Université Paris 1 Panthéon – Sorbonne

Mémoire de MASTER M2

SYSTÈMES D'INFORMATION ET DE CONNAISSANCE

Sous-parcours **Business Analysis**

Promotion 2023-2024

« ÉVALUATION ET AMÉLIORATION DE LA QUALITÉ DES DONNÉES ISSUES DU WEB SCRAPING DANS LE SECTEUR AUTOMOBILE »

RÉDIGÉ ET SOUTENU PAR : Thi Ha Trang NGUYEN

DIRECTEUR DE MÉMOIRE : Samuel PARFOURU

DATE DE SOUTENANCE : 01 octobre 2024

Tables des matières

Remerciements

Je dois l'achèvement de ce travail à l'accompagnement et au soutien de nombreuses personnes qui ont été présentes tout au long de mes recherches. La rédaction de ce mémoire symbolise pour moi une véritable réalisation personnelle. J'adresse mes remerciements à tous ceux qui ont contribué, directement ou indirectement, à ce projet.

Premièrement, ma gratitude va à madame Selmin NURCAN pour m'avoir offert l'occasion d'intégrer ce master. Sans mon acceptation initiale, ce travail n'aurait pas existé. Malgré un contexte exceptionnel, le déroulement de la formation a été optimal grâce à votre travail méticuleux sur l'organisation de ce programme.

De plus, j'exprime ma reconnaissance à l'ensemble de l'équipe pédagogique qui a enrichi mes connaissances. En particulier, mes remerciements vont à monsieur Samuel PARFOURU pour son immense bienveillance et sa disponibilité tout au long de l'élaboration de ce mémoire. Ses conseils m'ont permis de diriger correctement mes recherches et de finaliser ce travail.

Je souhaite également remercier monsieur Abderrahim KHELLOU et madame Khadidiatou NDIAYE pour ses accompagnements durant cette formation et pour m'avoir confié des missions. Ils m'ont aidé à progresser et à perfectionner mes compétences professionnelles. Toujours à l'écoute de mes demandes, ses conseils ont été d'une grande valeur et ont été une source de motivation pour mener à bien ce mémoire.

Enfin, je tiens à remercier mes proches et ma famille pour leur soutien moral dans les moments d'incertitude. Leurs encouragements lors de nos discussions m'ont permis d'y voir plus clair et de ne pas perdre de vue l'objectif de ce travail.

Abréviations

API : Application Programming Interfaces

CFAA: Computer Fraud and Abuse Act CFAA

DOM : Document Object Model

DQS : Data Quality Services

EDA : Analyse statistique exploratoire

ETL : Extraction, Transformation, Chargement

RGPD : Règlement Général sur la Protection des Données

SGBDR : Systèmes de bases de données relationnelles

UI : interface utilisateur

Liste de figures

Introduction générale

Dans le contexte de la transformation numérique rapide, les entreprises doivent gérer et utiliser les données de manière essentielle. L'arrivée de nouvelles technologies comme l'intelligence artificielle, l'internet des objets et le Big Data a profondément changé la façon dont les organisations fonctionnent. Cela offre de nouvelles opportunités pour améliorer les processus, optimiser les décisions stratégiques et augmenter la compétitivité. Cependant, ces progrès technologiques soulèvent également des défis importants, notamment concernant la qualité des données collectées et utilisées par les entreprises.

Les données sont aujourd'hui considérées comme l'un des actifs les plus précieux, car elles permettent aux entreprises de mieux comprendre leur environnement, d'anticiper les tendances du marché et de prendre des décisions éclairées. Cependant, il ne suffit pas d'accumuler de grandes quantités de données : il faut s'assurer que ces données sont précises, complètes et peuvent être exploitées. C'est dans ce contexte que la qualité des données prend toute son importance, en particulier pour les données issues du web scraping, une technique largement utilisée pour extraire des informations pertinentes à partir de sites web concurrents.

Ce rapport vise à explorer les techniques d'évaluation et de l'amélioration de la qualité des données issues du web scraping, en se concentrant particulièrement sur le secteur automobile. En basant sur mon expérience chez Renault Group, on examinera les diverses étapes requises pour s'assurer que les données collectées sont d'une qualité adéquate pour être utilisées dans des processus décisionnels essentiels.

Ce mémoire est structuré comme suit :

- **Chapitre I : Contexte et problématique**

Ce chapitre présente le contexte général du projet, avec une description de l'entreprise, des marques de Renault Group, et des services impliqués. Il explore les missions liées au web scraping et la problématique spécifique de la qualité des données scrappées. L'objectif du mémoire est également défini.

- **Chapitre II : État de l'art**

Ce chapitre traite du web scraping, ses concepts, son évolution, et ses principaux outils (BeautifulSoup, Scrapy, Selenium, etc.). Il aborde aussi les réglementations comme le RGPD et les droits d'auteur. Le chapitre se concentre ensuite sur la qualité des données scrappées, en identifiant les défis et les conséquences d'une mauvaise qualité. Il présente enfin des méthodes d'évaluation et d'amélioration des données, notamment via le nettoyage, la validation, et l'utilisation de frameworks comme Lean Six Sigma et des modèles de machine learning.

- **Chapitre III : Méthodologie**

Ce chapitre décrit la méthodologie employée pour évaluer et améliorer la qualité des données scrappées, avec une introduction aux différentes étapes du processus.

- **Chapitre IV : Proposition de solution et cas d'application**

Ce chapitre présente une approche pratique basée sur la méthodologie Lean Six Sigma pour identifier, mesurer, analyser, améliorer et contrôler la qualité des données scrappées. Des solutions spécifiques sont proposées et testées sur des cas concrets.

- **Chapitre V : Discussion et perspectives**

Ce dernier chapitre discute des résultats obtenus, des défis rencontrés et des implications de la solution proposée. Il ouvre également des perspectives pour des travaux futurs visant à optimiser davantage la qualité des données scrappées.

Chapitre I : Contexte et problématique

A. Contexte

1. Présentation de l'entreprise

1.1. Présentation générale

Fondée en 1898 par les frères Renault, l'entreprise Renault est aujourd'hui l'un des plus importants constructeurs automobiles au monde. Bien qu'elle ait débuté dans la production de voitures particulières, Renault s'est progressivement diversifiée, s'engageant également dans la fabrication de véhicules utilitaires, de camions, de tracteurs et de véhicules électriques. Reconnue pour son innovation, la marque Renault jouit d'une présence internationale solide.

Figure 1: technocentre de Renault Group

Tout au long de ses 125 années d'existence, Renault a su s'adapter aux évolutions du marché automobile mondial. Dès ses débuts, l'entreprise a fait preuve d'innovation, introduisant des avancées techniques telles que la boîte de vitesses à prise directe, et en remportant des courses qui ont contribué à sa renommée. Après la Seconde Guerre

mondiale, Renault a été nationalisée par le gouvernement français, ouvrant une nouvelle ère d'expansion industrielle et commerciale.

Dans les années 1980 et 1990, Renault s'est internationalisée en acquérant des parts dans des entreprises étrangères et en établissant des partenariats stratégiques, notamment avec Volvo et Nissan. Un tournant majeur a eu lieu en 1999 avec la création de l'Alliance Renault-Nissan, devenue l'une des plus grandes alliances de l'industrie automobile mondiale.

Renault a renforcé son influence mondiale en rejoignant l'Alliance. L'entreprise se distingue par son engagement en faveur de l'innovation, en particulier dans les domaines de la mobilité durable et de la transition énergétique. Avec des modèles emblématiques comme la Renault ZOE, le groupe est devenu un leader européen des ventes de véhicules électriques.

Renault s'engage également dans la conduite autonome et les services de mobilité connectée. En collaborant avec des entreprises technologiques, le groupe développe des solutions innovantes en matière de véhicules autonomes, de services de partage de véhicules et de gestion des flottes.

La stratégie de Renault repose sur trois axes principaux : l'électrification, la digitalisation et la croissance internationale. L'entreprise investit massivement dans la recherche et le développement pour rester à la pointe des technologies de l'automobile de demain.

Le Groupe Renault est présent dans 134 pays à travers le monde et exploite 36 usines, ce qui lui permet de produire près de quatre millions de véhicules par an. Le groupe emploie environ 180 000 personnes et génère un chiffre d'affaires de plus de 55 milliards d'euros (en 2023).

1.2. Les marques de Renault Group

Le portefeuille de marques de Renault inclut Renault, Dacia, Alpine et Mobilize. Chaque marque a un positionnement spécifique : Renault se concentre sur les véhicules de qualité destinés au grand public, Dacia offre des véhicules économiques, Alpine se spécialise dans les voitures de sport haut de gamme, et Mobilize se focalise sur les nouvelles solutions de mobilité.

Figure 2: les marques de Renault Group

Renault est également fortement implanté dans les pays émergents, notamment en Amérique latine, en Russie et en Afrique du Nord, où il adapte ses produits aux besoins locaux. La capacité du groupe à comprendre et à répondre aux besoins variés de ces marchés a été un facteur clé de sa croissance continue.

2. Présentation de services

Dans le cadre de mon alternance, je fais partie de l'équipe "Business Intelligence" qui est une composante de l'équipe "Data Analyst & Data Scientist". Cette équipe est rattachée à la "Direction Architecture & Data", elle-même intégrée à la "Direction de la Transformation Digitale". Cette dernière dépend de la "Direction Informatique & Digital", basée au Plessis-Robinson. La figure 1 illustre clairement cette structure hiérarchique.

Notre équipe compte 9 membres, incluant les data analysts, sous la responsabilité de Monsieur Abderrahim KHELLOU. Nous intervenons dans divers services tels que les finances, les ressources humaines, les achats, la fabrication, le pricing, etc. Au sein du département pricing, je travaille avec trois autres data analysts sous la supervision de Monsieur Cedric LAPIE. Nos activités se concentrent sur l'analyse des données relatives au pricing.

3. Présentation des missions

3.1. Analyser l'existant

Dans le cadre de mon alternance en tant que data analyst, j'ai été chargé d'une mission essentielle visant à améliorer la qualité des données collectées à partir de sites web de

la concurrence. Ma première tâche a consisté à analyser l'existant, c'est-à-dire les algorithmes et les règles déjà en place au sein de l'entreprise pour le traitement de ces données.

Plus précisément, une autre data analyst de l'équipe avait déjà développé un code Python dédié à l'estimation des prix des différentes versions de voitures proposées par nos concurrents. Mon rôle était de prendre en main ce code, de l'analyser en détail, et de comprendre toutes les règles qu'il applique. Ces règles concernent notamment la vérification de la qualité des données, telles que les prix, les modèles, et d'autres attributs associés aux véhicules.

L'analyse du code m'a permis de déchiffrer comment ces règles sont implémentées, en particulier celles qui assurent la cohérence et l'exactitude des informations collectées. Il était nécessaire pour moi de bien comprendre ces mécanismes afin de pouvoir évaluer leur efficacité et d'identifier d'éventuelles améliorations.

Après avoir étudié le code, j'ai procédé à son exécution pour en vérifier le bon fonctionnement. Au cours de cette étape, j'ai rencontré plusieurs erreurs. Ces erreurs étaient variées : certaines étaient liées à la logique du code, d'autres à des problèmes dans la manipulation des données. Chaque erreur m'a demandé de retracer sa source pour en comprendre la cause profonde.

Une fois les causes des erreurs identifiées, j'ai commencé de les corriger. Cela m'a parfois conduit à apporter des modifications au code existant, en ajustant des conditions logiques ou en affinant les règles de validation des données. Mon objectif était de m'assurer que ces corrections amélioraient la robustesse du code sans introduire de nouveaux problèmes.

Enfin, après avoir effectué les corrections nécessaires, j'ai réexécuté le code pour valider mes modifications et garantir que toutes les erreurs avaient été résolues. Cette étape de test a confirmé que les ajustements apportés avaient effectivement amélioré la qualité des données issues des sites web concurrents.

Ce projet a été l'occasion d'améliorer mes capacités d'analyse du code et de résolution de problèmes. Il m'a également permis de mieux comprendre les défis liés à la qualité des données dans un environnement concurrentiel.

3.2. Comprendre et analyser les données scrappées

Dans le cadre de mon alternance, je fais partie de l'équipe pricing, composée de 10 personnes basées en Inde. La mission principale de cette équipe est de scrapper les données de pricing provenant des sites web de nos concurrents. Ces données jouent un rôle crucial dans nos analyses, car elles nous permettent de suivre l'évolution des prix et des offres sur le marché.

Ma deuxième mission consiste à comprendre et analyser en profondeur les données que nous avons collectées. Cela implique de me familiariser avec les différents aspects des informations extraites, notamment les principes liés aux versions des véhicules, aux modèles spécifiques, et aux différentes structures de prix appliquées à chaque version. En particulier, je dois m'assurer de bien saisir les complexités des prix de vente, ainsi que des prix de leasing, qui peuvent varier en fonction des options et des configurations choisies.

Cette analyse me demande d'examiner les données sous plusieurs aspects pour identifier les tendances, les écarts et les opportunités qui pourraient être pertinentes pour nos stratégies de pricing. En comprenant ces principes fondamentaux, je contribue à améliorer nos analyses comparatives et à fournir des insights précis qui aident notre entreprise à se positionner de manière compétitive sur le marché.

3.3. Améliorer l'algorithme pour identifier les anomalies

Dans le cadre de ma troisième mission en tant que data analyst, j'ai été chargé d'améliorer l'algorithme existant afin de mieux identifier les anomalies et les incohérences dans les données que nous scrapons à partir des sites web concurrents. Ce travail est essentiel pour garantir que les données que nous utilisons pour nos analyses de pricing soient aussi précises et fiables que possible.

La première étape de cette mission a consisté à comprendre en profondeur l'algorithme déjà en place ainsi que la nature des données scrappées. Une fois que j'ai acquis une compréhension de ces éléments, j'ai commencé à identifier les règles qui permettraient de vérifier la qualité de ces données. Ces règles devaient couvrir divers aspects, tels que la détection de valeurs manquantes, d'incohérences entre les modèles de véhicules, les

versions, et les prix, ainsi que d'autres types d'erreurs qui pourraient fausser nos analyses.

Une fois ces règles définies, il était impératif de les faire valider par Madame Khadidiatou, une senior data analyst avec qui je travaille étroitement. Sa validation était importante pour s'assurer que les règles que j'avais établies étaient en adéquation avec les standards de qualité de notre équipe et qu'elles répondaient aux exigences spécifiques de notre projet.

Après avoir obtenu son approbation, j'ai poursuivi mon travail en intégrant ces règles directement dans le code Python, en utilisant la bibliothèque Pandas. Cette étape a nécessité non seulement de coder les règles de validation, mais aussi d'effectuer des ajustements continus pour optimiser l'algorithme et le rendre plus robuste face aux différents types de données que nous traitons.

Au cours de ce processus, j'ai rencontré de nombreuses erreurs et problèmes techniques. Certains de ces problèmes étaient liés à la complexité des données, tandis que d'autres étaient le résultat de l'intégration des nouvelles règles dans l'algorithme existant. J'ai dû analyser chaque erreur en profondeur, comprendre ses causes, et apporter les corrections nécessaires pour que l'algorithme fonctionne correctement.

Grâce à ce travail, j'ai non seulement amélioré l'algorithme, mais j'ai aussi renforcé mes compétences en résolution de problèmes et en manipulation de données avec Pandas. Cette mission a été un défi technique enrichissant qui a contribué à la fiabilité de nos analyses de données et, en fin de compte, à la qualité des décisions stratégiques prises par l'équipe pricing.

B. Problématique

La problématique de ce mémoire est présenté comme ci-dessous:

« Comment améliorer la qualité des données issues du web scraping sur les sites web concurrents dans le secteur automobile ? »

La décision de choisir cette problématique se base en plusieurs observations et nécessités spécifiques.

Tout d'abord, bien que l'état de l'art sur le web scraping et la qualité des données soit riche et diversifié, il traite souvent ces sujets de manière générale, sans se concentrer sur un secteur particulier. Cela limite la pertinence et l'applicabilité des solutions proposées lorsque l'on aborde des contextes spécifiques, tels que celui du secteur automobile. Ce secteur présente des caractéristiques uniques : les prix des véhicules, les promotions, et les configurations des modèles varient fréquemment, ce qui rend le web scraping particulièrement complexe. Les données extraites doivent être d'une qualité irréprochable pour permettre aux entreprises de réagir rapidement et de manière informée face aux actions de leurs concurrents.

En se concentrant sur le secteur automobile, cette étude vise à combler le manque de spécificité dans la recherche existante. L'objectif est d'adapter et d'améliorer les méthodes générales d'amélioration de la qualité des données pour répondre aux besoins et aux défis de ce secteur. Les décisions stratégiques dans l'automobile, comme les ajustements de prix, dépendent en effet de données précises et fiables.

Finalement, ce choix de problématique est motivé par la volonté d'apporter une contribution ciblée et utile, qui dépasse les approches généralistes en offrant des solutions adaptées aux défis spécifiques rencontrés dans l'industrie automobile.

C. Objectif du mémoire

L'objectif de ce mémoire est de résoudre un problème lié au web scraping, en particulier dans le secteur automobile. Ce secteur est très concurrentiel et la précision des données y joue un rôle stratégique. Le travail vise d'abord à explorer et analyser les méthodes existantes pour évaluer et améliorer la qualité des données obtenues par web scraping. En effet, les données collectées sur les sites web des concurrents peuvent souvent comporter des erreurs de format, des incohérences ou des lacunes, ce qui compromet leur fiabilité et leur utilité pour la prise de décision. Ce mémoire examine les différentes techniques de validation et d'amélioration de la qualité des données, comme le nettoyage et l'identification et correction des anomalies.

L'objectif est également de proposer des solutions concrètes adaptées au secteur automobile, où les données sur les prix des véhicules évoluent constamment et nécessitent une grande précision pour être vraiment utiles. En s'appuyant sur des études de cas pratiques et des exemples concrets, ce mémoire vise à offrir des recommandations que les entreprises automobiles pourront utiliser. Cela leur permettra d'améliorer la qualité de leurs données issues du web scraping, optimisant ainsi leurs stratégies commerciales et leur compétitivité sur le marché.

