

# Master 2 MIAGE MBDS

# Rapport TPA

BANTOS Elodie

DIALLO Mouctar

ACHOURI Abdenour

SOUARE Noumouke

GUO Yue

**Enseignants encadreurs :**

* Michel SYSKA
* Gabriel MOPOLO
* Nicolas PASQUIER
* Sergio SIMONIAN

### Introduction générale

En France, en 2022, les ventes de voitures ont connu une baisse record de 13.3% par rapport à 2021. Notre objectif est donc simple : dans un premier temps, nous chercherons à démystifier les raisons de la popularité – ou du manque de popularité – des voitures auprès des consommateurs, puis dans un second temps, nous établirons une prédiction permettant de trouver la voiture la mieux adaptée selon le profil du consommateur.

Nos missions sont donc les suivantes :

1. Dans un premier temps, nous devons récupérer les données pour les inclure dans le Data Lake.
2. Dans un second temps, nous allons nous référer au fichier CO2.csv afin de filtrer les données sous Hadoop avec le principe Map/Reduce.
3. Après cela, nous allons intégrer nos données extraites avec Hadoop dans notre Data Lake.
4. Dès que nos données sont toutes incluses dans notre Data Lake, nous allons créer un système de visualisation de nos données afin d’avoir une vision clarifiée de ce dont nous disposons.
5. Grâce à cette vision, nous pourrons choisir les données pertinentes pour notre base de données de recherche (filtrage des données).
6. Dans un dernier temps, nous allons créer un algorithme en R permettant de déterminer à partir de notre nouvelle base de données quels sont les véhicules qui vont être achetés selon le profil du client

Dans un premier temps, nous allons donc présenter notre projet en nous appuyant sur les points essentiels du contexte métier. Puis, nous allons expliquer comment nous nous sommes réparti les tâches selon nos facilités respectives en développement. Après cela, nous représenterons notre Data Lake sous forme de schéma puis nous expliquerons en détails comment nous avons décidé de construire notre Data Lake. Par la suite, nous décrirons les étapes de la mise en œuvre de l’activité autour de Hadoop Map Reduce, avant d’expliquer comment nous avons opté de visualiser les données avec des outils de DataViz puis nous passerons à l’analyse en profondeur de ces données. L’exploitation de cette analyse finale nous permettra de conclure sur les résultats obtenus pour l’entreprise, de lui proposer des solutions pérennes et adaptées grâce à l’élaboration d’un bilan personnalisé.

Présentation du projet

Répartition du travail entre les membres du groupe

**Répartition originelle du travail :**

DataViz - Yue et Abdenour

Data Lake en SQL - Mouctar

Spark - Elodie et Mouctar

PostgreSQL - Noumouke

**Gestion de projet :**

Daily du 07/11/2023

| Nom et prénom | Travail effectué | Difficultés rencontrées | Objectif de la semaine |
| --- | --- | --- | --- |
| BANTOS Elodie | -Ebauche d’une première structure de l’Architecture  -Définition du projet dans le rapport | Aucune | -Se répartir les tâches et démarrer le projet avec une structure valide |
| ACHOURI Abdenour | -Ebauche d’une première structure de l’Architecture  -Définition du projet dans le rapport |  | -Se répartir les tâches et démarrer le projet avec une architecture valide |
| SOUARE Noumouké | Correction de l’architecture du projet |  | -Se répartir les tâches et démarrer le projet avec une architecture valide |
| GUO Yue | Compréhension du sujet |  | -Se répartir les tâches et démarrer le projet avec une architecture valide |

Daily du 15/11/2023

| Nom et prénom | Travail effectué | Difficultés rencontrées | Objectif de la semaine |
| --- | --- | --- | --- |
| BANTOS Elodie | Mise en place du fichier Spark sur jupyter et début de l’analyse des données en Map/Reduce | Aucune | -Continuer sur le fichier Spark pour obtenir des données propres |
| ACHOURI Abdenour |  |  |  |
| SOUARE Noumouké | Correction de l’architecture du projet |  |  |

