

# Projet Big Data Analytics : Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèles de Véhicules

(Par Nicolas PASQUIER pour la partie Analyse de Données,  
Gabriel MOPOLO et Sergio SIMONIAN pour la partie Gestion des Données,

Marco WINKLER pour la partie Data Visualisation, deadline : **30 Avril 2024, date non négociable !!!**)

## 1. Contexte et Objectifs

La réalisation de ce projet va permettre d'évaluer les cours de *Big Data* (Pr. Gabriel MOPOLO et Sergio SIMONIAN), de *Data Visualisation* (Pr. Marco WINCKLER), le cours *Data Analytics & Machine Learning* (Pr. Nicolas PASQUIER) et voire même le cours de programmation WEB/Mobile (G. GALLI, ...).

Lire attentivement le dossier.

### Contexte du Projet

Vous avez été contacté par un concessionnaire automobile afin de l'aider à mieux cibler les véhicules susceptibles d'intéresser ses clients. Pour cela il met à votre disposition :

- Son catalogue de véhicules
- Son fichier clients concernant les achats de l'année en cours
- Un accès à toutes les informations sur les immatriculations effectuées cette année
- Une brève documentation des données
- Un vendeur (voir son interview ci-dessous)

Votre client sera satisfait si vous lui proposez un moyen afin :

- Qu'un vendeur puisse en quelques secondes évaluer le type de véhicule le plus susceptible d'intéresser des clients qui se présentent dans la concession
- Qu'il puisse envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sélectionnés par son service marketing (voir ci-dessous)

### Documentation des Données

Les fichiers de données à votre disposition vous sont décrits dans les tables ci-dessous. Pour chaque attribut du fichier, vous sont donnés son nom, son type (numérique, caractères, catégoriel ou booléen) sa description et son domaine de valeurs.

Certains attributs peuvent comporter des valeurs manquantes ou incorrectes (erreur de saisie par exemple). Celles-ci sont représentées par une cellule vide ou bien contenant une valeur hors du domaine de valeurs de la variable (valeurs « ? », « » ou « N/D » par exemple).

*Immatriculations.csv : informations sur les immatriculations effectuées cette année*

| Attribut        | Type       | Description                                 | Domaine de valeurs   |
|-----------------|------------|---|--|
| Immatriculation | caractères | Numéro unique d'immatriculation du véhicule | Texte au format « 9999 AA 99 »   |
| Marque          | caractères | Nom de la marque du véhicule                | Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundaï, Jaguar, Kia, Lancia, Mercedes , Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo |

|           |            |                           |   |
|-----------|------------|---------------------------|---|
| Nom       | caractères | Nom du modèle de véhicule | S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6 FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4 |
| Puissance |            | Puissance en chevaux Din  | [55, 507]   |
| Longueur  |            | Catégorie de longueur     | courte, moyenne, longue, très longue  |
| NbPlaces  | numérique  | Nombre de places          | [5, 7]  |
| NbPortes  | numérique  | Nombre de portes          | [3, 5]  |
| Couleur   | catégoriel | Couleur                   | blanc, bleu, gris, noir, rouge  |
| Occasion  | booléen    | Véhicule d'occasion ?     | true, false   |
| Prix      | numérique  | Prix de vente en euros    | [7500, 101300]  |

Catalogue.csv : catalogue de véhicules

| Attribut  | Type       | Description                  | Domaine de valeurs  |
|-----------|------------|------------------------------|---|
| Marque    | caractères | Nom de la marque du véhicule | Audi, BMW, Dacia, Daihatsu, Fiat, Ford, Honda, Hyundai, Jaguar, Kia, Lancia, Mercedes, Mini, Nissan, Peugeot, Renault, Saab, Seat, Skoda, Volkswagen, Volvo   |
| Nom       | caractères | Nom du modèle de véhicule    | S80 T6, Touran 2.0 FSI, Polo 1.2 6V, New Beatle 1.8, Golf 2.0 FSI, Superb 2.8 V6, Toledo 1.6, 9.3 1.8T, Vel Satis 3.5 V6, Megane 2.0 16V, Laguna 2.0T, Espace 2.0T, 1007 1.4, Primera 1.6, Maxima 3.0 V6, Almera 1.8, Copper 1.6 16V, S500, A200, Ypsilon 1.4 16V, Picanto 1.1, X-Type 2.5 V6, Matrix 1.6 FR-V 1.7, Mondeo 1.8, Croma 2.2, Cuore 1.0, Logan 1.6 MPI, M5, 120i, A3 2.0 FSI, A2 1.4 |
| Puissance | numérique  | Puissance en chevaux Din     | [55, 507]   |
| Longueur  | catégoriel | Catégorie de longueur        | courte, moyenne, longue, très longue  |
| NbPlaces  | numérique  | Nombre de places             | [5, 7]  |
| NbPortes  | numérique  | Nombre de portes             | [3, 5]  |
| Couleur   | catégoriel | Couleur                      | blanc, bleu, gris, noir, rouge  |
| Occasion  | booléen    | Véhicule d'occasion ?        | true, false   |
| Prix      | numérique  | Prix de vente en euros       | [7500, 101300]  |