Chapitre II : État de l'art

1. Le web scraping : concepts et techniques

1.1. Définition et principes du Web scraping

1.1.1. Historique et évolution du web scraping

Le web scraping, ou extraction de données web, est une technique qui consiste à collecter automatiquement des données à partir de sites web. Son évolution est étroitement liée à l'histoire du web lui-même, ainsi qu'aux progrès technologiques qui ont permis de rendre cette technique plus sophistiquée et plus accessible (Glez-Peña et al., 2014). Dans ce mémoire, j'explore les étapes clés de l'évolution du web scraping, de ses débuts modestes à son rôle important dans l'industrie actuelle.

- **Les Débuts du Web Scraping (Années 1990)**

Avec l'apparition de World Wide Web dans les années 1990, la quantité d'informations disponibles en ligne a explosé (Berners-Lee et al., 1994). Cependant, l'accès à ces données était initialement limité par les capacités simples des moteurs de recherche et l'absence d'outils spécialisés pour extraire des informations spécifiques. Les premières tentatives de scraping étaient largement manuelles : les utilisateurs ont copié et ont collé des informations directement à partir des pages web.

Les premières méthodes automatisées ont vu le jour à la fin des années 1990. Les développeurs ont commencé à écrire des scripts en Perl ou en Python pour extraire des données de manière systématique à partir de pages web (Hemenway & Calishain, 2003). Ces scripts ont reposé principalement sur des expressions régulières pour identifier des motifs dans le code HTML. Cela permet de cibler et d'extraire des informations spécifiques, telles que des adresses e-mail ou des listes de produits.

- **L'Évolution des Techniques de Web Scraping (Années 2000)**

Les techniques de web scraping ont connu un développement rapide au début des années 2000. Avec Python devenant un langage de programmation web largement adopté et l'émergence de bibliothèques telles que BeautifulSoup, les développeurs ont maintenant un accès plus facile au web scraping (Mitchell, 2015). L'outil BeautifulSoup, plus précisément, a rendu l'extraction de données beaucoup plus facile en offrant une

navigation aisée dans la structure DOM (Document Object Model) des sites web. Cela facilite grandement l'extraction de données complexes.

Au cours de cette période, le web scraping a également commencé à cibler des sites web plus dynamiques qui utilisaient du JavaScript pour générer du contenu en direct. Les techniques de scraping qui se limitent à analyser le HTML ne sont plus suffisantes et il est nécessaire de développer de nouvelles méthodes capables d'interagir avec les scripts JavaScript. L'utilisation de navigateurs sans tête tels que PhantomJS a été développée pour permettre la manipulation automatisée et la capture des données dynamiques ainsi que l'interaction avec les pages web.

De plus en plus, les entreprises ont commencé à utiliser le web scraping de manière stratégique. Un exemple concret de l'utilisation du web scraping dans le secteur du commerce électronique est lorsque les entreprises s'en servent pour suivre en temps réel les prix pratiqués par leurs concurrents, ce qui leur permet ensuite d'adapter leurs propres tarifs de manière plus compétitive (Vargiu & Urru, 2013).

- **Sophistication et Législation (Années 2010 à aujourd'hui)**

Depuis les années 2010, le web scraping est devenu une pratique courante dans de nombreux secteurs industriels, soutenue par des outils avancés comme Scrapy, Selenium, et Puppeteer. Ces outils offrent des fonctionnalités plus robustes, notamment la gestion des cookies, l'interaction avec les formulaires, et le traitement des CAPTCHA, qui sont souvent utilisés pour protéger les sites web contre le scraping (Mitchell, 2018).

Cependant, cette montée en popularité a également soulevé des inquiétudes d'ordre légal et éthique. Un grand nombre d'entreprises ont commencé à adopter des mesures pour sécuriser leurs données en ligne, telles que la mise en place de systèmes anti-scraping et l'introduction de poursuites judiciaires contre les entités qui récupéraient illégalement le contenu de leurs sites. En 2018, l'entrée en vigueur du Règlement Général sur la Protection des Données (RGPD) a apporté une dimension de complexité supplémentaire en Europe. Cela signifie que les entreprises doivent être plus prudentes dans leur collecte et utilisation des données utilisateur (Voigt & Von dem Bussche, 2017).

Malgré ces défis, le web scraping continue de prospérer, soutenu par des technologies émergentes comme le machine learning et l'intelligence artificielle, qui permettent d'améliorer la qualité et la précision des données collectées (Zhao, 2017). Aujourd'hui,

le web scraping n'est pas seulement une méthode de collecte de données, mais un élément essentiel de nombreuses stratégies d'entreprise, permettant aux entreprises d'analyser des marchés, de surveiller la concurrence et d'extraire des insights précieux à partir de vastes quantités de données disponibles en ligne.

1.1.2. Les différents types de web scraping

Le web scraping existe plusieurs approches pour réaliser et adapter à des besoins spécifiques. Parmi les plus courantes, on trouve le HTML parsing, la manipulation du DOM (Document Object Model), l'interaction avec les API, le headless browsing, et l'utilisation de frameworks de scraping spécialisés.

- **HTML Parsing**

Le HTML parsing est l'une des méthodes les plus basiques et les plus courantes de web scraping. Elle consiste à télécharger le code source HTML d'une page web, puis à l'analyser pour extraire les informations souhaitées. Cette méthode est idéale pour des pages statiques où les données sont directement intégrées dans le HTML.

Fonctionnement : Un parseur HTML, tel que BeautifulSoup en Python, est utilisé pour transformer le code HTML en un arbre d'éléments. À partir de cet arbre, il est possible de naviguer et d'extraire des éléments spécifiques, comme des titres, des liens, des images, ou des paragraphes de texte.

Avantages : Cette méthode est simple à mettre en œuvre et fonctionne bien avec des pages web simples.

Limites : Elle est limitée aux pages statiques. Si une page charge du contenu dynamiquement via JavaScript, cette méthode ne sera pas en mesure de capturer ces données.

- **Manipulation du DOM**

La manipulation du DOM est une approche plus avancée qui permet d'interagir directement avec l'arborescence des éléments d'une page web. Cette méthode est souvent utilisée en conjonction avec des outils comme Selenium, qui permet de charger et manipuler le DOM dans un navigateur web (Gundecha, 2018).

Fonctionnement : Contrairement au HTML parsing, la manipulation du DOM permet de contrôler un navigateur réel (ou un navigateur sans interface utilisateur, appelé

headless browser). Cela permet de naviguer à travers les éléments d'une page, de cliquer sur des boutons, de remplir des formulaires, et d'extraire des données après que le JavaScript a été exécuté.

Avantages : Cette méthode est particulièrement utile pour scraper des sites dynamiques où le contenu est chargé après le chargement initial de la page.

Limites : L'utilisation de Selenium pour manipuler le DOM peut être plus lente et plus complexe à configurer. De plus, elle nécessite plus de ressources en termes de puissance de calcul.

- **Interaction avec les API**

L'interaction avec les API est l'une des méthodes les plus modernes et efficaces pour récupérer des données à partir de sites web. De nombreux sites proposent des API (Application Programming Interfaces) qui permettent aux développeurs d'accéder directement aux données sans avoir besoin de scraper l'interface utilisateur (Geewax, 2021).

Fonctionnement : Les API fournissent une interface programmée qui permet de demander des données spécifiques au serveur du site web, souvent sous forme de JSON ou XML. Cette méthode est beaucoup plus propre et structurée que le scraping direct des pages HTML.

Avantages : Les API sont conçues pour être utilisées par des programmes, ce qui les rend souvent plus stables et fiables que le scraping traditionnel. Elles permettent également d'accéder à des volumes de données plus importants et plus organisés.

Limites : Toutes les informations disponibles sur un site web ne sont pas nécessairement accessibles via une API. De plus, certaines API peuvent imposer des restrictions d'utilisation ou des limites de taux d'accès.

- **Headless Browsing**

Le headless browsing est une technique qui consiste à utiliser un navigateur web sans interface graphique pour effectuer des tâches de scraping. Cela peut être particulièrement utile pour scraper des pages web nécessitant une interaction complexe, tout en évitant les coûts de performance liés à l'affichage de la page (Jung et al., 2021).

Fonctionnement : Un navigateur headless, comme Puppeteer ou PhantomJS, est utilisé pour charger des pages web, exécuter du JavaScript, et interagir avec la page comme le ferait un utilisateur humain. Une fois la page entièrement chargée et manipulée, les données peuvent être extraites du DOM.

Avantages : Permet de scraper des sites web complexes avec une performance relativement élevée tout en restant invisible pour les utilisateurs.

Limites : La configuration et l'utilisation des navigateurs headless peuvent être plus complexes et nécessitent une bonne compréhension des interactions utilisateur.

1.2. Les outils de Web scraping

Différents outils de web scraping se démarquent par leurs caractéristiques et leur capacité à s'adapter divers types de sites web et besoins. Parmi les frameworks les plus utilisés figurent BeautifulSoup, Scrapy et Selenium. En plus de cela, il existe d'autres outils tels que Puppeteer, Octoparse et ParseHub qui offrent également des solutions intéressantes pour répondre à des scénarios spécifiques.

1.2.1. BeautifulSoup

BeautifulSoup est l'un des outils de web scraping les plus connus, principalement utilisé pour le parsing d'HTML et XML. Cette bibliothèque Python est appréciée pour sa simplicité et son efficacité dans l'extraction de données à partir de pages web statiques. (Richardson, 2023)

- **Fonctionnalités principales**

BeautifulSoup permet de transformer une page web en un arbre d'éléments que l'on peut facilement explorer et manipuler. Il est possible de rechercher des éléments spécifiques (comme des balises ou des classes CSS) et d'en extraire le contenu.

L'un des points forts de BeautifulSoup est sa capacité à gérer des documents HTML qui ne sont pas parfaitement structurés. Cela en fait un outil utile pour scraper des pages qui peuvent poser des problèmes avec d'autres bibliothèques plus strictes.

BeautifulSoup est souvent utilisé en conjonction avec des bibliothèques comme Requests pour récupérer le contenu des pages, ou Pandas pour structurer les données extraites en dataframes.

- **Avantages** : Facile à apprendre et à utiliser, excellent pour les projets simples, bonne gestion des pages HTML mal formées.
- **Limites** : Ne convient pas aux sites web dynamiques où le contenu est généré par JavaScript, performances limitées pour les grands volumes de données.

1.2.2. Scrapy

Scrapy est un framework Python complet dédié au web scraping et au crawling. Contrairement à BeautifulSoup, Scrapy est conçu pour des projets de scraping de grande envergure, avec une architecture qui facilite l'extraction, le traitement et le stockage de grandes quantités de données. (Scrapy, 2023)

- **Fonctionnalités principales**

Scrapy est structuré de manière à encourager une organisation modulaire du code. Les projets sont composés de spiders (des robots de scraping) qui définissent comment naviguer sur les sites et extraire les données.

Scrapy est conçu pour être rapide et efficace, capable de gérer des centaines de requêtes simultanées grâce à sa gestion asynchrone des requêtes HTTP.

La fonctionnalité suivante est pipelines de traitement des données. Une fois les données extraites, Scrapy permet de les traiter via des pipelines, qui peuvent inclure des étapes de nettoyage, de validation et de stockage des données.

Scrapy peut être étendu via des middlewares et des extensions pour adapter le framework aux besoins spécifiques d'un projet, comme la gestion de sessions ou le traitement de captchas.

- **Avantages** : Très performant, idéal pour des projets complexes avec des volumes de données importants, architecture modulaire et extensible.
- **Limites** : Courbe d'apprentissage plus raide que BeautifulSoup, nécessite une bonne compréhension des concepts de programmation orientée objet et de la gestion asynchrone des tâches.

1.2.3. Selenium

Selenium est un outil de test automatisé pour les applications web, mais il est également largement utilisé pour le web scraping, en particulier pour les sites qui génèrent du contenu via JavaScript. Contrairement à BeautifulSoup et Scrapy, Selenium contrôle un navigateur web réel, ce qui lui permet d'interagir avec les pages web de manière très similaire avec un utilisateur humain.

- **Fonctionnalités principales**

La première fonctionnalité est l'interaction avec le javascript. Selenium peut exécuter des scripts JavaScript, interagir avec des éléments dynamiques (comme des menus déroulants, des formulaires, etc.), et extraire des données après le rendu complet de la page. (SeleniumHQ, 2023)

La deuxième fonctionnalité est l'automatisation des navigations. Selenium permet de naviguer à travers plusieurs pages, de cliquer sur des boutons, de remplir des formulaires, et même de prendre des captures d'écran, ce qui le rend extrêmement flexible.

La dernière fonctionnalité est le support multi-navigateurs. Selenium peut être utilisé avec différents navigateurs web, tels que Chrome, Firefox, Safari, et Edge, offrant ainsi une grande polyvalence dans les environnements de test.

- **Avantages** : Capacité à gérer des pages web dynamiques, contrôle total du navigateur, permet de scraper des sites complexes avec des interactions utilisateur.
- **Limites** : Plus lent que Scrapy ou BeautifulSoup en raison de la surcharge liée à l'exécution d'un navigateur complet, plus complexe à configurer et à maintenir.

1.2.4. Puppeteer

Puppeteer est un outil de scraping relativement récent développé par Google. Il est basé sur le moteur de Chrome, Chromium, et permet d'automatiser des interactions utilisateur à travers un navigateur sans interface graphique (headless browser). Bien qu'il soit similaire à Selenium, Puppeteer est souvent considéré comme plus performant pour des tâches spécifiques. (Puppeteer, 2023)

Fonctionnalités principales

Premièrement, c'est le contrôle de chrome/chromium. Puppeteer permet de contrôler Chrome ou Chromium pour naviguer sur des pages web, remplir des formulaires, cliquer sur des éléments, et bien plus encore, le tout sans afficher d'interface graphique.

Deuxièmement c'est la fonctionnalité de screenshots et PDFs. Outre le scraping, Puppeteer peut capturer des captures d'écran ou générer des PDFs de pages web, ce qui est utile pour des tâches de reporting automatisé.

Ensuite, comme Selenium, Puppeteer peut manipuler le DOM après le rendu complet de la page, ce qui le rend idéal pour les sites dynamiques.

- **Avantages** : Plus rapide et plus léger que Selenium pour des tâches spécifiques, bon support pour les sites dynamiques, API moderne et bien intégrée avec Node.js.
- **Limites** : Limité au moteur de Chrome/Chromium, nécessite des connaissances en JavaScript/Node.js pour une utilisation efficace.

1.2.5. Octoparse et ParseHub

Octoparse et ParseHub sont des outils de web scraping basés sur des interfaces graphiques, conçus pour les utilisateurs qui préfèrent éviter la programmation. Ces outils permettent de scraper des données via un système de glisser-déposer, rendant le web scraping accessible à un public plus large.

- **Fonctionnalités principales**

Octoparse et ParseHub offrent une interface visuelle qui permet de configurer des tâches de scraping en cliquant simplement sur les éléments d'une page web. Pas besoin de compétences en programmation pour commencer. (Octoparse, 2023)

Ces outils peuvent aussi gérer des sites dynamiques et nécessitant des interactions utilisateur, bien que leur flexibilité soit inférieure à celle des outils basés sur du code.

Ils permettent d'avoir l'automatisation et programmation des tâches. Les utilisateurs peuvent planifier des tâches de scraping et automatiser des extractions de données à intervalles réguliers.

- **Avantages** : Facile à utiliser pour les non-programmeurs, rapide à mettre en place pour des tâches simples, pas besoin d'installation complexe.
- **Limites** : Moins flexible que des outils comme Scrapy ou Selenium, certaines limitations dans la personnalisation et l'automatisation avancée.

Le choix de l'outil de web scraping dépend des besoins spécifiques du projet. Pour des tâches simples et des pages statiques, BeautifulSoup est normalement suffisant. Pour des projets plus complexes qui sont nécessaires une gestion efficace des données et une performance élevée, Scrapy est recommandé. Selenium et Puppeteer sont indispensables pour scraper des sites dynamiques ou nécessitant des interactions utilisateur avancées. Bien que moins adaptés aux projets avancés, des solutions sans code telles qu'Octoparse et ParseHub permettent enfin à tous d'accéder facilement au web scraping. La croissance continue de ces outils illustre l'importance grandissante du web scraping dans divers domaines.Réglementations concernant le web scraping

1.3. Réglementations concernant le web scraping

1.3.1. Le Règlement Général sur la Protection des Données (RGPD)

Le RGPD, qui est l'une des législations les plus strictes en matière de protection des données personnelles, a été adopté en mai 2018. Le RGPD, qui est valable au sein de l'Union européenne, impose des règles strictes concernant la collecte, le traitement et le stockage des données personnelles. Même si le RGPD n'a pas pour objectif principal de réglementer le web scraping, il a cependant des répercussions significatives sur cette pratique, particulièrement lorsque des données personnelles sont collectées sans le consentement préalable des utilisateurs.

Selon le RGPD, les informations personnelles sont définies comme toutes données qui peuvent permettre d'identifier une personne physique. Lorsqu'une entreprise pratique le web scraping et récupère des informations personnelles telles que les noms, adresses e-mail ou autres données permettant l'identification d'un individu, elle doit se conformer aux obligations du RGPD. Cela comprend le fait d'obtenir explicitement le consentement des personnes concernées, de divulguer clairement l'utilisation des données et d'assurer une sécurité suffisant pour protéger ces informations.

Si le RGPD n'est pas respecté, il existe des sanctions strictes telles que des amendes qui peuvent aller jusqu'à 20 millions d'euros ou représenter 4 % du chiffre d'affaires annuel

mondial de l'entreprise, en fonction du montant le plus élevé. Il est donc important pour les entreprises qui envisagent de pratiquer le web scraping d'évaluer attentivement les types de données qu'elles collectent et de s'assurer qu'ils respectent les exigences du RGPD. (Albrecht, 2020).

1.3.2. Propriété Intellectuelle et Droit d'Auteur

Le web scraping pose également des questions liées à la propriété intellectuelle, en particulier au droit d'auteur. Le contenu des sites web, y compris les textes, images, et bases de données, est souvent protégé par le droit d'auteur. Cela signifie que l'extraction de ces données sans autorisation peut constituer une violation de la propriété intellectuelle.