Daily du 21/11/2023

| Nom et prénom | Travail effectué | Difficultés rencontrées | Objectif de la semaine |
| --- | --- | --- | --- |
| GUO Yue | Tentative d’installation de l’environnement du Data Lake | Changement de tâche car a des problèmes d’environnement sur le Data Lake | Installation et réglage des problèmes relatifs à l’environnement Docker |
| BANTOS Elodie | Début de la mise en place du fichier Spark pour la lecture du fichier CO2.csv | Difficulté à séparer les données relatives à Bonus/Malus du fichier | -Installation Docker  -Finir le fichier Spark et commencer les tests d’intégration du fichier Spark dans le Data Lake |
| ACHOURI Abdenour | Étude des modèles de prédiction attendus.  Prise en main du jeu de données. | Aucune | -Installation Docker |
| SOUARE Noumouké |  | Aucune | -Installation Docker  -Recherche sur l’ETL pour faire transférer les données d’une table à l’autre |
| DIALLO Mouctar |  |  | -Installation Docker |

Daily du 28/11/2023

| Nom et prénom | Travail effectué | Difficultés rencontrées | Objectif de la semaine |
| --- | --- | --- | --- |
| GUO Yue |  |  |  |
| BANTOS Elodie | Spark terminé, en attente de la validation de la pull request, réalisation du rendu côté SPARK | Aucune | Aider pour le data lake ou le data warehouse |
| ACHOURI Abdenour |  |  |  |
| SOUARE Noumouké |  |  |  |
| DIALLO Mouctar |  |  |  |

Architecture du Data Lake

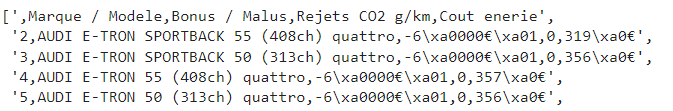
Construction du Data Lake par étapes

Hadoop Map/Reduce

Le fichier à traiter est le fichier CO2.csv et l’outil employé afin de réaliser cette opération est l’outil Spark. En aval de toute opération de modification d’un fichier, il est intéressant de procéder à une période d’analyse de celui-ci afin de déterminer comment nous allons agencer notre code.

Le fichier CO2.csv est un fichier qui semble avoir été mal formaté à la suite d’extraction de données. Cela peut arriver par exemple lorsque l’encodage d’origine est en UTF-8 (qui supporte mieux les particularités des langages Européens) et que l’on le transite vers un encodage en latin-1.

En se penchant sur les données de notre fichier, on remarque que certaines anomalies ne sont pas gênantes comme des accents remplacés par des caractères spéciaux, qui peuvent être facilement interprétables par un œil humain. Cependant, ces anomalies d’encodage s’avèrent bien plus handicapantes lorsqu’il s’agit de tarifs, comme les bons/malus, ou d’estimations, comme les émissions de CO2.



Sur l’illustration ci-dessus, nous pouvons voir par exemple que les prix des Bonus ou malus sont faussés, et qu’il est impossible de définir combien de “0” correspondent au prix réel.

Après quelques recherches, nous avons découvert que la suite d’éléments “xa0” représentait en réalité un espace insécable en Latin-1. Nous en avons donc déduit que cela signifiait que nous devions traiter notre fichier en Spark de manière à supprimer cette expression régulière.

Pour ce faire, nous avons tout d’abord différencié chaque ligne de notre fichier de manière à ce que celles-ci apparaissent en listes en utilisant la fonctionnalité split(‘,’), afin que chaque données à l’intérieur de celles-ci soient considérées par notre outil spark comme un indice de cette liste.

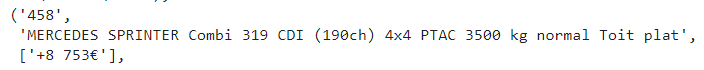


Par exemple, sur la ligne précédente, les numéros de ligne apparaissent à l’indice 0, tandis que les Bonus/Malus sont considérés comme apparaissant à l’indice 2.

Cette étape étant terminée, nous avons décidé d’opérer directement sur nos données, en les séparant chacunes les unes des autres afin d’avoir une meilleure visibilité sur celles-ci avant de supprimer chaque expression régulière en Latin-1. Dans un premier temps, nous avons traité les données relatives au Bonus/malus puis nous avons traité les données du coût en énergie.

Voici un aperçu du code utilisé et des données propres :

CO2\_clean\_rejet **=** CO2\_Separated\_data**.**map(**lambda** x: (x[0], x[1], [value**.**replace('\xa0', ' ')**.**strip() **for** value **in** x[2]**.**split(',')], x[3], x[4]))



Une bonne analyse du document et l’utilisation de l’outil Spark ont donc permis d’éviter d’éventuelles erreurs de traitement de données dans notre data lake et nous permettra une analyse plus poussée de nos données en R après la création de notre data warehouse.

Visualisation des données avec des outils de DataViz

Analyse des données avec des outils de Machine Learning

Conclusion générale

Références et bibliographies

Annexes