Clients\_N.csv<sup>1</sup> : fichier clients concernant les achats de l'année en cours

| Attribut | Type | Description | Domaine de valeurs |
|----------|------|-------------|--------------------|
|----------|------|-------------|--------------------|

<sup>1</sup> Le numéro N (1, 2, 3, ...) du fichier Clients\_N.csv à utiliser dépend de votre numéro de groupe.

|                    |            |  |   |
|--------------------|------------|--|---|
| Age                | numérique  | Age en années du clients                                   | [18, 84]  |
| Sexe               | catégoriel | Genre de la personne                                       | M, F  |
| Taux               | numérique  | Capacité d'endettement du client en euros (30% du salaire) | [544, 74185]  |
| SituationFamiliale | catégoriel | Situation familiale du client                              | Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule |
| NbEnfantsAcharge   | numérique  | Nombre d'enfants à charge                                  | [0, 4]  |
| 2eme voiture       | booléen    | Le client possède déjà un véhicule principal ?             | true, false   |
| Immatriculation    | caractères | Numéro unique d'immatriculation du véhicule                | Texte au format « 9999 AA 99 »                          |

*Marketing.csv : clients sélectionnés par le service marketing*

| Attribut           | Type       | Description  | Domaine de valeurs                                      |
|--------------------|------------|--|---|
| Age                | numérique  | Age en années du clients                                   | [18, 84]  |
| Sexe               | catégoriel | Genre de la personne                                       | M, F  |
| Taux               | numérique  | Capacité d'endettement du client en euros (30% du salaire) | [544, 74185]  |
| SituationFamiliale | catégoriel | Situation familiale du client                              | Célibataire, Divorcée, En Couple, Marié(e), Seul, Seule |
| NbEnfantsAcharge   | numérique  | Nombre d'enfants à charge                                  | [0, 4]  |
| 2eme voiture       | booléen    | Le client possède déjà un véhicule principal ?             | true, false   |

### Informations Données par le Concessionnaire

L'interview du gestionnaire de la concession automobile nous a permis de définir le contexte et les objectifs de l'application :

« Les différents véhicules de notre catalogue répondent à des besoins différents. Certains sont petits afin de mieux circuler en ville, d'autres ont de l'espace pour transporter toute une famille tandis que certains sont plus puissants et destinés à une clientèle plus fortunée. Nous souhaitons définir différentes catégories de véhicules afin de mieux comprendre les désirs des clients et proposer aux nouveaux clients le véhicule le plus adapté à leurs besoins. ».

La réalisation de ce projet va nécessiter la mise en œuvre des notions acquises durant les cours concernant les thématiques des Big Data, de la Data Visualisation et du Data Mining & Machine Learning.

## 2. Projet **Big Data Analytics** avec les **Lacs de Données** et les tables externes : **BDA/DL**

Dans ce projet, vous devez :

- En 2.1 Gérer les données avec l'architecture BDA/DL
- En 2.2 Mettre en œuvre la visualisation des données
- En 2.3 Effectuer l'analyse de données en préparant les données en entrées des outils d'analyses (R, ...) avec SQL.

### 2.1 Architecture du projet BDA/DL Attendues par G. MOPOLO et S. SIMONIAN

Le schéma ci-dessus vous aidera à construire un lac de données avec plusieurs composants de votre choix. Vous devez choisir entre un des deux lacs de données suivants :

- 1) Le lac de données HIVEQL. Ici, le moteur SQL HIVE d'Hadoop sera le frontal de votre lac. Toutes les données devront être accessible depuis HIVE pour construire les modèles d'analyse
- 2) Ou le lac de données SQL. Ici, un moteur SQL du marché (Oracle, SQL server, MySQL, PostgreSQL, etc.) sera le frontal de votre lac. Toutes les données devront être accessible depuis HIVE pour construire les modèles d'analyse.