Le droit d'auteur protège la façon dont les informations sont présentées, organisées et écrites sur un site web. Cependant, les données brutes comme les prix ou les statistiques ne sont pas protégées. Par exemple, un site web qui montre des listes de produits avec des descriptions détaillées peut être protégé par le droit d'auteur. Cela signifie qu'il est illégal de copier ou d'extraire ces descriptions sans autorisation. D'un autre côté, les faits ou données brutes comme les prix des produits ne sont pas protégés par le droit d'auteur. Cela est dû au principe selon lequel les faits ne peuvent appartenir à personne.

Certaines régions, comme l'Union européenne, offrent une protection spéciale aux bases de données. Cette protection donne aux créateurs de bases de données un droit exclusif d'utiliser et de réutiliser les données collectées, si la création de la base de données a nécessité un investissement important en termes de ressources humaines, techniques ou financières. Donc, le fait d'extraire des données de bases de données protégées pourrait violer ces droits et obliger l'entreprise à faire face à une action en justice. (Rosati, 2019).

1.3.3. Conditions d'Utilisation et Accès Non Autorisé

Les conditions d'utilisation des sites web sont un autre domaine juridique important qui peut affecter la légalité du web scraping. De nombreux sites incluent dans leurs conditions d'utilisation des clauses qui interdisent explicitement le scraping ou l'extraction automatisée de données. Bien que ces conditions soient souvent ignorées par les utilisateurs, elles peuvent être appliquées dans le cadre de litiges.

Aux États-Unis, le Computer Fraud and Abuse Act (CFAA) a été utilisé pour poursuivre des individus et des entreprises qui ont accédé à des systèmes informatiques ou des sites web en violation de leurs conditions d'utilisation. Par exemple, dans l'affaire United States v. Nosal, le CFAA a été invoqué pour punir l'accès non autorisé à des informations protégées par les conditions d'utilisation d'une entreprise. Toutefois, des décisions récentes, comme l'affaire hiQ Labs, Inc. v. LinkedIn Corp., ont limité l'application du CFAA dans les cas de scraping de données publiquement accessibles, bien que le statut juridique exact reste débattu (Streiff, 2020).

En dehors des États-Unis, d'autres juridictions ont des lois similaires pour protéger les systèmes informatiques contre l'accès non autorisé, et ces lois peuvent également s'appliquer au web scraping. Les entreprises doivent donc être conscientes des conditions d'utilisation des sites qu'elles scrappent et obtenir des autorisations lorsque cela est nécessaire pour éviter des conséquences juridiques.

Le cadre juridique entourant le web scraping est complexe et en constante évolution. Les réglementations telles que le RGPD, les lois sur la propriété intellectuelle, et les conditions d'utilisation des sites web créent un environnement où les entreprises doivent naviguer avec prudence. Pour minimiser les risques juridiques, il est crucial pour les entreprises de s'assurer qu'elles respectent toutes les lois applicables, d'obtenir des autorisations si nécessaire, et de consulter des conseillers juridiques pour évaluer les implications potentielles de leurs pratiques de scraping.

2. La qualité des données : définition et défis

2.1. Définitions de la qualité des données

La qualité des données est un concept multidimensionnel qui se réfère à la capacité des données à répondre aux besoins des utilisateurs et à l'objectif pour lequel elles ont été créées. Autrement dit, des données sont considérées de bonne qualité si elles sont "adaptées à l’usage" attendu, c'est-à-dire qu'elles répondent aux attentes et aux exigences de ceux qui les utilisent (Sebastian-Coleman, 2013). Ce principe de "pertinence à l'usage" est au cœur de la définition de la qualité des données.

La qualité des données est évaluée à travers plusieurs dimensions spécifiques, chacune mesurée par des indicateurs appelés métriques (Cichy et al, 2019). Ces dimensions sont interconnectées et permettent d'évaluer la valeur des données dans divers contextes. (Ehrlinger et al, 2022)

Précision : Cette dimension montre dans quelle mesure les données représentent fidèlement les valeurs réelles qu'elles sont censées représenter. Par exemple, si on collecte des prix de voitures sur des sites web concurrents, les prix extraits doivent être exacts pour être considérés comme de bonne qualité.

Pertinence : La pertinence concerne à l'adéquation des données aux besoins actuels et potentiels des utilisateurs. Une donnée pertinente est celle qui apporte une bonne réponse aux questions que l'utilisateur veut résoudre. Dans le cadre de la comparaison de prix, les données doivent être correspond aux besoins spécifiques de l'analyse, comme suivre les tendances de marché ou détecter des opportunités d'ajustement des prix.

Actualité : L'actualité, ou la fraîcheur des données, mesure le temps écoulé entre la collecte de l'information et l'événement qu'elle décrit. Par exemple, dans un secteur où les prix changent fréquemment, comme l'industrie automobile, il est important que les données soient récentes pour être valides.

Cohérence : La cohérence concerne à la capacité des données à être combinées de manière fiable à d'autres jeux de données sans contradictions. Cela signifie que, dans plusieurs contextes ou avec différentes sources, les données doivent rester uniformes et compréhensibles. Par exemple, si des prix différents sont collectés pour le même modèle de voiture sur plusieurs sites, ces différences doivent pouvoir être expliquées avec la façon logique (ISO 8000) (Benson, 2009).

2.2. Défis spécifiques à la qualité des données issues du web scraping

Le web scraping est une méthode qui permet d'extraire des données à partir de sites web. Cependant, cette méthode comporte plusieurs défis qui concernent la qualité des données. Ces défis sont augmentés par la diversité des sources en ligne, la structure des données et les limitations techniques. Voici quelques-uns des principaux défis que les utilisateurs de web scraping rencontrent.

2.2.1. Hétérogénéité des sources et des formats de données

L'un des défis majeurs du web scraping est la diversité des sources de données. Les sites web utilisent différentes structures pour afficher leurs informations, ce qui rend l'extraction et l'intégration des données très complexes. Certains sites web sont basés sur des formats tels que HTML, XML ou JSON, tandis que d'autres utilisent des contenus non structurés, comme des articles de blog ou des avis clients. Ces variations

nécessitent l'utilisation d'outils adaptés pour chaque type de source, augmentant le risque d'erreurs lors de la collecte des données.

En outre, les incohérences dans les formats de données collectées, par exemple les différences dans les unités de mesure ou la manière dont les dates sont exprimées, peuvent influencer la qualité des données. Les chercheurs doivent souvent recourir à des techniques de normalisation pour résoudre ce problème (Berti-Équille, 2006). Pour garantir une bonne intégration des données, il est essentiel de mettre en place des protocoles d'extraction qui identifient et corrigent ces anomalies.

2.2.2. La dynamique des données

Les sites web évoluent constamment, que ce soit par des mises à jour de contenu ou des modifications de leur structure. Cette nature dynamique rend difficile l'automatisation à long terme du web scraping. Un script qui fonctionne aujourd'hui peut ne plus marcher demain si un site change la manière dont il organise ses données. Cela crée des défis importants pour maintenir et surveiller les processus de collecte de données.

Pour réduire ces problèmes, il est recommandé de planifier des sessions de scraping régulières pour suivre les modifications fréquentes des sites, surtout pour les sites qui publient des données importantes comme les prix ou les informations de marché. Des techniques comme la rotation d'IP et l'utilisation de services pour contourner les CAPTCHAs peuvent aider à éviter les blocages, mais doivent être utilisées avec précaution pour respecter les réglementations des sites web (Breunig et al., 2000).

2.2.3. Absence de standardisation et ambiguïté des données

La collecte de données web peut être difficile à cause du manque de normalisation des formats de données. Les sites web, contrairement aux bases de données structurées, n'ont pas de normes universelles pour présenter les données. Cela peut entraîner des incohérences, surtout pour les données non structurées ou semi-structurées. Par exemple, une même information comme une adresse peut être présentée différemment selon les sites. Cela est difficile à les intégrer dans une base de données.

De plus, les données collectées peuvent contenir des erreurs ou des ambiguïtés, en particulier quand elles proviennent de sources générées par les utilisateurs (par exemple des avis ou des forums). Les erreurs de saisie, les fautes d'orthographe, les abréviations

et les symboles non standard peuvent affecter la précision des données et compliquer leur analyse. Pour résoudre ce problème, il faut mettre en place des procédures de nettoyage de données, telles que la déduplication, pour assurer leur fiabilité (Batini et al., 2004).

2.2.4. Gestion des volumes de données

Le scraping à grande échelle peut collecter de grandes quantités de données. Cependant, les outils de traitement des données ne sont pas toujours adaptés à ces volumes importants. Cela pose des défis, comme la gestion du temps de traitement, l'optimisation des performances et la réduction des erreurs dans les données collectées. De plus, la vérification de la qualité des informations extraites est une préoccupation.

Pour résoudre ces problèmes, il faut utiliser des outils adaptés, comme des plateformes d'extraction-transformation-chargement (ETL) qui facilitent l'analyse et le nettoyage des données à grande échelle. Des solutions de réduction de dimensions peuvent également être utilisées pour simplifier l'analyse des données volumineuses tout en minimisant la perte d'informations critiques (Caruso et al., 2000).

2.2.5. Validité temporelle des données

Un autre défi important du web scraping est la péremption rapide des données collectées. Certaines informations, comme les prix ou les stocks, peuvent devenir obsolètes en quelques heures, voire quelques minutes, selon les secteurs. Cette temporalité des données pose un problème majeur pour les utilisateurs qui nécessitent des informations en temps réel.

Il est donc crucial de maintenir la fraîcheur des données en mettant en place des processus d'actualisation régulière et en surveillant en continu les sources. Les techniques d'extraction continue (ou streaming) peuvent également être adoptées pour garantir que les informations collectées restent pertinentes et à jour (Batini et al., 2004)
.

La qualité des données issues du web scraping est influencée par une multitude de facteurs, allant de la diversité des sources à la nature dynamique des données. Pour surmonter ces défis, il est essentiel de mettre en place des protocoles stricts de gestion de la qualité, incluant des techniques de normalisation, de nettoyage et de surveillance

continue. En appliquant ces bonnes pratiques, les utilisateurs de web scraping peuvent maximiser la fiabilité et l'utilité des données collectées tout en minimisant les risques associés à la collecte de données à partir de sources hétérogènes.

2.3. Conséquences de la mauvaise qualité des données

La mauvaise qualité des données peut avoir des impacts considérables sur une organisation. Cela influence aux processus décisionnels, les opérations quotidiennes et la compétitivité d'une entreprise. Les principales conséquences se manifestent sur plusieurs plans : la perte de temps et d'argent, la perte de la confiance dans les systèmes et les données, et la prise de décisions erronées basées sur des informations inexactes.

- **Perte de temps et coûts financiers élevés**

La mauvaise qualité des données oblige les organisations à investir du temps et de l'argent pour corriger les erreurs. Cela entraîne des coûts supplémentaires importants. Certaines études montrent que la mauvaise qualité des données peut coûter des milliards aux entreprises dans le monde. Ces coûts incluent non seulement les ressources nécessaires pour identifier et corriger les erreurs, mais aussi les pertes financières dues à des décisions basées sur des données incorrectes (Berti-Équille, 2006).

Par exemple, dans un contexte commercial, les erreurs dans les données des clients peuvent entraîner des envois erronés, la perte de clients ou la dégradation des relations commerciales. Ne pas gérer la qualité des données dès le départ engendre donc des coûts directs et indirects. Cela pousse les entreprises à mettre en place des processus coûteux pour assurer l'intégrité des données tout au long de leur cycle de vie.

- **Décisions stratégiques biaisées**

Lorsque les décideurs utilisent des données de mauvaise qualité, ils risquent de prendre des décisions qui ne correspondent pas à la réalité de l'entreprise ou du marché. Des données incomplètes ou erronées peuvent cacher des tendances importantes. Cela peut entrainer une prise de décision qui n'est pas optimale ou correcte. Cela peut affecter des aspects critiques de l'entreprise, comme la tarification, le marketing et la gestion des stocks. (Batini et al., 2004)

Par exemple, dans une entreprise de vente au détail, des données inexactes sur les niveaux de stock peuvent entraîner des ruptures de stock ou un surstock, augmentant

ainsi les coûts de gestion. En outre, des prévisions basées sur des données de mauvaise qualité risquent d'influencer négativement aux objectifs financiers et aux stratégiques de l'entreprise.

- **Perte de confiance dans le système**

La mauvaise qualité des données peut faire perdre la confiance dans les systèmes d'information de l'organisation. Quand les utilisateurs remarquent des erreurs ou des incohérences fréquentes dans les données, ils deviennent réticents à utiliser ces systèmes pour leur travail. Cela entraîne non seulement de l'inefficacité, mais aussi des coûts cachés liés à la baisse de productivité. (Breunig et al., 2000).

De plus, la confiance des partenaires commerciaux ou des autorités peut aussi être affectée. Les entreprises doivent fournir des données précises et fiables pour respecter les exigences légales ou les attentes des parties prenantes. Un manque de conformité à cause de données inexactes peut mener à des sanctions légales ou des amendes, augmentant ainsi les risques financiers pour l'organisation.

- **Atteinte à la réputation de l'entreprise**

La mauvaise qualité des données peut également ternir la réputation d'une entreprise sur le marché. Les clients, les partenaires et les parties prenantes s'attendent à ce que les entreprises disposent de systèmes fiables qui permettent de prendre des décisions éclairées et précises. Lorsqu'une organisation ne parvient pas à répondre à ces attentes, elle risque de perdre la confiance de ses clients et de subir des dommages irréparables à sa réputation.

Dans le cas d'entreprises opérant dans des secteurs hautement régulés comme la finance ou la santé, la qualité des données revêt une importance encore plus cruciale. Des erreurs dans les données de patients, par exemple, peuvent entraîner des traitements médicaux incorrects, affectant ainsi directement la santé des individus. De telles situations peuvent non seulement entraîner des poursuites judiciaires, mais également des pertes importantes en termes de réputation et de crédibilité (Berti-Équille, 2006).

- **Complexité accrue dans la gestion des données**

La mauvaise qualité des données peut aussi rendre la gestion des systèmes d'information plus difficile. Lorsque des données incorrectes ou incohérentes circulent au sein de plusieurs systèmes, leur correction nécessite souvent une révision complète des

processus existants. Il faut identifier et corriger ces erreurs, parfois dans des environnements complexes où les données sont liées.

L'intégration défaillante de données de sources multiples peut créer des incohérences. Cela complexifie encore les analyses et l'utilisation de ces informations. Cela peut ralentir la mise en place de nouvelles initiatives commerciales ou freiner l'innovation technologique, car il faut d'abord améliorer la gestion des données.

La mauvaise qualité des données freine les entreprises, en termes de coûts, de temps, de prise de décision, de productivité, de réputation et de gestion des systèmes. Il est donc essentiel que les entreprises mettent en place des stratégies de contrôle qualité des données, de leur création à leur utilisation finale, pour limiter ces conséquences négatives.

3. Méthodes d'évaluation de la qualité des données

3.1. Techniques de validation des données

La validation des données est un processus essentiel pour assurer l'exactitude, la cohérence et la fiabilité des informations au sein d'un système d'information. Les techniques de validation peuvent être classées en plusieurs catégories, depuis des méthodes manuelles jusqu'aux systèmes automatisés utilisant des algorithmes complexes. Ces techniques sont appliquées à différentes étapes du cycle de vie des données. Ci-dessous sont présentées quelques-unes des principales techniques utilisées dans l'industrie.

- **Validation syntaxique et intégrité des données**

La validation syntaxique vérifie que les données respectent les formats prédéfinis, comme les types de données (nombres, texte, dates, etc.) et les règles spécifiques à chaque attribut (numéros de sécurité sociale, codes postaux, etc.). Elle vérifie également les règles d'intégrité comme l'unicité ou la non-nullité d'une clé primaire. Cela permet de s'assurer que les données respectent les règles du modèle relationnel et que les informations de la base de données sont cohérentes. (Berti-Équille, 2004).

Par exemple, les systèmes de bases de données relationnelles (SGBDR) utilisent des contraintes d'intégrité pour garantir que les données saisies sont cohérentes. Il est important que chaque clé étrangère d'une table soit liée à une clé primaire valide dans

une autre table. Cette vérification est cruciale pour s'assurer que les relations entre les tables de la base de données restent valides et non corrompues.

- **Contrôle statistique et analyse exploratoire des données**

L'analyse statistique exploratoire (EDA) est une technique utilisée pour détecter les anomalies dans un ensemble de données. Elle se base sur des résumés statistiques comme les moyennes, les écarts-types et les quantiles pour évaluer la distribution des données et identifier d'éventuelles irrégularités. Les représentations visuelles comme les histogrammes et les diagrammes de dispersion permettent de mettre en évidence les valeurs aberrantes ou isolées, qui peuvent indiquer des erreurs de saisie ou de transmission des données.

L'audit statistique des données permet aussi de vérifier si les distributions attendues des attributs spécifiques sont respectées, en utilisant des méthodes paramétriques comme les tests de normalité ou les tests du chi-carré. Ces techniques fournissent des indicateurs précieux pour déterminer si les données sont plausibles et conformes aux attentes initiales.

- **Audit des données et suivi des processus**

L'audit des données consiste à analyser systématiquement les enregistrements d'une base de données pour vérifier leur conformité avec les règles de gestion établies. Cette technique permet de détecter les erreurs et de suivre leur évolution dans le temps. L'audit utilise des méthodes comme la comparaison des données à des références externes ou des bases de données de référence. (Berti-Équille, 2004).

Le suivi des processus est une méthode complémentaire. Elle concerne à échantillonner et suivre les données au fur et à mesure de leur progression dans les chaînes de traitement d'information. Cela permet de vérifier l'intégrité et la cohérence des données à chaque étape, réduisant ainsi les risques d'erreurs lors des opérations.

- **Vérification par rapport à des référentiels externes**

La vérification par comparaison avec des sources externes ou des référentiels est couramment utilisée dans les systèmes d'information pour valider la qualité des données. Cette approche compare les enregistrements d'une base avec des bases de

référence externes reconnues, comme des répertoires professionnels ou des bases de données nationales.

Par exemple, dans le secteur des télécommunications, les données des clients sont souvent comparées à des bases de données de référence telles que le SIRET (répertoire des entreprises en France) pour assurer leur validité et leur actualité. Toutefois, cette méthode présente des limites, car elle repose sur la qualité des données de référence utilisées. Si ces dernières sont erronées ou obsolètes, le processus de validation peut conduire à de fausses corrections.

