctionnalité non bloquante pour le projet, qui pourra être rajoutée par la suite par l'équipe du pôle data du gouvernement si nécessaire.

La première solution consistant à augmenter les heures de travail est fortement déconseillée, car elle fait persister le risque de fatiguer les équipes et de dégrader la qualité du travail livré.

La solution la plus adaptée serait donc de profiter de la réunion de la fin du sprint 2 avec le client pour lui exposer les faits, et lui suggérer de prendre une décision entre la rallonge budgétaire pour un sprint supplémentaire(qui sera budgétisée en amont de la réunion), ou le choix de retirer une fonctionnalité non bloquante (une liste de fonctionnalités non bloquantes sera réalisée et fournie au client.) Il sera ainsi libre de prendre sa décision de manière éclairée.

Voici un logigramme permettant de schématiser cette situation:



C3.4.1 : Mettre en place un système de veille technologique et réglementaire en matière de science des données et d'Intelligence Artificielle à l'aide de recherches documentaires, de plateformes de partage, de webinars afin d'être alerté des évolutions qui impacteraient les pratiques métier.

La veille technologique est indispensable dans le domaine des sciences des données et de l'intelligence artificielle. En effet, les évolutions sont constantes, et il est fortement recommandé de prendre conscience de ces nouveaux outils afin de pouvoir répondre au mieux aux problématiques des clients.

Pour réaliser cette veille, plusieurs méthodes sont mises en place:

- La première méthode consiste à suivre des profils d'influenceurs sur Linkedin sur les sujets techniques qui nous intéressent. Il est aussi possible de rejoindre des groupes de partage sur des professions ou des sujets ciblés.

- La seconde méthode consiste à suivre des meetings et des podcasts sur ces sujets. Pour cela, certaines chaînes Youtube sont spécialisées sur certains sujets.

- Enfin, la troisième méthode consiste à se rendre à des salons ou à des présentations de veille. Ayant la chance d'habiter dans une grande ville, j'ai pu assister à des conférences présentées à la "Mêlée Numérique", un réseau d'acteurs du numérique basé à Toulouse.

Ces différentes méthodologies sont idéales pour être réalisées ensemble, afin de couvrir un périmètre plus grand et de pouvoir croiser les informations. En effet, le croisement des informations permet de vérifier la véracité de chacune des sources, et de pouvoir en tirer des conclusions éclairées.

Lors d'une des conférences de la "Mêlée Numérique" auxquelles j'ai assisté, le thème était l'IA et l'écologie dans les futures voitures connectées. Cette conférence m'a fait prendre conscience que l'intelligence artificielle, malgré sa capacité révolutionnaire à réaliser des tâches qui semblent complexes, est aussi très gourmande en énergie, ce qui ne va pas forcément dans le sens de l'écologie.

L'impact engendré sur mes pratiques métier m'a fait prendre conscience qu'il est préférable de privilégier des pipelines utilisant des algorithmes simples lorsque les cas d'usages sont possibles, plutôt que d'utiliser des pipelines utilisant des modèles de traitement de texte(LLM).
Cette bonne pratique permet de réduire la consommation énergétique, de réduire l'impact écologique, tout en s'inscrivant dans les engagements des normes RSE des entreprises.

C3.4.2 : Intégrer dans ses pratiques métier les enjeux en termes de données responsables, de responsabilité sociétale et environnementale (RSE), de sécurité, d'éthique et de confidentialité des données en se tenant informé des évolutions du cadre juridique, à travers une recherche documentaire ou en étant accompagné par des juristes afin d'agir dans le respect de la législation.

La responsabilité sociétale et environnementale des entreprises (RSE) s'applique dans tous les domaines de métiers, et principalement dans les métiers touchant la science des données au vu des consommations énergétiques que celle-ci peut impliquer ainsi que des données sensibles qui peuvent être traitées.

Les enjeux du RSE dans le domaine de la science des données consistent au fait de s'assurer que les projets soient Éthique, Responsable et Durable

Éthique dans le domaine où une IA ou bien un LLM ne doit pas contenir de biais qui pourrait causer des décisions injustes ou discriminantes.

Responsable sur les données qui sont récupérées et traitées, afin que celles-ci soient récupérées de manière légale et respectent les normes RGPD.

Durable dans une optique de préservation de l'environnement, en limitant les consommations énergétiques superflues qui pourraient être évitées.

En prenant en compte les différents enjeux du RSE dans le pilotage du projet data que nous sommes en train de réaliser, la part **Éthique** semble peu impactante sur cette application car le sujet reste purement technique, dans le domaine des prix des essences en France.

La part de **Responsabilité** sur les données travaillées est importante car nous devons prendre en compte la loi RGPD. Cependant, les données sont déjà anonymisées et en OpenData. Il faudra donc rester vigilant, mais le travail d'anonymisation a normalement déjà été fait en théorie. Une vérification devra quand même être faite.

C'est sur la partie **Durable** qu'il faudra le plus se concentrer car nous sommes sur des grosses quantités de données (39 millions de données pour le jeu de données des prix des essences issus de chaque station essence au format journalier). L'extraction et la transformation de ces données ont un coût énergétique relativement élevé et il faudra se concentrer sur une optimisation de cette partie du projet afin de limiter les émissions de CO2 causées par ces transformations.

L'action mise en œuvre pour améliorer la Durabilité du RSE consiste à ne pas charger l'intégralité des données chaque jour, mais à récupérer uniquement les données nécessaires à la mise à jour dans la limite de conception des sources de données d'où nous récupérons les informations.

L'estimation des coûts de cette action est négligeable. Elle consiste simplement à donner une attention particulière lors de la conception de l'algorithme d'extraction, de transformation et de chargement en bases de données des informations traitées. L'estimation serait donc à environ une demi journée de travail d'un Data-Engineer, où il se concentrera sur l'optimisation et l'architecture de sa pipeline de données afin de répondre aux critères voulus.

Les résultats attendus sont un temps d'extraction, de transformation et de chargement beaucoup plus rapide des mises à jours des données récupérées de manière quotidienne, en chargeant un minimum de données redondantes par rapport aux données déjà existantes sur la base de données.
Cela diminue la consommation d'électricité, préserve les data centers du pôle data du gouvernement qui fournit les données, et diminue la quantité de CO2 produite de manière journalières pour ces traitements.

Conclusion

J’ai beaucoup appris de la gestion de projet data que cela soit dans mes expériences passées, au sein de mon entreprise, au sein des projets réalisés à l'école supérieure, ou bien durant la réalisation de ce mémoire.

La gestion de projet est un univers complexe et passionnant, et je pense que tout projet d’envergure qui se déroule correctement commence avant tout par une gestion de projet viable et minutieuse.

Je vous remercie du temps que vous avez passé à lire mon mémoire.