#### 2.1.1 Les sources de données

Voici la description des différentes sources de données :

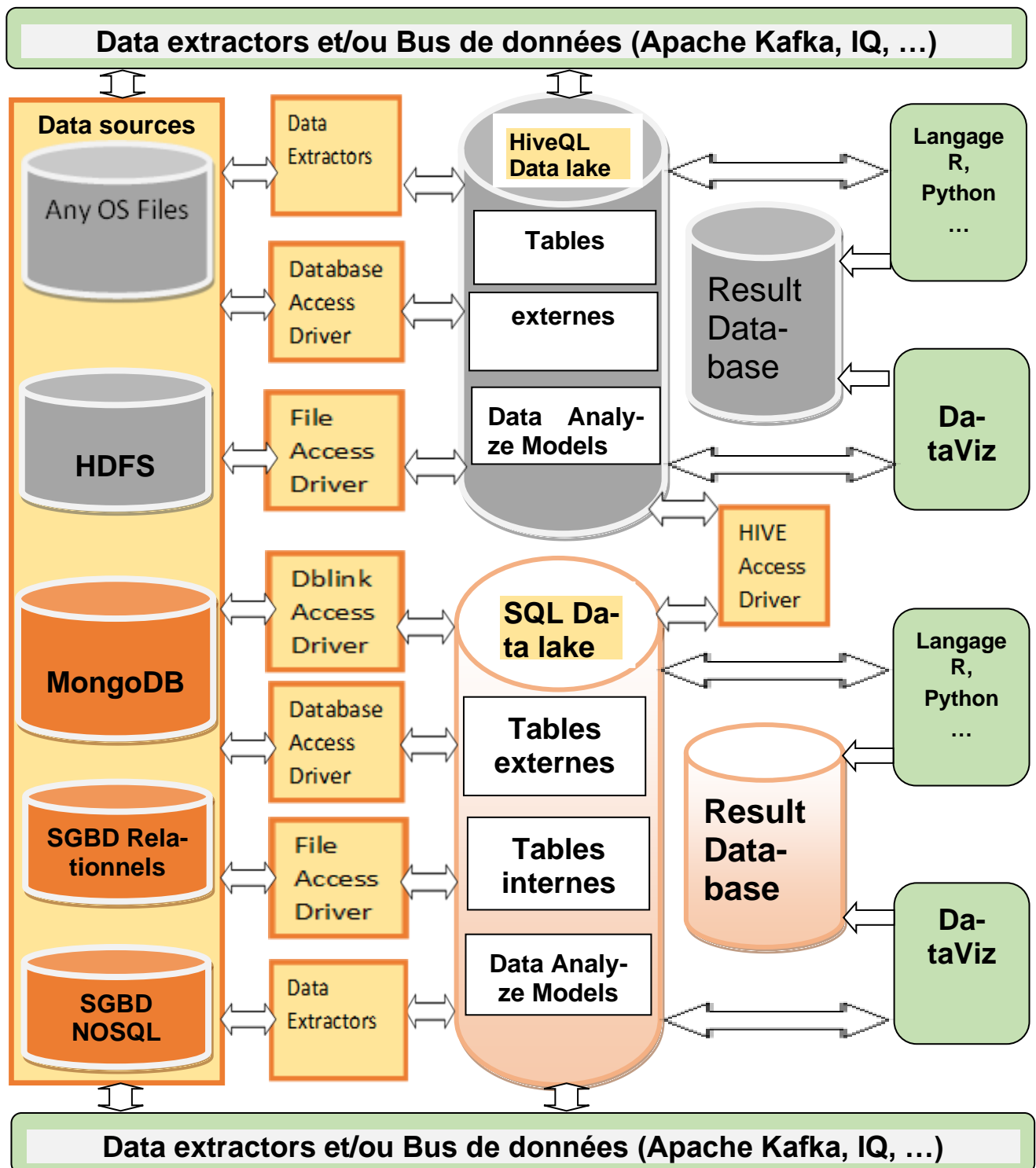
- 1) La source de données HDFS (Hadoop Data File System). HDFS est le file system distribué d'Hadoop. Cette source est obligatoire et doit apparaître dans chaque architecture. Le fichier CO2.csv, au minimum, doit y résider. Vous pouvez y déposer d'autres fichiers
- 2) La source Any OS Files (Windows Files system, Mac OS files system, Linux Files System). Cette source n'est à utiliser que si la source HDFS pose un problème
- 3) La source de données MongoDB. Cette source est obligatoire dans chaque projet., L'alimentation de cette source se fera via un extracteur de données (voir la section concernée plus loin)
- 4) La source de données Autre SGBD NoSQL. Cette source est obligatoire dans chaque projet. L'alimentation de cette source se fera via un extracteur de données (voir la section concernée plus loin). Cependant, le choix du moteur NoSQL est libre.