3.2. Comparaison avec des sources de données fiables

La qualité des données dépend souvent de leur source. Il est important d'utiliser des méthodes statistiques robustes pour évaluer et comparer ces données. Les données provenant de sources fiables comme les gouvernements ou les organisations internationales sont généralement plus cohérentes et précises que celles recueillies sur des plateformes collaboratives ou des sources non vérifiées. La mesure de la variance est une technique statistique couramment utilisée pour évaluer la qualité des données de différentes sources.

Les sources de données ouvertes, comme les gouvernements ou les organisations internationales, assurent un certain contrôle dans la collecte et la gestion des données. Les chercheurs s'y réfèrent souvent car elles suivent des protocoles rigoureux. Cependant, dans le cas des données géographiques collectées par des volontaires (VGI), une étude menée par Senaratne et al. (2017) a montré que ces contributions, bien que utiles, présentent plus d'erreurs en raison du manque de supervision.

Les données publiques doivent être complètes et cohérentes. Un cadre de travail a été proposé pour comparer des ensembles de données provenant de différentes sources (Vetrò et al., 2016). Cela permet de détecter les erreurs et d'identifier les incohérences. Ces méthodes renforcent la transparence et la véracité des données collectées.

La mesure de la variance est un outil clé pour comparer des données provenant de plusieurs sources. En analysant la dispersion des données, on peut déterminer si elles sont homogènes ou hétérogènes. Une faible variance indique une bonne cohérence entre les sources, ce qui est un bon signe de fiabilité. Une variance élevée signale des

différences importantes, ce qui peut poser une question de la qualité ou l'exactitude des données. (Gudivada et al., 2017).

Les techniques d'audit sont une méthode efficace pour évaluer la qualité des données. Elles consistent à vérifier la cohérence des données en les comparant aux documents sources. Cela permet d'identifier les divergences et de s'assurer que les données reflètent correctement les informations qu'elles représentent (Roos et al., 1982). Dans le domaine de la santé, les audits ont révélé des différences importantes entre les informations médicales collectées par les hôpitaux (Iezzoni, 1997).

Pour obtenir des données fiables, il faut utiliser plusieurs sources et appliquer des techniques de vérification croisées. Comparer des jeux de données de différentes sources peut aider à détecter les erreurs ou les informations manquantes. Les méthodes de collecte de données ouvertes telles que celles proposées par Vetrò et al. (2016), recommandent aussi de valider les données en les comparant à plusieurs jeux de données.

Pour garantir la fiabilité des résultats, il est essentiel d'utiliser des méthodes efficaces d'évaluation de la qualité des données. La variance statistique et les techniques d'audit sont des outils utiles pour s'assurer que les données sont précises et complètes. Cela aide les chercheurs et les décideurs à vérifier que les informations utilisées sont fiables, pertinentes et adaptées à leurs situations particulières.

3.3. Outils et frameworks pour l'évaluation de la qualité

L'évaluation de la qualité des données est un domaine clé dans les systèmes d'information, la gestion des bases de données et l'analyse des grandes quantités de données (Big Data). Les outils et frameworks dédiés à cette tâche permettent non seulement d'identifier les erreurs et incohérences dans les données, mais aussi de les corriger, de les nettoyer et de s'assurer de leur validité pour un usage fiable

3.3.1. Selenium : Automatisation et tests de qualité

Selenium est un outil puissant pour l'automatisation des tests de navigation et d'interaction avec les pages web. Utilisé principalement pour le web scraping et le test d'applications web, Selenium se distingue des autres outils par sa capacité à interagir

avec des pages dynamiques générées par JavaScript, souvent inaccessibles aux techniques traditionnelles de scraping.

L'un des principaux avantages de Selenium est sa capacité à exécuter des tests de qualité de manière automatique, en simulant des interactions humaines avec une interface utilisateur (UI). Cela permet non seulement de valider l'extraction des données, mais aussi de s'assurer que les informations extraites sont exactes et complètes. Cette approche est utile dans les scénarios où les données peuvent être soumises à des erreurs humaines ou à des anomalies dans le processus d'interaction avec le site web.

- **Utilisation de Selenium pour interagir avec des pages dynamiques**

Contrairement à des outils comme BeautifulSoup, qui extraient des informations à partir du code HTML statique, Selenium est capable de gérer les pages dynamiques, c'est-à-dire celles qui sont générées ou mises à jour en temps réel à l'aide de technologies comme JavaScript ou Ajax. Cela permet d'accéder à des données souvent invisibles pour les simples parsers HTML.

Par exemple, dans le cadre du scraping de données de formulaires en ligne, où une page peut se rafraîchir ou des éléments peuvent apparaître après une interaction (comme le choix d'une option dans un menu déroulant), Selenium peut être configuré pour simuler ces actions et récupérer les résultats à chaque étape.

Dans ce cadre, Selenium permet également de tester les données en fonction des actions réalisées. Si un formulaire est rempli, Selenium peut vérifier automatiquement si les informations affichées après la soumission du formulaire correspondent aux données attendues. Cette validation est essentielle pour s'assurer que les processus de scraping ne contiennent pas d'erreurs, comme des champs mal formatés ou des informations manquantes.

- **Validation de l'intégrité des données avec Selenium**

L'un des aspects les plus intéressants de Selenium pour l'évaluation de la qualité des données est la capacité à vérifier l'intégrité des données extraites. Par exemple, si l'on scrape des formulaires remplis par des utilisateurs ou des transactions financières sur un site, Selenium peut être configuré pour vérifier que les informations collectées sont exactes et correspondent aux données saisies ou attendues. Cela implique des comparaisons directes entre les informations affichées sur la page web et les résultats de l'extraction.

En outre, Selenium permet de gérer les cas d'erreurs potentiels. Par exemple, lorsqu'une page web génère un message d'erreur ou si une soumission de formulaire échoue, Selenium peut être configuré pour identifier ces erreurs, enregistrer les logs et assurer un retour d'informations immédiat au scraper. Cela aide à maintenir une haute qualité des données en évitant la propagation d'erreurs dans les systèmes en aval.

3. Tests fonctionnels et qualité des données

Selenium va au-delà de la simple extraction des données ; il est également utilisé pour exécuter des tests fonctionnels automatisés. Ces tests permettent de s'assurer que le processus de collecte des données fonctionne correctement à chaque étape du scraping. Par exemple, dans le cadre d'un projet de scraping pour un site de commerce électronique, Selenium pourrait être utilisé pour vérifier la disponibilité des produits, simuler l'achat d'un produit, et ensuite valider que les données du panier et les informations de commande sont exactes.

Ces tests fonctionnels garantissent que l'intégrité du processus de scraping est maintenue. Si à n'importe quel moment les informations sont manquantes ou mal collectées, Selenium renverra un message d'alerte, ce qui permet aux développeurs de réagir rapidement et de corriger les erreurs avant que celles-ci n'affectent la qualité globale des données. De cette manière, Selenium devient un outil critique dans les environnements où des données de haute qualité sont indispensables, par exemple pour des analyses de marché ou des études concurrentielles.

- **Avantages de Selenium pour l'évaluation de la qualité des données**

Les principaux avantages de l'utilisation de Selenium pour l'évaluation de la qualité des données incluent :

- **Capacité à traiter des pages dynamiques** : Selenium peut interagir avec des éléments générés dynamiquement, ce qui en fait un outil indispensable pour le scraping des sites modernes utilisant des technologies comme JavaScript ou Ajax.
- **Tests automatisés** : Selenium permet de simuler des scénarios utilisateur complexes et d'automatiser les tests de qualité des données à chaque étape du processus de collecte.

- **Vérification de l'intégrité des données** : Selenium peut comparer les informations extraites aux données attendues ou saisies, garantissant ainsi la précision des résultats.
- **Rapport d'erreurs** : Lorsque des anomalies surviennent lors du scraping (par exemple, des erreurs de chargement de page ou des informations manquantes), Selenium peut générer des rapports d'erreurs détaillés et alerter les équipes techniques.

- **Limites et défis**

Bien que Selenium soit un outil puissant pour l'automatisation et les tests de qualité, il présente quelques défis. D'une part, il nécessite un environnement de navigateur réel (comme Chrome ou Firefox) pour simuler la navigation, ce qui peut entraîner des contraintes en termes de performance et de temps d'exécution, notamment lorsque de nombreuses pages doivent être traitées. De plus, Selenium peut être bloqué par certains sites web qui détectent les bots, ce qui nécessite des ajustements pour éviter que le scraping ne soit interrompu.

3.3.2. Frameworks personnalisés pour l'évaluation de la qualité des données

En effet, dans des secteurs tels que le commerce électronique, la finance ou encore la recherche scientifique, les données extraites peuvent provenir de plusieurs sources, ce qui augmente la complexité de leur traitement (Khder, 2021). C'est dans ce contexte que l'utilisation de frameworks personnalisés pour évaluer la qualité des données est devenue primordiale. Ces frameworks sont conçus pour s'adapter aux besoins spécifiques des utilisateurs et permettent une évaluation plus fine de la fiabilité, de la précision et de la cohérence des données collectées.

- **Personnalisation en fonction des besoins spécifiques**

Un framework personnalisé est une solution sur mesure développée pour répondre à des exigences précises dans la collecte et l'évaluation des données. Contrairement aux outils génériques, ces frameworks sont configurés pour évaluer des critères de qualité spécifiques en fonction des données collectées (Pandey et al., 2020). Par exemple, dans le secteur du commerce électronique, les prix des produits ou les avis clients changent fréquemment. Un framework personnalisé peut surveiller la fréquence des mises à jour des données afin de s'assurer que celles-ci restent à jour et pertinentes. Il peut également

vérifier la fiabilité des sources en évaluant leur historique et leur fréquence de mise à jour, afin d'exclure les sources peu fiables.

Dans le secteur du commerce électronique, l'utilisation de frameworks personnalisés permet de gérer les informations provenant de plusieurs sources telles que les plateformes de vente en ligne et les avis clients. Un cadre personnalisé peut comparer les prix des produits entre différents sites web pour s'assurer de la cohérence des informations et signaler les écarts lorsque des prix incohérents sont détectés (Khder, 2021). De plus, ces frameworks peuvent inclure des règles spécifiques pour identifier les informations non pertinentes, telles que les avis générés automatiquement ou les faux avis.

- **Vérification de la qualité des données**

Un autre avantage des frameworks personnalisés est leur capacité à vérifier la qualité des données en temps réel. Cela inclut la détection de problèmes comme les incohérences entre les sources, les données incomplètes ou les données redondantes (Pandey et al., 2020). En cas d'anomalies, le framework peut envoyer des alertes ou corriger automatiquement certaines erreurs. Par exemple, dans le domaine financier, la collecte de données sur les marchés boursiers en temps réel peut inclure des informations provenant de plusieurs plateformes avec des légers décalages temporels. Un framework personnalisé peut identifier ces écarts et s'assurer que les données utilisées sont cohérentes avant qu'elles ne soient intégrées dans les modèles d'analyse prédictive (Meschenmoser et al., 2016).

Dans le secteur financier, les données des marchés boursiers et des transactions bancaires doivent être exactes et à jour pour garantir la validité des analyses. Les frameworks personnalisés peuvent surveiller les données en temps réel et comparer les informations entre plusieurs bases de données. Lorsque des écarts sont détectés, le framework peut intervenir pour corriger automatiquement les erreurs avant qu'elles n'impactent les résultats financiers (Pandey et al., 2020). Cette vérification en temps réel garantit que les données utilisées dans les décisions commerciales sont fiables.

- **Cohérence entre plusieurs sources**

Un des défis majeurs dans le web scraping est de garantir la cohérence des données entre plusieurs sources. Les frameworks personnalisés permettent de détecter les incohérences potentielles entre des sources multiples, d’analyser les divergences et de

normaliser les données (Khder, 2021). Par exemple, lorsqu'une même donnée est extraite de plusieurs sites, comme dans le cas de la surveillance des prix des produits, un framework personnalisé peut analyser les variations de données, signaler les écarts et suggérer des corrections. Cela permet d'améliorer la fiabilité des analyses en garantissant que les données sont uniformes.

Dans le secteur de la logistique, les données provenant de plusieurs plateformes, comme la disponibilité des navires pour le transport maritime, doivent être cohérentes pour permettre une planification efficace. Les frameworks personnalisés permettent de vérifier que ces données sont exactes et synchronisées entre les différentes plateformes afin d'éviter les erreurs dans la gestion des ressources (Meschenmoser et al., 2016).

- **Technologies sous-jacentes aux frameworks personnalisés**

Les frameworks personnalisés utilisent souvent des technologies avancées comme l'apprentissage automatique, les algorithmes de détection d'anomalies, et les modèles prédictifs pour évaluer la qualité des données (Pandey et al., 2020). L'utilisation de ces technologies permet aux frameworks de s'adapter aux changements dans les données et d'apprendre en continu. Par exemple, des algorithmes d'apprentissage automatique peuvent être utilisés pour identifier des schémas de données anormaux, tandis que des modèles prédictifs peuvent être utilisés pour estimer la qualité des données futures en fonction de leur historique.

Les frameworks personnalisés offrent une solution flexible et efficace pour l'évaluation de la qualité des données dans des secteurs où les besoins spécifiques exigent des niveaux élevés de précision et de cohérence. En intégrant des outils avancés pour la vérification, la correction automatique et l'analyse des données, ces frameworks permettent aux entreprises de garantir la fiabilité des informations avant leur utilisation dans des processus décisionnels critiques (Khder, 2021). Le développement de ces frameworks sur mesure est donc essentiel pour maximiser l'efficacité des systèmes de collecte de données à grande échelle et pour s'assurer que les décisions prises à partir de ces données reposent sur des informations de haute qualité.

4. Techniques et méthodes d'amélioration de la qualité des données

4.1. Nettoyage et prétraitement des données

Le nettoyage et le prétraitement des données sont des processus indispensables pour garantir l'intégrité et la qualité des informations utilisées dans divers systèmes.

4.1.1. Méthodes de nettoyage des données

Le nettoyage des données est un processus qui vise à éliminer les erreurs, corriger les incohérences et garantir la qualité des informations dans un ensemble de données. Les principales méthodes incluent (Berti-Équille, 2004) :

- **Détection et élimination des doublons**

Les doublons sont des données redondantes, souvent issues de la collecte de données multiples ou de l'intégration de plusieurs sources. Ils peuvent fausser les analyses et les résultats. Des algorithmes spéciaux sont utilisés pour détecter et supprimer ces enregistrements répétés. Ces algorithmes comparent différents champs comme les noms, adresses ou numéros d'identification pour trouver les doublons potentiels. Des techniques d'appariement d'enregistrements permettent de regrouper les données liées au même objet, mais provenant de sources différentes.

- **Correction des erreurs typographiques et syntaxiques**

Les erreurs de saisie manuelle (comme les fautes de frappe) et les formats incohérents (dates mal formatées, unités monétaires incorrectes) sont fréquents. Le nettoyage consiste à standardiser ces formats en appliquant des règles de normalisation (ex : s'assurer que toutes les dates sont au format "JJ-MM-AAAA"). Des outils de correction peuvent détecter ces anomalies et proposer des corrections automatiques basées sur des règles ou des dictionnaires.

- **Imputation des valeurs manquantes**

Les valeurs manquantes peuvent avoir un impact négatif sur les analyses, parce qu'elles introduisent des biais dans les résultats. Des techniques d'imputation permettent de remplacer ces données manquantes par des estimations raisonnables, telles que la moyenne d'une colonne, la médiane, ou encore des prédictions issues de modèles

statistiques. Ces approches garantissent que l'ensemble des données est complet avant toute analyse.

- **Détection des valeurs aberrantes**

Les valeurs aberrantes sont des observations très différentes des autres données. Cela peut être dû à des erreurs lors de la saisie ou de la collecte des données. Des techniques statistiques, comme l'analyse de la distribution des données, permettent d'identifier ces valeurs aberrantes. Elles peuvent ensuite être corrigées ou supprimées selon leur nature.

4.1.2. Méthodes de prétraitement des données

Le prétraitement est une étape nécessaire qui permet de préparer les données pour une utilisation efficace. Voici quelques-unes des méthodes les plus couramment utilisées (Berti-Équille, 2004) :

- **Normalisation des données**

Les données collectées peuvent avoir des formats différents, comme des devises, des unités de mesure ou des dates variées. La normalisation est le processus qui permet de convertir ces données dans un format cohérent et standardisé. Cela facilite leur traitement et leur intégration dans des systèmes communs. Par exemple, les données financières peuvent être converties dans une seule monnaie, et les dates peuvent être mises dans un format standard comme "JJ-MM-AAAA".

Ce processus est particulièrement utile lors de l'intégration de données provenant de plusieurs bases de données. Une fois les données normalisées, elles deviennent plus faciles à analyser et à exploiter, car les incohérences dues aux différences de format sont éliminées.

- **Filtrage des données**

Le filtrage des données permet de retirer les données inutiles ou incorrectes. Ce processus est souvent utilisé pour supprimer les observations incomplètes ou erronées qui pourraient biaiser les résultats des analyses. Par exemple, si certaines données présentent trop de valeurs manquantes ou des valeurs aberrantes (trop éloignées des valeurs attendues), elles peuvent être exclues des analyses.

Les techniques de filtrage des données incluent également la suppression des doublons, où plusieurs enregistrements identiques sont éliminés, ou la suppression des valeurs

aberrantes, qui sont des données qui s'écartent considérablement des autres valeurs dans le jeu de données.

- **Transformation des données**

La transformation des données consiste à modifier la structure des données brutes pour les rendre compatibles avec les outils d'analyse. Ce processus inclut la modification des formats de données ou la réorganisation de celles-ci pour mieux répondre aux besoins des systèmes d'information.

Par exemple, on peut diviser une adresse complète en plusieurs champs distincts (rue, ville, code postal, pays). Cela permet des analyses géographiques plus précises, comme l'identification des tendances par région. D'autres transformations peuvent inclure l'agrégation des données (par exemple, passer du quotidien au mensuel pour les ventes) ou la conversion des unités de mesure (par exemple, de miles en kilomètres).

4.1.3. Outils de nettoyage et de prétraitement des données

- **Outils ETL (Extraction, Transformation, Chargement)**

Les outils ETL permettent d'extraire des données brutes de différentes sources, de les transformer selon des règles prédéfinies, puis de les charger dans d'autres systèmes. Ces outils sont très utiles pour les entreprises qui gèrent de grands volumes de données provenant de plusieurs systèmes. Ils permettent d'uniformiser les formats, de corriger les erreurs et de transformer les données selon les besoins analytiques.

Talend : Outil open source offrant une grande flexibilité dans la création de processus ETL, avec des fonctions avancées pour la transformation et le nettoyage des données (Talend, 2006)

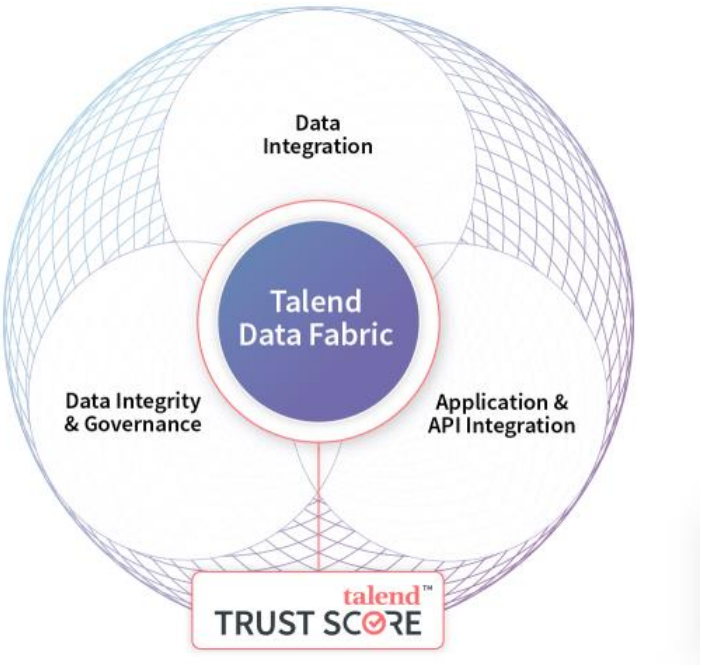

Figure 3: Les fonctions de Talend (Talend, 2006)

Informatica : Un des outils les plus utilisés dans les grandes entreprises, offrant une large gamme de fonctionnalités d'intégration et de gestion des données, y compris des capacités avancées de nettoyage (Informatica, 2011).

Apache Nifi : Solution open source conçue pour l'automatisation des flux de données avec des capacités de nettoyage, transformation et distribution (Apache Nifi, 2020).

Ces outils facilitent l'automatisation des tâches de transformation des données, tout en assurant la cohérence et la fiabilité des informations lors de leur transfert entre différents systèmes (Strong et al., 1997).

- **Potter's Wheel**

Potter's Wheel est un outil interactif qui permet aux utilisateurs de détecter et de corriger en temps réel les anomalies dans les bases de données. Il offre diverses fonctions comme la fusion, la division ou la suppression de données. Grâce à une interface facile à utiliser, Potter's Wheel permet aux utilisateurs d'appliquer directement des transformations sur les données et de voir instantanément les résultats.

Cet outil est particulièrement utile pour les tâches de nettoyage nécessitant une intervention humaine pour affiner ou valider les corrections proposées (Raman et al,

2001). Potter's Wheel a été largement utilisé pour les bases de données relationnelles, permettant une vérification dynamique et interactive des modifications de données.

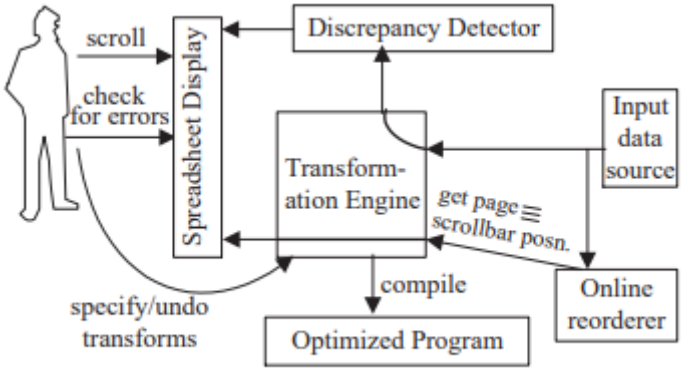

Figure 2: Potter's Wheel Architecture

Figure 4: L'architecture de Potter's Wheel (Raman et al,2001)

- **Microsoft SQL Server 2012**

Microsoft SQL Server 2012 est une base de données complète qui comprend des fonctionnalités robustes pour le nettoyage des données grâce à Data Quality Services (DQS). DQS permet d'assurer la qualité des données en les analysant, en les nettoyant et en les standardisant à l'aide de règles prédéfinies dans une base de connaissances (Kumari et al, 2013).

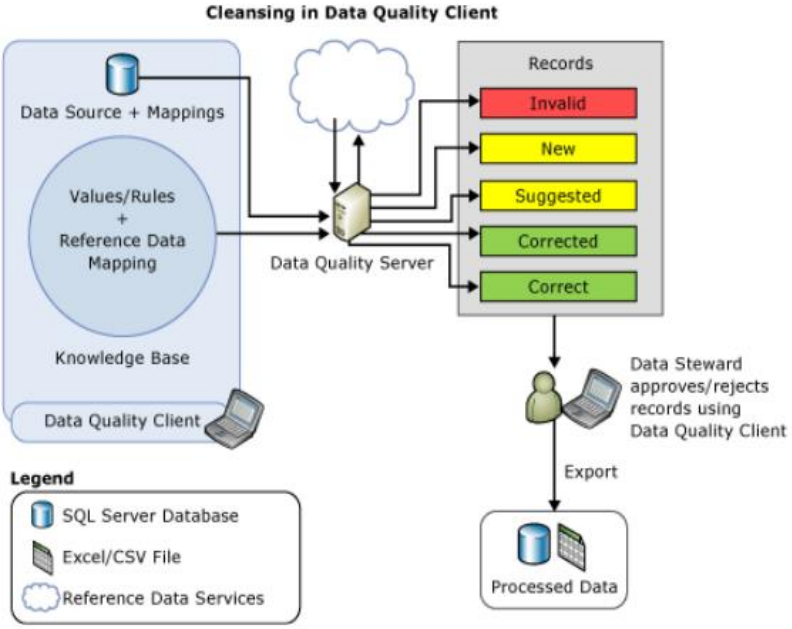

Figure 5: Cleaning in Data Quality Client (Kumari et al, 2013)

Fonctionnalités de nettoyage de données avec DQS

DQS introduit un processus de nettoyage qui repose sur une méthode assistée par ordinateur et une méthode interactive. Le processus commence par l'analyse des données par rapport à une base de connaissances, qui contient des règles et des références pour évaluer la validité des informations. Par exemple, DQS peut standardiser des abréviations ("St." devient "Street") ou enrichir des adresses manquantes ou incorrectes en s'appuyant sur des bases de données de référence.

Processus assisté par ordinateur : Ce processus utilise la base de connaissances pour automatiquement analyser les données, détecter les erreurs et proposer des corrections ou des remplacements. Il assure que les données respectent les standards établis et identifie les informations invalides ou incomplètes.

Processus interactif : Une fois l'analyse automatique terminée, un utilisateur (souvent un data steward) peut approuver, rejeter ou modifier les suggestions de correction proposées. Cette étape permet d'ajuster les données de manière plus fine, garantissant une qualité maximale avant leur validation finale.

Avantages du nettoyage avec DQS

DQS permet de normaliser et enrichir les informations en appliquant des règles de domaine (par exemple, transformation des termes courants ou remplissage des éléments manquants comme les codes postaux). Cela améliore l'intégrité des bases de données en s'assurant que toutes les données sont conformes à des standards précis.

Le DQS classe les données en différentes catégories (valides, invalides, nouvelles, suggérées, corrigées, correctes), facilitant l'identification des erreurs et anomalies. Cela permet une gestion simplifiée des erreurs et une correction rapide des informations défectueuses.

4.2. Lean six sigma

Lean Six Sigma est une méthode qui combine les principes de Lean et de Six Sigma pour améliorer la qualité des processus, réduire les inefficacités et garantir des résultats précis. Appliqué à l'amélioration de la qualité des données, Lean Six Sigma permet de structurer et d'optimiser les processus de collecte, de gestion et d'analyse des données

afin de minimiser les erreurs, d'augmenter l'efficacité et de garantir la fiabilité des informations. (Gupta, 2019).

- **Lean : Élimination du gaspillage**

Le Lean se concentre principalement sur l'élimination des gaspillages dans les processus. Il identifie les étapes inutiles ou non optimisées qui n'apportent pas de valeur au processus de collecte et de traitement des données. L'objectif est de supprimer les redondances, de réduire les délais et d'assurer une meilleure gestion des ressources, tout en optimisant le flux de travail. (George, 2012).

Dans le contexte de la gestion des données, Lean peut être utilisé pour automatiser des tâches répétitives, comme le nettoyage de données manuelles ou la détection de doublons, ce qui réduit le temps de traitement et améliore l'efficacité. L'automatisation de certaines étapes importantes permet d'éliminer les erreurs humaines et de garantir un meilleur flux des données.

- **Six Sigma : Réduction des variations et des erreurs**

Le Six Sigma se concentre sur la réduction des erreurs et des variations dans un processus. Il vise à améliorer la précision et la qualité du processus. Le Six Sigma utilise des méthodes statistiques pour identifier et résoudre les problèmes. L'objectif est de réduire le taux d'erreur à moins de 3,4 erreurs pour un million d'opportunités. (Pillet, 2003).

Lorsqu'il est appliqué à la qualité des données, le Six Sigma aide à identifier les causes des problèmes de données, comme les incohérences entre les sources ou les erreurs de saisie. Il permet de mettre en place des solutions pour minimiser ces variations.

Dans le cadre du web scraping, le Six Sigma peut être utilisé pour analyser les écarts entre les informations recueillies et les données de référence. Cela permet de corriger les erreurs et d'améliorer la précision des données avant leur intégration dans les systèmes d'analyse.

- **La méthodologie DMAIC dans Lean Six Sigma**

Une des approches les plus courantes dans Lean Six Sigma est la méthode DMAIC, un processus en cinq étapes utilisées pour améliorer et optimiser les processus :

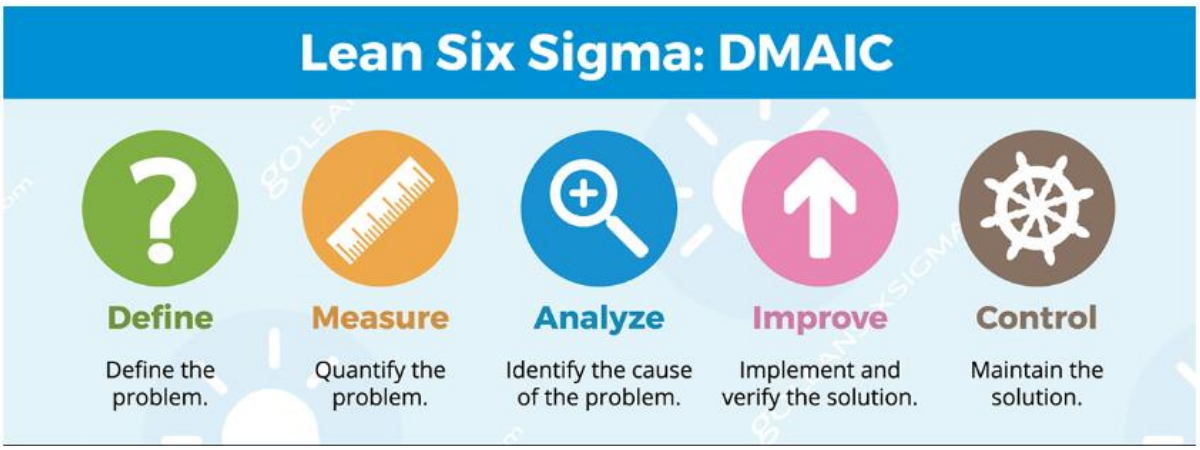

Figure 6: Lean Six Sigma : DMAIC

Définir : Identifier les problèmes et objectifs liés à la qualité des données.

Mesurer : Quantifier les problèmes de qualité des données à l'aide d'indicateurs de performance.

Analyser : Identifier les causes profondes des problèmes de qualité des données.

Innover (Improve) : Mettre en place des solutions pour améliorer la qualité des données.

Contrôler : Maintenir les améliorations apportées grâce à un suivi continu.

- **Exemples d'application de Lean Six Sigma à la qualité des données**

❖ ***Amélioration des processus de nettoyage de données***

En appliquant Lean Six Sigma, les organisations peuvent automatiser et optimiser le processus de nettoyage des données. Cela permet d'éliminer les doublons, de corriger les incohérences et d'améliorer les formats de données. Ainsi, le temps nécessaire pour rendre les données exploitables est réduit.

❖ ***Réduction des erreurs dans la collecte de données***

Dans les environnements où les données sont collectées manuellement ou par scraping, Lean Six Sigma aide à réduire les erreurs de saisie et les incohérences entre les sources. Cela se fait en standardisant les processus de collecte et en surveillant la qualité.

❖ ***Optimisation des bases de données***

Lean Six Sigma peut aussi être utilisé pour améliorer la gestion des bases de données. Il permet d'optimiser la structure des tables et des champs, ce qui garantit un accès plus rapide aux données et une meilleure fiabilité des informations stockées.

- **Bénéfices de Lean Six Sigma pour la qualité des données**

D'abord, il permet de réduire des erreurs. En utilisant des outils statistiques, Lean Six Sigma permet de réduire considérablement le taux d'erreurs dans les données.

Ensuite, Lean six sigma aide à améliorer de la fiabilité. Grâce à l'analyse des causes profondes et à la mise en œuvre de solutions correctives, les données deviennent plus cohérentes et fiables.

Et puis, Lean Six Sigma élimine les étapes superflues, réduisant ainsi les temps de traitement et améliorant l'efficacité des processus de gestion des données.

Finalement, il permet de préciser les analyses. Des données de meilleure qualité garantissent des analyses plus précises, ce qui permet de prendre des décisions mieux informées.

4.3. Utilisation de modèles de machine learning

L'utilisation du machine learning pour améliorer la qualité des données se base sur des algorithmes et des modèles prédictifs. Ceux-ci peuvent identifier, corriger et prévenir les erreurs dans les données. Ces techniques sont devenues essentielles dans de nombreux domaines, à cause de la complexité croissante des jeux de données et de la grande quantité d'informations générées chaque jour. (Zhang et al., 2010).

- **Détection des anomalies**

Les modèles de machine learning peuvent détecter les anomalies dans les jeux de données, comme les valeurs aberrantes, les incohérences ou les erreurs qui ne correspondent pas aux schémas attendus. Des techniques comme les réseaux neuronaux, les forêts d'arbres décisionnels et les algorithmes de clustering sont souvent utilisées pour cette tâche. (Chandola et al., 2009).

Dans l'industrie financière, l'utilisation de l'apprentissage automatique permet de détecter des transactions suspectes ou frauduleuses en identifiant les écarts dans les données par rapport aux modèles normaux de comportement (Phua et al., 2010).

- **Nettoyage automatisé des données**

Les techniques de machine learning peuvent être appliquées pour nettoyer les données, en supprimant automatiquement les doublons, en corrigeant les valeurs manquantes, et en standardisant les formats. Des algorithmes supervisés sont formés sur des jeux de données propres pour apprendre à corriger automatiquement les erreurs dans de nouveaux jeux de données (Chu et al., 2016).

Les entreprises de e-commerce utilisent des modèles ML pour améliorer la qualité des données sur les produits en corrigeant automatiquement les descriptions ou les prix en fonction de sources fiables (Dong et al., 2013)

- **Amélioration de la précision des données par prédiction**

Les algorithmes de machine learning peuvent être utilisés pour prédire des valeurs manquantes ou estimer des informations incorrectes en se basant sur des tendances détectées dans les jeux de données. Cela est particulièrement utile lorsque des données sont incomplètes ou manquantes, permettant ainsi d'améliorer la qualité des analyses et décisions basées sur ces données (García et al., 2015).

Dans le domaine médical, le machine learning peut être utilisé pour remplir des informations manquantes dans les dossiers de santé électroniques, en se basant sur des patients ayant des caractéristiques similaires (Beaulieu-Jones et al., 2017).

- **Validation et ajustement des données**

Les modèles peuvent être employés pour valider les données en comparant les nouvelles données aux modèles existants. Lorsque des incohérences sont détectées, des ajustements automatiques peuvent être réalisés, réduisant ainsi les erreurs humaines. De plus, des techniques telles que le feedback adaptatif permet d'améliorer continuellement la précision des modèles en fonction de nouvelles données (Taleb et al., 2015).

Chapitre III : Méthodologie proposée pour l'évaluation et l'amélioration de la qualité des données

1. Introduction

Dans l'industrie automobile, le web scraping est souvent utilisé pour collecter des informations sur les prix, les modèles, les spécifications techniques, les stocks et les options de personnalisation des véhicules. Cependant, la qualité de ces données n'est pas toujours garantie, à cause des changements fréquents sur les sites internet sources et des erreurs possibles dans les processus d'extraction automatique. Ce chapitre propose d'appliquer la méthode DMAIC de Lean Six Sigma pour évaluer et améliorer la qualité des données obtenues par web scraping. La méthode DMAIC offre une approche structurée et logique pour résoudre les problèmes de qualité des données à la source et assurer des améliorations durables.

2. Méthodologie

Pour appliquer la méthode DMAIC (Définir, Mesurer, Analyser, Innover, Contrôler) de Lean Six Sigma aux étapes d'évaluation et d'amélioration de la qualité des données issues du web scraping dans le secteur automobile, il faut restructurer les processus pour qu'ils suivent une démarche logique et efficace.