#### 2.1.2 L'alimentation du data lake

Vous avez deux possibilités pour rendre les données disponibles dans votre lac de données : via les access driver (accès directs aux données depuis la source sans les déplacer) et via les extracteurs de données (Accès temps différés : les données sont actualisées à intervalle définies).

Les data extractors peuvent être :

- **ETL (Extract Transform and Load)**. A l'origine ces outils ont été conçus pour alimenter un datawarehouse
- **ELT (Extract Load and Transform)**. A l'origine ces outils ont été conçus pour alimenter un data lake
- **Un chargeur maison**. Il s'agit d'un programme chargé d'extraire et recopier les données dans le lac de données (à utiliser en même temps ou alternative d'un ETL/ELT)
- Un chargeur du SGBD source ou du SGBD du datalake
- Un chargeur de l'environnement Hadoop (Sqoop, Flume, ...).



Les access drivers :

- Database access driver : ce sont des drivers qui vous permettront d'accéder aux données d'une base de données à partir du data lake sans déplacer les données de cette base
- Hive access driver : ce sont des drivers qui vous permettront d'accéder aux données d'une base de données HIVE à partir du data lake sans déplacer les données de cette base
- Data file access driver : ce sont des drivers qui vous permettront d'accéder aux fichiers du file system à partir du data lake sans charger ces fichiers
- DBlink access driver : ce sont des drivers qui vous permettront d'accéder aux données distribuées d'une base de données à partir du data lake sans déplacer les données à travers des tuyaux appelés Database Link.

Vous devez utiliser au moins un de ces access drivers dans votre architecture.

### 2.1.3 Le Bus de données

Un bus de données joue un rôle important dans l'échange de données en temps réel entre les sous-systèmes. Il permet de gérer la latence pour limiter les attentes et organiser la collecte efficace des données. Un bus garantit le découplage entre producteurs de données (sources dans l'architecture) et consommateurs (data lake dans l'architecture). Un supporte des consommateurs multiples.

L'utilisation d'un bus de données dans votre architecture sera un plus. **Il est obligatoire s'il y a un groupe de de plus de 4 étudiants.**

### 2.1.4 Les données du Data lake

Votre lac de données contiendra :

- Les données physiques (data lake physique) chargées avec vos data extractors
- Des données virtuelles (data lake virtuelle) accessibles via vos access drivers
- Et des données des modèles d'analyses que vous devez construire (à partir des data lake physique et virtuel) avant d'utiliser des outils de machine learning, deep learning, etc. Ces données doivent être présentées sous forme de vues SQL.

### 2.1.5 Les outils d'analyse et visualisation

Vous utiliserez pour la visualisation, le reporting et l'analyse : les outils qui vous ont été enseignés en cours. Vous pouvez aussi utiliser des outils équivalents de votre choix avec l'accord de l'enseignant responsable.

Ces outils reçoivent en entrée les données notamment les données des modèles.

### 2.1.6 Result database

Les résultats des analyses et de reporting devront être stockées dans une base de résultats. Vous devez choisir vous-même la nature de cette base.

### 2.1.7 Application WEB de pilotage de l'environnement du data lake

Vous pouvez développer une application web pour piloter votre environnement du data-lake. Ce développement sera un bonus cette année donc c'est facultatif.