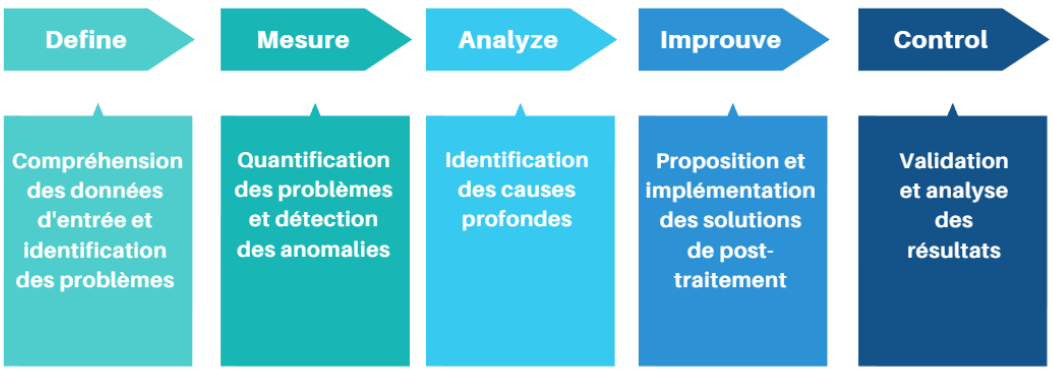

Figure 7: schéma des étapes de la méthodologie

- **Définir (Define) : Compréhension des données d'entrée et identification des problèmes**

La première étape de DMAIC comprend identifier les problèmes spécifiques liés à la qualité des données et définir les objectifs de qualité. Dans le secteur automobile, les données issues du web scraping incluent souvent des informations importantes telles que les prix des véhicules, les modèles, les spécifications techniques les offres de leasing, ou encore les options de personnalisation. Il faut s'assurer que ces données sont complètes, exactes et cohérentes pour permettre de prendre des bonnes décisions.

On doit identifier les différentes sources de données (sites de constructeurs, sites d'annonces, bases de données tierces), leur structure et leur qualité initiale. On va définir les critères de qualité spécifiques, comme la précision des informations sur les prix et les options des véhicules.

Les acteurs de cette étape jouent un rôle essentiel dans la définition des besoins, l'identification des problèmes. Ils sont illustrés dans l'image suivant :

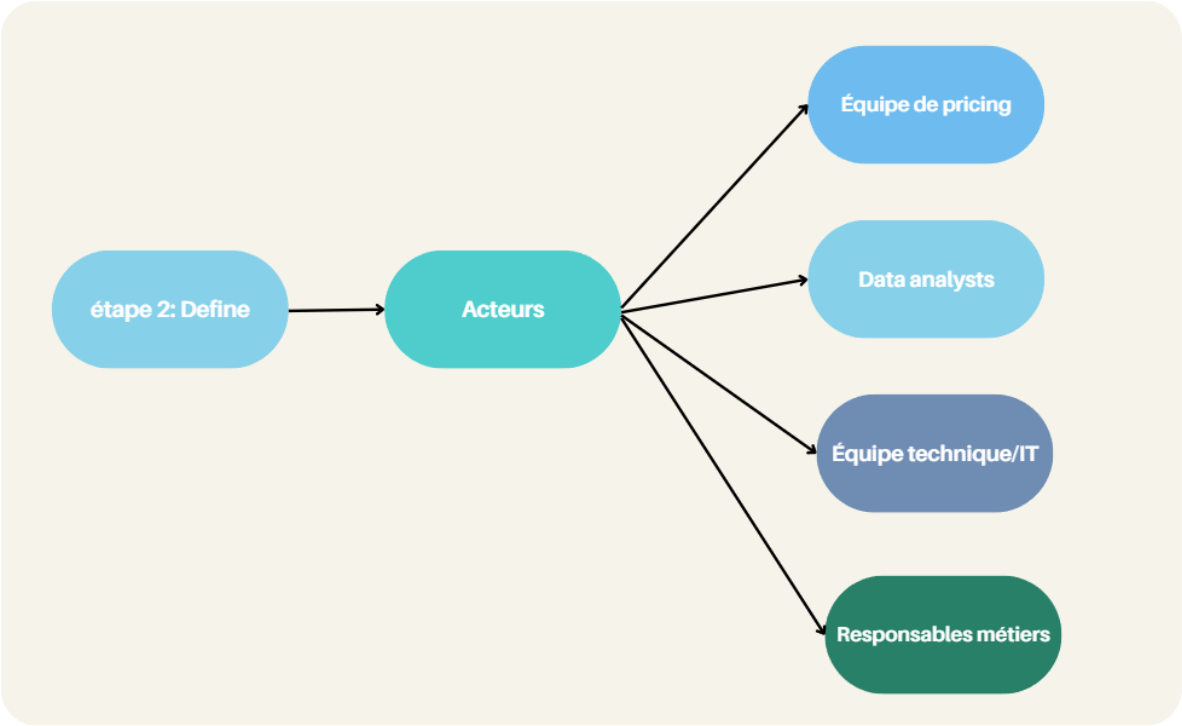

Figure 8: les acteurs de l'étape "Define"

L'équipe de pricing est chargée de définir les objectifs de prix des véhicules et de s'assurer que les données reflètent la réalité du marché. Les data analyst sont responsables de l'analyse des données issues du web scraping et de la détection des premières anomalies. L'équipe technique/informatique est en charge du maintien des outils de web scraping, de la collecte des données et de la gestion de l'infrastructure de la base de données. Les responsables métiers définissent les besoins et s'assurent que les données permettent de répondre aux exigences stratégiques de l'entreprise. Enfin, les fournisseurs de données tiers peuvent être impliqués pour garantir la fiabilité des données provenant de sources externes.

Plusieurs outils sont nécessaires pour collecter, organiser et analyser les données à cette étape. Ces outils permettent d'extraire les données de sources pertinentes et de vérifier leur fiabilité. Les outils de web scraping, comme Selenium, Scrapy et BeautifulSoup, servent à extraire automatiquement des données à partir de sites web externes, tels que ceux des concurrents. Les données collectées sont ensuite stockées dans des bases de données, par exemple sur la plateforme Google Cloud Platform, pour être analysées et structurées. Des outils de visualisation des données, par exemple Tableau ou Power BI, peuvent être utilisés pour visualiser les premières tendances des données et détecter les anomalies avec la façon visuelle.

- **Mesurer (Measure) : Quantification des problèmes et détection des anomalies**

Après avoir défini les objectifs, la phase de mesure permet d'évaluer le niveau actuel de qualité des données dans le secteur automobile. Il s'agit notamment d'identifier les anomalies qui affectent la qualité des informations, telles que des incohérences dans les prix des véhicules d'une même marque, des données manquantes ou des doublons. Des indicateurs de performance (KPI) sont utilisés pour quantifier ces problèmes, par exemple le taux d'erreurs dans les spécifications techniques.

Détection des anomalies : on peut utiliser des outils d'analyse de données et des algorithmes pour identifier les anomalies dans les informations collectées via le web scraping. On peut mesurer la qualité des données en fonction de critères définis pour le secteur automobile, comme le taux d'erreurs dans les caractéristiques des modèles ou les variations de prix entre les sources.

Plusieurs acteurs sont essentiels pour la réussite de cette phase de mesure. Chaque rôle contribue à la détection et à la quantification des anomalies dans les données collectées.

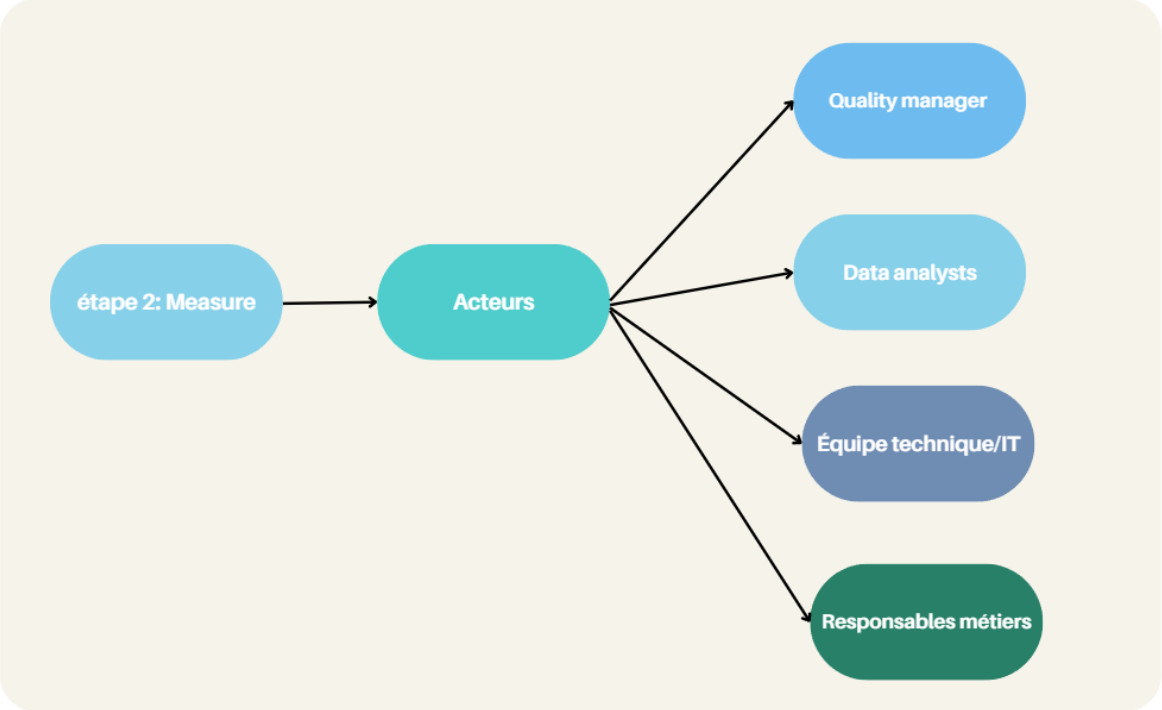

Figure 9: Les acteurs de l'étape "Measure"

Les data analyst sont chargés d'analyser les données, en utilisant des outils analytiques pour détecter les incohérences, les erreurs. Ensuite, l'équipe informatique/technique s'assure que les scripts de collecte de données fonctionnent correctement et met à jour les systèmes en cas de modifications des sites web concurrents. Ils interviennent également pour ajuster les processus de collecte quand des problèmes ou des incohérences, des erreurs sont identifiés. Les responsables métiers (prix, marketing) définissent les indicateurs clés de performance pertinents pour évaluer la qualité des données. Par exemple, pour l'équipe des prix, le taux d'écart de prix entre les sources est un indicateur essentiel. Les responsables qualité valident les critères de qualité et supervisent les analyses pour s'assurer que les méthodes utilisées respectent les normes.

On peut utiliser les outils de visualisation des données, comme Tableau, Power BI ou Google Data Studio qui permettent de voir les tendances et les anomalies dans les données. Cela aide à détecter les problèmes comme les variations de prix ou les incohérences dans les spécifications techniques. Les outils d'analyse de données, comme le langage de programmation Python avec des bibliothèques telles que Pandas, NumPy et Scikit-learn, sont utilisés pour analyser en détail les données. Cela permet d'identifier les valeurs extrêmes, les doublons et les données manquantes.

Dans l'étape de mesure, on peut utiliser plusieurs techniques pour s'assurer que les données collectées via le web scraping sont fiables.

- La validation syntaxique vérifie que les données sont dans les bons formats (par exemple, les prix doivent être des nombres corrects, et les caractéristiques techniques doivent correspondre à des valeurs prédéfinies).
- L'intégrité des données est contrôlée pour s'assurer de la cohérence entre les champs (par exemple, les prix correspondent aux spécifications du modèle sans anomalies, et les doublons sont éliminés).
- L'analyse statistique et exploratoire des données peut être utilisée pour détecter les valeurs aberrantes en utilisant la distribution des données (écart-type, médiane, moyenne) et des visualisations.
- Les données collectées sont comparées à des sources de référence fiables (comme les constructeurs ou les concessionnaires) pour identifier toute incohérence ou différence importante.

- **Analyser (Analyze) : Identification des causes profondes**

La phase d'analyse concerne à identifier les causes profondes des problèmes de qualité des données. Dans le secteur automobile, les erreurs peuvent provenir de changements fréquents dans les structures de sites web, de mises à jour irrégulières des informations par les sources, ou d'incohérences dans la collecte automatique. L'étape de structuration des données permet d'organiser les informations de manière logique afin d'éliminer les erreurs à la source.

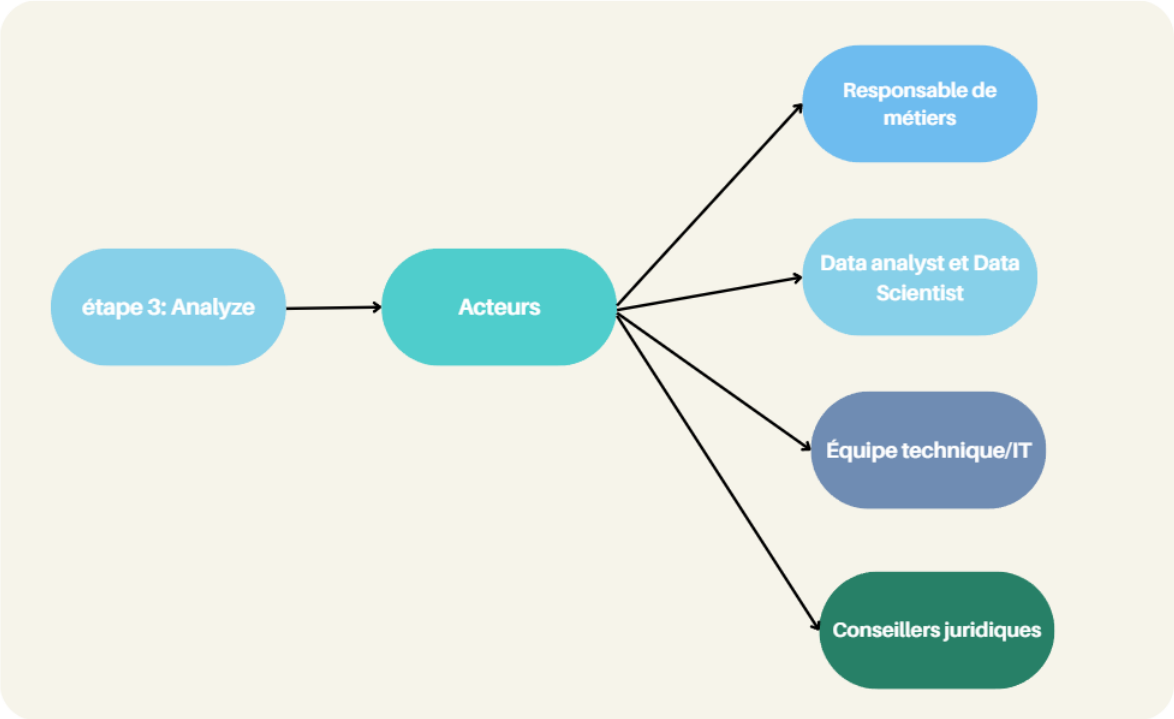

Figure 10: Les acteurs de l'étape "Analyze"

Les data scientists et les data analysts sont responsables d'analyser en profondeur les données. Ils utilisent des outils analytiques et des méthodes statistiques pour détecter les sources d'erreurs et proposer des solutions basées sur les résultats. L'équipe technique/IT examine les scripts de scraping et les infrastructures de collecte pour identifier les problèmes techniques qui peuvent influencer la qualité des données, par exemple des scripts obsolètes suite à des changements sur les sites concurrents. Les responsables métiers (pricing, marketing…) aide à définir les critères critiques pour la qualité des données. Ils contribuent à prioriser les problèmes de qualité les plus importants pour les décisions stratégiques. Les conseillers juridiques peuvent être utilisé pour identifier les risques légaux liés aux changements fréquents dans les politiques de protection des données des sites sources, dans le cadre des réglementations de scraping.

- **Innover (Improuve) : Proposition et implémentation des solutions de post-traitement**

Dans l'étape Innover (Improve), le nettoyage et le prétraitement des données sont nécessaires pour garantir la qualité des informations collectées via le web scraping et pour économiser le temps de traitement des données. Le nettoyage des données vise à identifier et à corriger les anomalies telles que les doublons, les valeurs manquantes, et les incohérences dans les spécifications techniques ou les prix. On peut utiliser les

méthodes présentées dans l'état de l'art pour nettoyer les données. Cela permet de réduire les erreurs manuelles et d'accélérer le processus d'analyse. Cela aide à éviter de perdre du temps sur des données incorrectes.

Cette phase vise à mettre en place des solutions innovantes pour résoudre les problèmes identifiés. Dans le secteur automobile, cela peut inclure l'automatisation de la correction des erreurs dans les données collectées par web scraping, ou l'intégration de systèmes qui permet de nettoyer les données en temps réel pour supprimer les incohérences et assurer la cohérence des informations.

Proposition de solutions de post-traitement : appliquer des solutions de traitement automatique, comme la correction des anomalies détectées, la suppression des doublons, la normalisation des spécifications des modèles de véhicules, et le comblement des valeurs manquantes ; utiliser des algorithmes de machine learning pour anticiper et corriger les erreurs futures.

Implémentation des solutions : intégrer des outils automatisés pour traiter et corriger les données issues du web scraping de manière continue, garantissant des informations fiables sur les prix, les stocks, et les caractéristiques des véhicules.

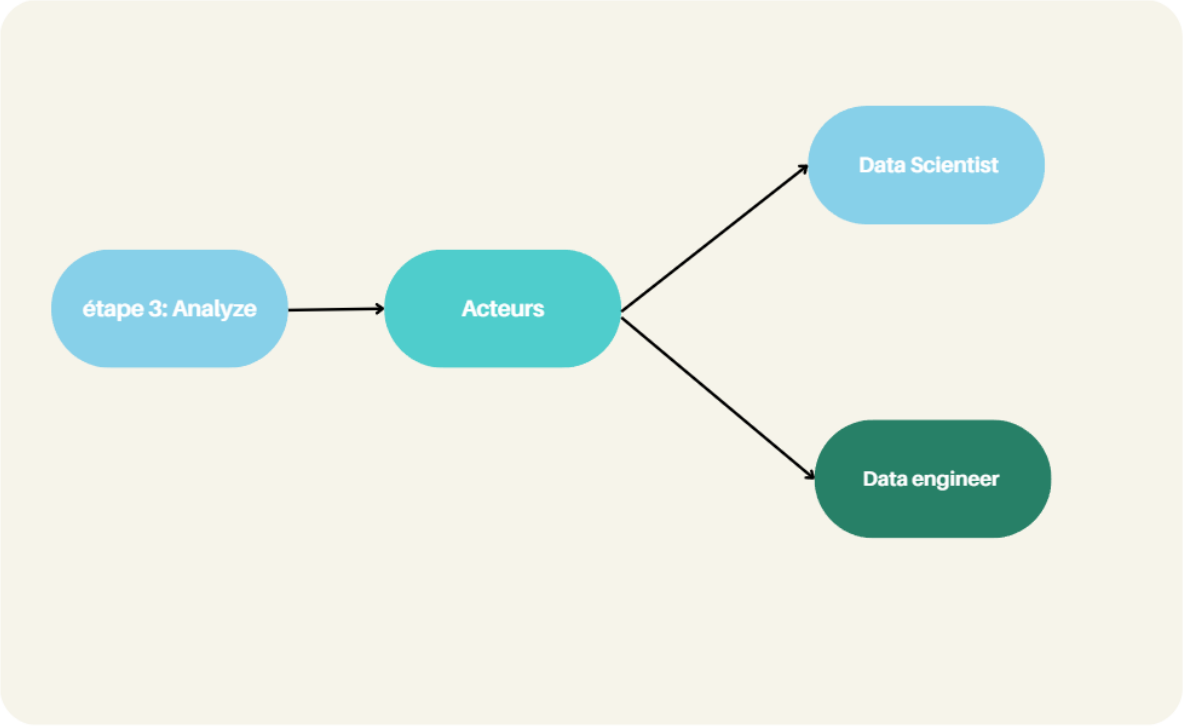

Figure 11: Les acteurs de l'étape "Improuve"

Dans cette étape, les data engineer et data scientist jouent un rôle clé dans la création et la mise en place de solutions automatisées pour traiter les données. Ils utilisent des algorithmes de nettoyage et la programmation pour corriger des erreurs.

On peut utiliser Python qui permetde nettoyer les données en détectant et corrigeant les incohérences, comme la suppression des doublons, la correction des anomalies et le comblement des valeurs manquantes. Elles permettent également de normaliser les spécifications techniques des véhicules.

- **Contrôler (Control) : Validation et analyse des résultats**

La phase de contrôle signifie garantir que les améliorations mises en place sont efficaces et que la qualité des données reste stable dans le temps. Dans le secteur automobile, cela comprend la mise en place de systèmes de surveillance continue des données collectées pour s'assurer que les informations sur les véhicules (prix, modèles, options) restent à jour et exactes.

Validation et analyse des résultats : mettre en place des processus de surveillance en continu pour garantir que la qualité des données reste conforme aux objectifs définis ; utiliser des KPI pour mesurer l'impact des solutions d'amélioration (ex. : réduction du taux d'erreurs dans les prix des véhicules, amélioration de la cohérence des informations) ; mettre en place des contrôles automatisés pour garantir la pérennité des résultats obtenus.

Dans l'étape de contrôle, comparer les données collectées par web scraping à des sources fiables est essentiel pour maintenir la qualité des données à long terme. Cette méthode est comparer régulièrement les prix, les modèles et les options des véhicules collectés en ligne avec des sites web des concurrents. En automatisant cette comparaison, on peut rapidement détecter les écarts et s'assurer que les données collectées reflètent bien la réalité du marché.

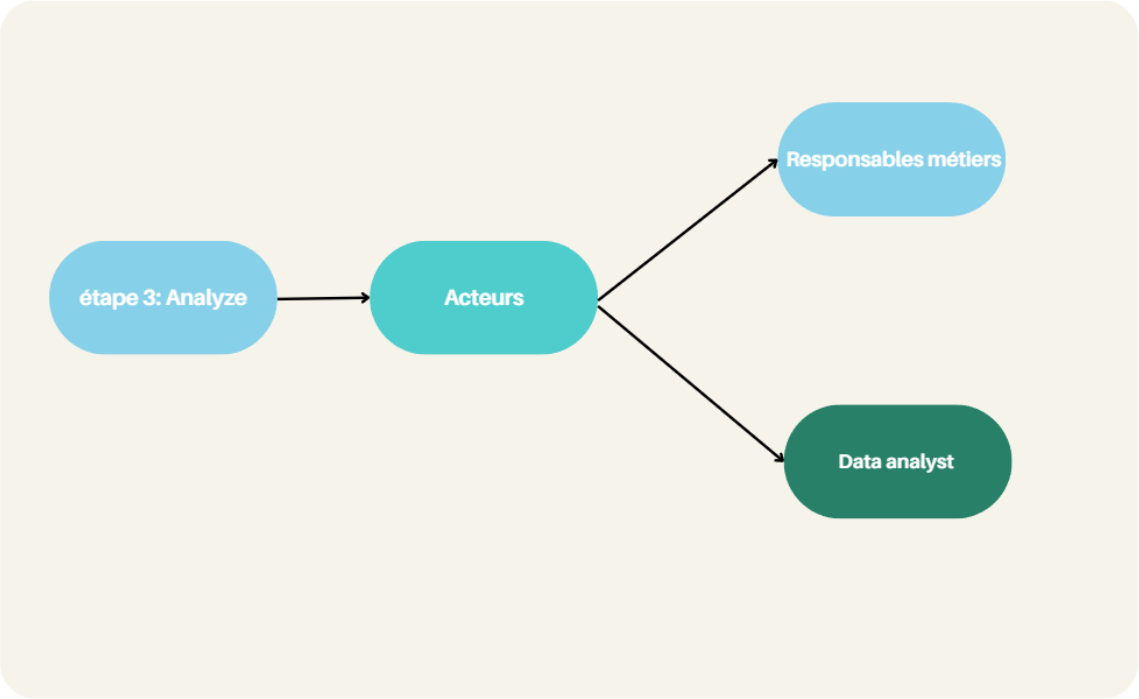

Figure 12: Les acteurs de l'étape "Control"

Les data analyst sont responsables de l'analyse continue des indicateurs de performance clés (KPI) et de la qualité des données. Ils s'assurent que les améliorations sont maintenues et que les données restent cohérentes et fiables.

Les responsables métiers surveillent l'efficacité des améliorations du point de vue commercial. Ils s'assurent que les données traitées répondent aux besoins stratégiques de l'entreprise, notamment dans les départements de pricing et de marketing.

Chapitre IV : Proposition de solution et cas d'application

Dans ce chapitre, chaque étape de la méthodologie sera détaillée avec un cas d'application chez Renault Group. Le département de pricing de Renault utilise le web scraping pour collecter des données sur les prix des véhicules de ses concurrents sur leurs sites web. Dans ce contexte, nous illustrerons comment chaque étape de la méthodologie est mise en œuvre chez Renault, en mettant l'accent sur les pratiques et les solutions spécifiques adoptées pour améliorer la qualité des données extraites.

1.Définir (Define): Compréhension des données d'entrée et identification des problèmes

Dans le cadre du département pricing de Renault, la première étape de la méthode DMAIC, Définir, consiste à comprendre les données d'entrée collectées via le web scraping sur les sites de concurrents et à identifier les problèmes potentiels. Renault utilise un processus de collecte automatisée des données concurrentielles, qui sont ensuite enregistrées et stockées dans Google Cloud Platform (GCP). Ces données sont structurées dans des colonnes prédéfinies telles que id_version, version, pays, marque, modèle, moteur, fuel_type, prix TTC, prix discounted, etc. Cependant, pour assurer la qualité des données avant leur utilisation, il est essentiel de procéder à une évaluation approfondie.

Les données d'entrée proviennent principalement des sites web des concurrents de Renault. Ces informations incluent les spécifications techniques des véhicules (modèle, moteur, version, type de carburant) et les prix (prix TTC, prix remisé). Les données sont organisées selon des colonnes structurées comme suit :

- id_version : Identifiant unique pour chaque version du modèle de véhicule.
- version : Nom ou description de la version du véhicule.
- pays : Le pays d'origine du concurrent ou le marché ciblé.

- marque : Nom de la marque automobile concurrente.
- modèle : Nom du modèle du véhicule.
- moteur : Type de moteur (ex. : diesel, essence, hybride).
- fuel_type : Type de carburant utilisé (essence, diesel, électrique).
- prix TTC : Prix toutes taxes comprises affiché pour le véhicule.
- prix discounted : Prix après remise ou promotion appliquée.

Ces données sont stockées dans des bases de données sur GCP, où elles sont disponibles pour des analyses de pricing et de positionnement concurrentiel.

Bien que les données soient bien structurées, plusieurs problèmes potentiels peuvent exister lors de leur collecte et de leur traitement. Les équipes de métiers peuvent analyser et identifier les problèmes. Voici quelques exemples de problèmes spécifiques liés aux données web scrapées dans le cadre du pricing chez Renault :

- **Incohérence des prix** : Les prix peuvent varier beaucoup d'une source à une autre en raison de différences dans la mise à jour des informations. Par exemple, un modèle de véhicule peut avoir un prix affiché sur un site concurrent qui n'est plus d'actualité.
- **Données manquantes** : Il est possible que certaines informations ne soient pas présentes sur tous les sites web concurrents. Par exemple, certains modèles peuvent ne pas inclure de prix remisé, ou certaines spécifications techniques, comme le type de carburant ou la version du moteur, peuvent être absentes.
- **Changements fréquents sur les sites web des concurrents** : Les sites des concurrents peuvent changer régulièrement la structure de leurs pages web, ce qui peut affecter la qualité et la précision des données collectées (ex. : nouveaux champs ajoutés, modification de l'ordre des données).

2.Mesurer (Measure) : Quantification des problèmes et détection des anomalies

Dans le cadre du département pricing de Renault, la deuxième étape de la méthode DMAIC, Mesurer, consiste à identifier et quantifier les problèmes de qualité dans les

données collectées via le web scraping sur les sites des concurrents. Cette phase concerne à mesurer l'ampleur des anomalies, notamment les incohérences de prix. Ces anomalies peuvent fausser l'analyse comparative des prix et des caractéristiques techniques des véhicules.

- **Incohérences de prix** :

Les incohérences de prix sont souvent dues à des remises ou options incluses par défaut. Cela peut être causé par des changements sur le site web, comme un champ de prix qui affiche désormais un prix remisé ou incluant une option, alors que ce qui nous intéresse est le prix de base sans ces ajouts. Cela entraîne des variations de prix qu'il faut analyser en effectuant de nouvelles simulations sur le site.

Ces anomalies sont souvent détectées par l'équipe technique et data analyst.

3.Analyser (Analyze) : Identification des causes profondes

Dans le département pricing de Renault, l'étape Analyser de la méthode DMAIC vise à identifier les causes des anomalies détectées dans les données collectées via le web scraping. Ces anomalies incluent des incohérences de prix et des changements fréquents dans les informations comme le moteur, le modèle ou la version. Il est important de comprendre et d'analyser ces erreurs afin d'améliorer le processus de collecte et de traitement des données.

Causes principales des incohérences de prix

- **Mises à jour irrégulières des sites concurrents** :

Les concurrents mettent souvent à jour leurs prix de manière irrégulière, en fonction des promotions, des stocks ou de leur politique commerciale locale. Lorsque le scraping est effectué à différents moments, les prix collectés peuvent ne pas montrer les dernières modifications.

Par exemple, un modèle de véhicule affiché avec un prix initial de 25 000 € sur un site concurrent peut être modifié après une promotion, mais si Renault collecte les données avant la mise à jour, l'analyse tarifaire sera basée sur des informations obsolètes.

- **Problèmes de synchronisation des données entre les sources** :

Les données proviennent de plusieurs sites concurrents qui ne sont pas mis à jour de manière synchronisée. Les prix peuvent varier selon un seul site de référence, la marque et le pays.

Par exemple, le même modèle peut être à 22 000 € sur un site et 23 500 € sur un autre. Cela est dû au fait que les mises à jour des prix ne se font pas en même temps sur tous les sites. Cette absence de synchronisation entraîne des incohérences dans l'analyse des prix réalisée par Renault.

4.Innover (Improve) : Proposition et implémentation des solutions de post-traitement

Dans le cadre de l'étape Innover de la méthode DMAIC, Renault met en place des solutions de post-traitement automatisées pour corriger les anomalies détectées dans les données collectées via le web scraping. Une principale problématique identifiée dans le département des prix est la variation significative des prix d'un mois à l'autre. Lorsque les prix d'un modèle de voiture varient de plus de 10% entre un mois et le mois précédent, cela est considéré comme une anomalie. Renault utilise un algorithme Python pour comparer automatiquement les prix et signaler ces incohérences.

L'analyse comparative des prix chez Renault vise à détecter les écarts entre les prix d'un mois à l'autre. Cette méthode est essentielle pour surveiller la stabilité des prix et s'assurer que les changements de tarifs reflètent bien les tendances du marché et les politiques commerciales. Un algorithme de comparaison des prix a été développé pour automatiser ce processus.

Fonctionnement de l'algorithme :

- L'algorithme compare le prix actuel d'un modèle de véhicule avec son prix du mois précédent. Si la différence dépasse 10 %, cela est automatiquement identifié comme une anomalie.
- L'algorithme génère une alerte pour signaler les prix aberrants et déclenche un processus de vérification.
- Cela permet à Renault de réagir rapidement aux incohérences, soit en ajustant le prix pour refléter les tendances du marché, soit en vérifiant si les données collectées sont incorrectes.

5.Contrôler (Control) : Validation et analyse des résultats

Renault procède à la validation des données collectées et à l'analyse des résultats pour s'assurer que les anomalies de prix détectées sont corrigées et que la qualité des informations est maintenue dans le temps. On utilise des analyses comparatives pour détecter les incohérences dans les prix TTC de certains modèles de véhicules (ici, le modèle ABC).

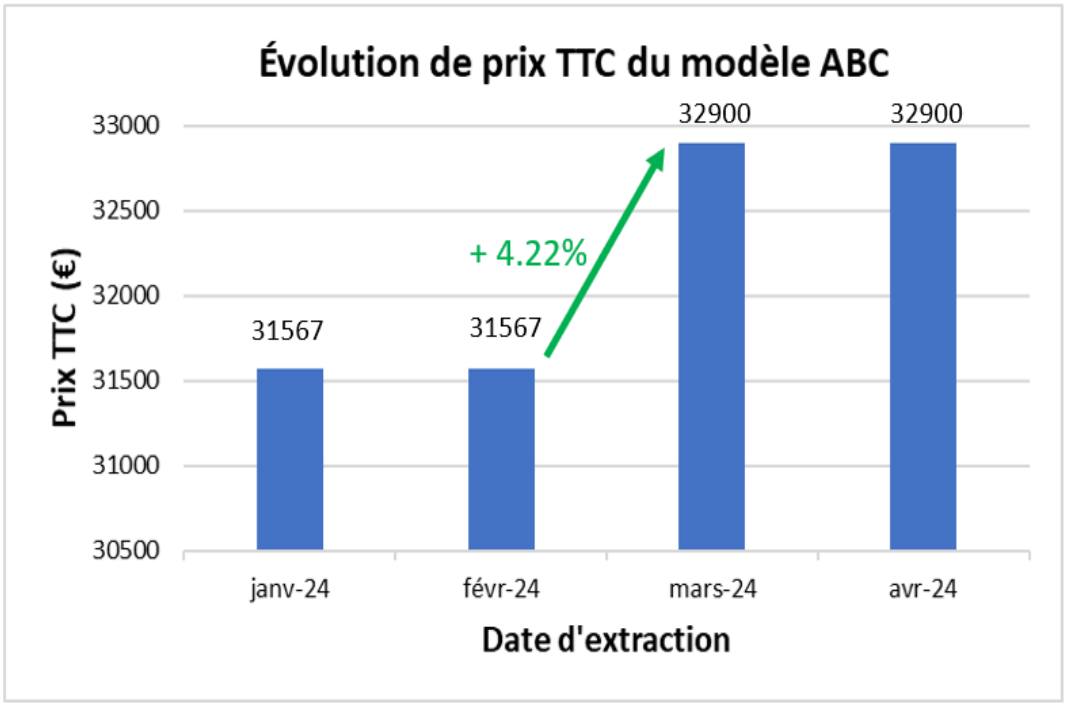

Figure 13: Évolution de prix TTC du modèle ABC

La Figure 13 fournit un exemple de l'évolution du prix TTC du modèle ABC sur une période de quatre mois, permettant de mettre en évidence des variations significatives entre ces mois, ce qui aide à détecter les anomalies potentielles et à informer l'équipe de pricing. Dans cette figure, on voit que :

- **Prix TTC du modèle ABC en janvier 2024 et février 2024** : Le prix reste stable à 31 567 €.

- **Prix TTC en mars 2024** : Le prix augmente légèrement à 32 900 €, soit une augmentation de 4,22 % par rapport au mois précédent.

Cette évolution est **cohérente** et logique. Une variation de **4,22 %** est normale dans un contexte de marché fluctuant, reflétant peut-être une petite hausse de prix due à des facteurs tels que les ajustements liés à l'inflation ou aux nouvelles promotions.