## 2.1.8 Travail à rendre sur l'ingénierie des données dans l'architecture DBA/DL

- Un document contenant :
  - Le descriptif du projet.
  - L'architecture DBA/DL avec la ventilation des données par base ou systèmes de fichiers
  - La description détaillée et précise de votre architecture.
- Des programmes et des scripts
  - Les scripts et les programmes de chargement des données dans chacune de vos sources (Bases de données ou systèmes de fichiers)
  - Les scripts et programmes de chargement des données dans le data lake physique
  - Les scripts et/ou programmes d'accès aux données depuis le data lake virtuel
  - Les scripts et/ou programmes contenant les requêtes SQL vous permettant de construire progressivement vos modèles d'analyse
  - Tout programme ou script permettant de restituer l'ensemble de votre activité
  - Les sources des applications si avez développé quelques-unes
  - Les scripts d'installation et de configuration pour tous les composants mis en œuvre

## 2.1.9 HADOOP MAP REDUCE (cours Sergio SIMONIAN)

### Histoire / Détail

Après avoir construit votre DATA LAKE le Concessionnaire vous appelle et vous fait part que certaines données étaient perdues avant votre intervention – notamment les détails sur l'émission CO2 / le coût d'énergie / la valeur de Bonus/Malus pour la taxation par marque et modèle de voiture. Il est possible que ses données seraient utiles pour améliorer la qualité de vos modèles prédictives. En cherchant sur Internet vous avez trouvé un fichier CO2.csv. C'est une autre base des données qui a certaines informations qui peuvent vous aider mais elle n'est pas parfaite. Elle ne contient pas tous les marques et modèles des voitures qui sont dans le catalogue du Concessionnaire. De plus le format de stockage est différent (la marque et le modèle sont dans une même colonne), il y a des valeurs manquantes (colonne Bonus/Malus) et des valeurs erronées (colonne Bonus/Malus par exemple contient '-6 000€ 1' à la place de '-6 000€').

Le but est d'écrire un programme map/reduce avec Hadoop ou Spark qui va permettre d'adapter le fichier CO2.csv pour intégrer ses informations complémentaires dans la ou les tables catalogue du Concessionnaire (ajouter des colonnes "Bonus / Malus", "Rejets CO2 g/km", "Coût Énergie").

### Notes :

- Les modèles des voitures du fichier CO2.csv n'ont pas beaucoup de valeurs en commun par rapport à la table catalogue – on voudrait utiliser une valeur moyenne d'émission CO2 (de même pour les autres colonnes : "Bonus / Malus", "Coût Énergie") pour la marque de voiture concernée.
- Pour les marques de voitures qui ne sont pas dans le fichier CO2.csv on voudrait insérer la moyenne d'émission CO2 (de même pour les autres colonnes) de tous les marques de véhicules qui sont présents des deux côtés.
- Pour l'import / export des données vous pouvez utiliser des connecteurs Hadoop vu dans le cours 3 ou/et découvrir l'outil Hadoop Sqoop (<http://sqoop.apache.org/>) qui est projet Apache simplifiant cette tâche.
- Le fichier CO2.csv se trouve avec les autres ressources disponibles.

### Travail à rendre :

- Code source du programme map/reduce utilisé pour l'adaptation du fichier CO2.csv et son intégration dans la table catalogue.
- Un rapport au format PDF décrivant votre démarche pour l'adaptation du fichier CO2.csv et son intégration dans la table catalogue incluant les commandes / programmes que vous avez exécutés et leur résultat / output

## 2.2 Techniques de Data Visualisation du sous projet BDA/DL et Activités Attendues par M. WINCKLER (**NE CONCERNE PAS MBDS EMSI**)

Il faudrait analyser les données du domaine d'application et les catégoriser :

- Décrire la chaîne de traitement (« visualisation pipeline ») que, à partir de données brutes, permet de créer une représentation graphique et interactive avec l'ensemble de données.
- Décrire les utilisateurs visés ;
- Décrire les objectifs de visualisation et les tâches utilisateurs ;
- Développer de (au moins) trois techniques de visualisations avec l'API D3JS selon les contraintes suivantes :
  - o Utiliser une base de données multidimensionnelles (minimum 5 attributs différents) ;
  - o Prévoir de l'interaction avec l'ensemble des données (ex. navigation, sélection, filtres, etc.) ;
  - o Deux niveaux de visualisation, c'est-à-dire une vision globale (« overview ») plus contexte ;
  - o Donner la possibilité de charger des ensembles de données indépendants de l'application
- Faire un prototype exécutable avec D3JS qui démontre que l'utilisateur peut explorer l'ensemble des données pour atteindre les buts fixés.

## 2.3 Analyse des Données du sous projet BDA/DL par les Techniques de Data Mining, Machine Learning et Deep Learning et Activités Attendues par N. PASQUIER ou un le professeur qui le remplace

Voici la description d'un processus générique possible pour réaliser cette analyse. Ce processus peut être étendu et particularisé en utilisant d'autres étapes ou techniques, afin par exemple d'optimiser l'approche ou de vérifier la cohérence des résultats obtenus durant les différentes étapes par exemple.