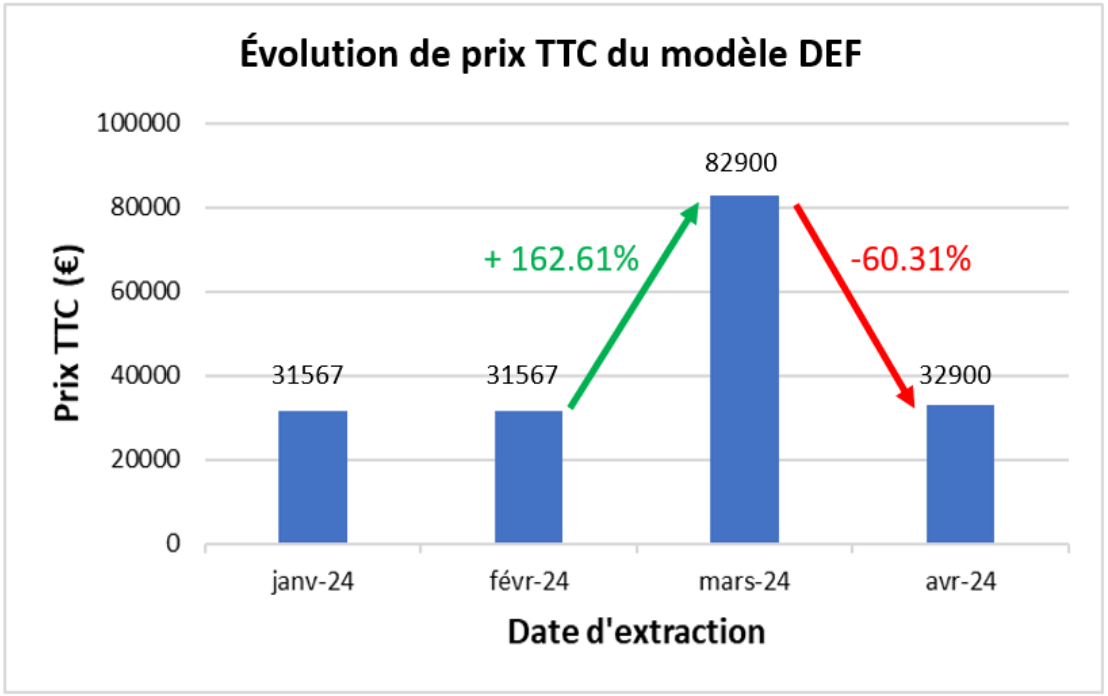

Figure 14:Évolution de prix TTC du modèle DEF

Dans cette figure, on a les remarques suivantes :

- **Prix TTC du modèle ABC en janvier et février 2024** : Le prix reste à 31 567 €.
- **Prix TTC en mars 2024** : Le prix monte de manière spectaculaire à 82 900 €, soit une augmentation de 162,61 %, ce qui est anormal.
- **Prix TTC en avril 2024** : Le prix retombe à 32 900 €, avec une baisse de 60,31 %.

Cette variation est clairement anormale et indique une incohérence de prix. Une hausse de plus de 162 % en un mois suivie d'une baisse significative de 60 % le mois suivant est très inhabituel et suggère soit une erreur de collecte de données (par exemple, une mauvaise extraction des informations), soit une erreur de saisie sur le site source du concurrent.

L'étape Contrôler permet à Renault de vérifier et de corriger les problèmes détectés dans les prix. Cela se fait grâce à des algorithmes qui comparent automatiquement les prix et à une surveillance continue des variations de prix. L'exemple des écarts de prix pour les modèles ABC et DEF montre l'importance de cette vérification. Elle aide l'équipe en charge des prix à repérer rapidement les incohérences importantes, à prendre de meilleures décisions et à ajuster la stratégie de prix pour rester compétitif sur le marché.

Chapitre V : Discussion et perspectives

A. Discussion

Dans tout projet de recherche, il est important de reconnaître les limitations qui peuvent influencer les résultats ou restreindre l'application des conclusions. Dans le cadre de ce mémoire sur l'évaluation et l'amélioration de la qualité des données issues du web scraping dans le secteur automobile, plusieurs limitations doivent être prises en compte :

- **Dépendance aux sources externes**

Le web scraping permet d'extraire des informations à partir de sites web tiers, notamment ceux des concurrents dans l'industrie automobile. Cependant, ces sites peuvent subir des changements soudains (comme des mises à jour de pages, des modifications de l'architecture des données, des blocages par CAPTCHA) qui perturbent la collecte des données. Cette dépendance aux sources externes crée un risque que certaines données deviennent inaccessibles ou inexactes sans préavis, ce qui peut affecter la qualité des analyses.

- **Fiabilité des données collectées**

Les données web scrapées peuvent parfois être sujettes à des erreurs, par exemple à cause de l'absence de mise à jour des informations sur les sites des concurrents, d'erreurs techniques dans les scripts de scraping, ou d'incohérences entre différentes sources. Il est donc nécessaire de considérer que les analyses effectuées dans ce mémoire dépendent de la qualité initiale des données collectées, qui peut ne pas être toujours parfaite.

- **Limitation des algorithmes utilisés**

Les algorithmes de détection d'anomalies, de comparaison de prix et de normalisation des spécifications utilisés dans ce projet ont montré leur efficacité, mais ils présentent aussi des limitations. Par exemple, les algorithmes de comparaison des prix se basent sur des seuils définis manuellement (par exemple, une variation de plus de 10 %), ce qui peut ne pas couvrir tous les cas possibles de variations anormales.

Les modèles de machine learning ou les algorithmes de normalisation utilisés peuvent ne pas détecter toutes les subtilités ou les incohérences dans les données, surtout en présence de variations complexes dans la nomenclature ou de données bruitées.

B. Perspectives

Ce mémoire a permis d'aborder l'évaluation et l'amélioration de la qualité des données issues du web scraping dans le secteur automobile, en utilisant la méthode DMAIC de Lean Six Sigma. Les résultats obtenus sont encourageants et montrent que les solutions proposées permettent de détecter et de corriger efficacement les anomalies dans les prix. Cependant, plusieurs perspectives peuvent être envisagées pour améliorer et étendre les travaux réalisés dans ce projet.

L'approche adoptée dans ce mémoire, notamment l'utilisation d'algorithmes de détection d'anomalies et de normalisation des données, peuvent être appliquée à d'autres secteurs d'activité au-delà de l'automobile. Par exemple, des industries telles que le commerce électronique, l'immobilier ou même le secteur des énergies renouvelables pourraient bénéficier de ces méthodes pour surveiller et analyser les prix de manière dynamique, tout en garantissant la qualité des données collectées. On pourrait adapter les solutions et les algorithmes développés dans ce mémoire pour traiter d'autres types de données, avec des caractéristiques propres à chaque secteur.

Les algorithmes de machine learning utilisés dans ce projet pour détecter les anomalies et normaliser les spécifications techniques peuvent être améliorés en explorant des techniques plus avancées, telles que les réseaux de neurones ou les modèles d'apprentissage supervisé. Cela permettrait d'accroître la capacité du système à identifier des schémas plus complexes dans les données, en particulier lorsque les anomalies sont moins évidentes. On pourrait intégrer des algorithmes d'intelligence artificielle avancés, tels que les réseaux neuronaux profonds, pour améliorer la précision de la détection des anomalies et rendre le système encore plus autonome.

Actuellement, le modèle se concentre principalement sur la détection des anomalies dans les prix et les données techniques. Cependant, il serait possible d'aller plus loin en développant des modèles prédictifs capables d'anticiper les variations de prix en fonction de facteurs externes, tels que les tendances du marché, les promotions des concurrents ou les fluctuations économiques mondiales. Ces modèles prédictifs offriraient une valeur ajoutée au département pricing en lui permettant d'anticiper les ajustements de prix nécessaires avant même que les anomalies ne se manifestent. On pourrait mettre en place des modèles de prédiction des prix pour anticiper les variations

futures, permettant ainsi à Renault de mieux ajuster sa stratégie tarifaire en fonction des tendances anticipées.

Conclusion générale

Ce mémoire a permis de faire une étude approfondie sur l'évaluation et l'amélioration de la qualité des données web scraping dans le secteur automobile, en particulier chez Renault. Face aux défis liés à la collecte et à l'analyse des données dans un environnement numérique, cette recherche s'est concentrée sur le développement de méthodes efficaces pour garantir la fiabilité et la pertinence des informations collectées en ligne.

On a exploré des concepts théoriques, des outils pratiques et des cadres d'évaluation pour répondre à la question de la qualité des données. Le web scraping, bien que largement utilisé pour extraire des données, a des limites liées à la variabilité des sources, aux erreurs de collecte et à la réglementation. Ces éléments affectent la qualité des données, nécessitant des méthodes de validation et de correction spécifiques.

La méthodologie DMAIC de Lean Six Sigma a servi de cadre pour structurer l'évaluation et l'amélioration des processus liés à la qualité des données chez Renault. Le contrôle a permis de résoudre concrètement les problèmes identifiés. Pour cela, des outils comme Selenium, Scrapy et des algorithmes Python ont été utilisés pour détecter et corriger les anomalies. Des techniques de machine learning ont également été intégrées pour améliorer la précision et l'efficacité des corrections.

Cependant, il est essentiel de reconnaître certaines limites de cette étude, telles que la dépendance aux sources externes et la nécessité d'une mise à jour régulière des outils de scraping pour suivre les modifications des sites web des concurrents. En outre, bien que les solutions proposées soient efficaces, elles peuvent nécessiter des ajustements supplémentaires pour garantir leur évolutivité et leur application dans d'autres secteurs.

Les perspectives présentées dans ce mémoire ouvrent la voie à des évolutions prometteuses. Par exemple, l'intégration de modèles prédictifs pour anticiper les variations de prix, et l'utilisation de technologies comme la blockchain pour garantir l'intégrité et la traçabilité des données. Ces évolutions permettront de renforcer la fiabilité des données collectées et d'améliorer la stratégie tarifaire de Renault dans un environnement concurrentiel.

En conclusion, ce mémoire a démontré l'importance de l'amélioration continue des processus de collecte et de traitement des données, ainsi que la nécessité d'adopter des

solutions technologiques avancées pour garantir des décisions éclairées basées sur des données de haute qualité. Les résultats obtenus posent les bases d'une stratégie de pricing robuste, capable de s’adapter aux évolutions rapides du marché et de répondre aux exigences du secteur automobile à l'ère du digital.

Références

[Albrecht, 2020] J. P. Albrecht. How the GDPR Will Change the World. *European Data Protection Law Review*, 6(1), 2020, pp. 20-28.

[Barzin, 2020] F. Barzin. L'extraction de données pour refléter les activités du marché immobilier namurois en temps réel et l'intégrer dans un système de support à la décision pour aider les collectivités locales à prendre des décisions informées, Mémoire, Université de Namur, 2020.

[Batini et al, 2004] C. Batini, T. Catarci, M. Scannapieco. *A survey of data quality issues in cooperative information systems* in the Proceedings of the International Conference on Conceptual Modeling (ER), 2004.

[Beaulieu-Jones et Moore, 2017] B. K. Beaulieu-Jones, J. H. Moore. Missing data imputation in the electronic health record using deeply learned autoencoders, in *Pacific symposium on biocomputing 2017*, World Scientific, pp. 207-218, 2017.

[Benson, 2009] P. Benson. NATO Codification System as the Foundation for ISO 8000, the International Standard for Data Quality. *Accessed: 2024-05-03*. 2009.

[Berners-Lee et al., 1994] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, A. Secret. The World-Wide Web. *Communications of the ACM*, 37(8), 1994, pp. 76-82.

[Berti-Équille, 2004] L. Berti-Équille. Ingénierie des systèmes d'information. Qualité des données.

[Berti-Équille, 2006] L. Berti-Équille. *Modelling and measuring data quality for quality-awareness in data mining*, Springer, 2006.

[Breunig et al, 2000] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander. *LOF: Identifying density-based local outliers* in the Proceedings of the ACM SIGMOD International Conference, 2000, pp. 93-104.

[Caruso et al, 2000] F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, P. Missier. *Telcordia's database reconciliation and data quality analysis tool* in the Proceedings of the International Conference on Very Large Databases (VLDB), 2000, pp. 615-618.

[Chandola et al., 2009] V. Chandola, A. Banerjee, V. Kumar. Anomaly detection: A survey, *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1-58, 2009.

[Chu et al., 2016] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang. Data cleaning: Overview and emerging challenges, in *Proceedings of the 2016 international conference on management of data*, pp. 2201-2206, 2016.

[Cichy et al, 2019] C. Cichy, S. Rass. An Overview of Data Quality Frameworks. *IEEE Access*, 7, 2019, pp. 24634-24648.

[Dong et Srivastava, 2013] X. L. Dong, D. Srivastava. Big data integration, in *Proceedings of the 2013 IEEE 29th international conference on data engineering (ICDE)*, IEEE, pp. 1245-1248, 2013.

[Ehrlinger et al, 2022] L. Ehrlinger, W. Wöß. A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, 5, 2022.

[García et al., 2015] S. García, J. Luengo, F. Herrera. Data preprocessing in data mining, *Springer*, 2015.

[Geewax, 2021] J. J. Geewax. *API Design Patterns*. Manning Publications, 2021, pp. 3-22.

[George, 2012] M. L. George. Lean Six Sigma pour les services comment utiliser la vitesse Lean et la qualité Six Sigma pour améliorer vos services et transactions, *Revue Française de Gestion Industrielle*, 2012.

[Glez-Peña et al., 2014] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, F. Fdez-Riverola. Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), 2014, pp. 788-797.

[Goulet, 2012] M. Goulet. Hiérarchiser les dimensions de la qualité des données : analyse comparative entre la littérature et les praticiens en technologies de l'information, 2012, pp. 23-24.

[Gundecha, 2018] U. Gundecha. *Selenium WebDriver with Python: Learn how to automate web applications for testing*. Packt Publishing, 2018.

[Gupta et al., 2019] S. Gupta, S. Modgil, A. Gunasekaran. Big data in lean six sigma: a review and further research directions, *International Journal of Production Research*, vol. 58, no. 3, pp. 947–969, 2019.

[Hemenway et Calishain, 2003] K. Hemenway, T. Calishain. *Spidering Hacks: 100 Industrial-Strength Tips & Tools*. O'Reilly Media, Inc., 2003.

[Iezzoni, 1997] L. I. Iezzoni. Assessing quality using administrative data dans *Annals of Internal Medicine*, vol. 127, no. 8, pp. 666-674, 1997.

[Jung et al., 2021] W.-C. Jung, J. Kim, N. Park. Web-Browsing Application Using Web Scraping Technology in Korean Network Separation Application. *Symmetry*, MDPI, 13, 2021, p. 1550.

[Khder, 2021] M. A. Khder. Web scraping or web crawling: State of art, techniques, approaches and application, *International Journal of Advances in Soft Computing*, vol. 11, pp. 45-56, 2021.

[Kumari et al, 2013] S. Kumari, N. Shama. Comparative Study of Data Cleaning Tools, *International Journal of Computer Applications*, vol. 76, no. 7, pp. 15-19, 2013.

[Meschenmoser et al, 2016] P. Meschenmoser, N. Meuschke, M. Hotz, B. Gipp. Scraping scientific web repositories: challenges and solutions for automated content extraction, *D-Lib Magazine*, vol. 22, no. 7/8, pp. 12-24, 2016.

[Mitchell, 2015] R. Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 2015.

[Mitchell, 2018] R. Mitchell. *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, Inc., 2018.

[Pandey et al, 2020] C. A. Pandey, C. A. John, D. Nanjundan. Portfolio builder and career recommender by scraping data using Flask and Tensorflow, *Academia.edu*, 2020.

[Phua et al., 2010] C. Phua, V. Lee, K. Smith, R. Gayler. A comprehensive survey of data mining-based fraud detection research, *arXiv preprint arXiv:1009.6119*, 2010.

[Pillet, 2003] M. Pillet. Six Sigma une rupture dans l'approche qualité, *Colloque ISMQ*, 2003.

[Raman et al, 2001] V. Raman, J. M. Hellerstein. Potter's Wheel: An Interactive Data Cleaning System, *Proceedings of the VLDB Conference*, 2001.

[Roos et al., 1982] L. L. Roos, N. P. Roos, S. M. Cageorge, J. P. Nicol. How good are the data?. *Reliability of one health care data bank dans Medical Care*, vol. 20, no. 3, pp. 283-299, 1982.

[Rosati, 2019] E. Rosati. *Copyright and the Court of Justice of the European Union*. Oxford University Press, 2019, pp. 45-67.

[Sebastian-Coleman, 2013] L. Sebastian-Coleman. Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. *Morgan Kaufmann*, 2013.

[Senaratne et al., 2017] H. Senaratne, A. Mobasheri, A. L. Ali. A review of volunteered geographic information quality assessment methods dans *International Journal of Geographical Information Science*, vol. 31, pp. 139-165, 2017.

[Streiff, 2020] M. Streiff. Web Scraping and the Law: Navigating Copyright and Data Privacy Issues. *Harvard Journal of Law & Technology*, 33(1), 2020, pp. 45-70.

[Taleb et al., 2015] I. Taleb, R. Dssouli, M. A. Serhani. Big data pre-processing: A quality framework, in *Proceedings of the 2015 IEEE international congress on big data*, IEEE, pp. 191-198, 2015.

[Vargiu et Urru, 2013] E. Vargiu, M. Urru. Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*, 2(1), 2013, pp. 44-54.

[Vetrò et al., 2016] A. Vetrò, L. Canova, M. Torchiano, D. L. Orellana, M. Acua. Open data quality measurement framework: Definition and application to Open Government Data dans *Government Information Quarterly*, vol. 33, pp. 325-337, 2016.

[Voigt et Von dem Bussche, 2017] P. Voigt, A. Von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 1st Ed., Cham: Springer International Publishing, 2017.

[Zhang et al., 2010] Y. Zhang, N. Meratnia, P. Havinga. Outlier detection techniques for wireless sensor networks: A survey, *IEEE communications surveys & tutorials*, vol. 12, no. 2, pp. 159-170, 2010.

[Zhao, 2017] B. Zhao. Web scraping. In *Encyclopedia of Big Data*, Springer, Cham, 2017, pp. 1-3.

[Gudivada et al., 2017] V. Gudivada, A. Apon, J. Ding. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations dans *International Journal on Advances in Software*, vol. 10, no. 1-2, pp. 25-35, 2017.

Sitographie

Apache Nifi. (2020). Flow-Based Programming for the Enterprise. Apache Nifi. Disponible sur : https://nifi.apache.org/

Google Inc. (2023). Puppeteer Documentation. Puppeteer. Disponible sur https://pptr.dev/https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Informatica. (2011). Informatica PowerCenter: User Guide. Informatica. Disponible sur https://docs.informatica.com/content/dam/source/GUID-2/GUID-2940FB03-64D3-4C9E-A304 54D03A93554E/7/en/PWX_1056_PowerCenterUserGuideForPowerCenter_en.pdf

Octoparse Team (2023). Octoparse Documentation. Octoparse. Disponible sur :https://helpcenter.octoparse.com/

Richardson, L. (2023). Beautiful Soup Documentation. Crummy. Disponible sur

Scrapy Team (2023). Scrapy Documentation. Scrapy. Disponible sur https://docs.scrapy.org/en/latest/

SeleniumHQ Team (2023). SeleniumHQ Documentation. SeleniumHQ. Disponible sur https://www.selenium.dev/documentation/Talend

Talend (2006). Modern data management that drives real value. Disponible sur https://www.talend.com/