### 1) Analyse exploratoire des données :

L'analyse exploratoire des données vous permettra d'identifier d'éventuels problèmes dans les données (valeurs incohérentes, codage des valeurs manquantes, etc.) et découvrir d'éventuelles propriétés de l'espace des données (valeurs doublons, variables liées, variables d'importance particulière ou bien inutiles, etc.).

Appliquez pour cela les différentes méthodes d'analyse exploratoire des données vues en cours (statistiques descriptives, histogrammes, nuages de points, boîtes à moustaches, etc.).

### 2) Identification des catégories de véhicules :

Vous devez à partir des informations du *Catalogue* identifier des catégories de véhicules (citadine, routière, sportive, etc.) en fonction de leur taille, puissance, prix, etc. Ces catégories doivent correspondre à divers besoins de la part des clients (une grande voiture pour les familles nombreuses, une petite voiture pour circuler en ville, etc.).

Ces catégories de véhicules constitueront les classes à prédire durant les étapes suivantes du processus.

### 3) Application des catégories de véhicules définies aux données des *Immatriculations* :

Les données d'*Immatriculations* contiennent les informations sur les véhicules vendus cette année. L'objectif est d'attribuer à chacun de ces véhicules la catégorie qui lui correspond en utilisant le modèle définissant les catégories de véhicules généré précédemment.



#### 4) Fusion des données *Clients* et *Immatriculations* :

Les données *Clients* contiennent les informations sur les clients ayant les véhicules vendus cette année. L'objectif est de faire la fusion entre les données des *Clients* et des *Immatriculations* afin d'obtenir sur une même ligne l'ensemble des informations sur le client (âge, sexe, etc.) et sur le véhicule qu'il a acheté (avec sa catégorie).

Cet ensemble de données servira lors des étapes suivantes pour l'apprentissage de la catégorie de véhicules (variable cible) la plus adaptée à un client selon ses caractéristiques (variables prédictives).

#### 5) Création d'un modèle de classification supervisée pour la prédiction de la catégorie de véhicules :

L'objectif de cette étape est de créer à partir du résultat de la fusion précédente un classifieur (modèle de classification supervisée) permettant d'associer aux caractéristiques des clients (âge, sexe, etc.) une catégorie de véhicules.

Testez les différentes approches et algorithmes (arbres de décision, random forests, support vector machines, réseaux de neurones, deep learning, etc.), avec pour chaque algorithme plusieurs paramètres testés, afin d'obtenir un classifieur aussi performant que possible.

L'évaluation et la comparaison des performances de chaque configuration algorithmique (un algorithme et un paramétrage spécifiques) testée sera réalisée grâce aux matrices de confusion et mesures d'évaluation calculées à partir des résultats des tests des classifieurs.

#### 6) Application du modèle de prédiction aux données *Marketing* :

Les données *Marketing* contiennent les informations sur les clients pour lesquels on souhaite prédire une catégorie de véhicules.

L'objectif est de prédire pour chacun de ces clients la catégorie de véhicules qui lui correspond le mieux en utilisant le classifieur généré durant l'étape précédente.

### Travail à Rendre sur la Partie Data Mining, Machine Learning et Deep Learning

Vous devez déposer dans la boîte de dépôt du projet :

- Un rapport au format PDF décrivant les réalisations pour la gestion et l'analyse des données.

Ce rapport doit décrire :

- Les choix effectués lors du projet en termes de gestion et d'analyse des données.
- Le(s) processus suivis.
- Les modèles de connaissances générés et l'interprétation de ces modèles.
- Les résultats que vous obtenez pour les clients sélectionnés par le service marketing.

- Les codes sources utilisés pour l'analyse des données et la génération des résultats que vous présentez dans le rapport.

Cette partie doit contenir l'ensemble des scripts que vous avez créés pour analyser et générer les modèles de connaissances à partir des données.

## 3 Organisation du projet

**Pour réaliser ce projet vous devez former une équipe de 4 étudiants.**

Nous sommes ouverts au cas où vous avez une autre architecture à nous proposer (sql ne sera pas sacrifié). Elle doit être validée avec Gabriel MOPOLLO MOKE

Une soutenance sera programmée à la fin du projet