Projet d'Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèle

1. Contexte et objectifs

Ce projet va permettre d'évaluer le cours Big Data et le cours Data Analytics.

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 dans la section 2)

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)

2. Informations données par le vendeur

« 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. ».

3. 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

Attribut	Type	Description	Domaine de valeurs
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, 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	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[1] : fichier clients concernant les achats de l'année en cours

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
Immatriculation	caractères	Numéro unique d'immatriculation du véhicule	Texte au format « 9999 AA 99 »

1. *Le numéro N = [0..19] du fichier Clients_N.csv à utiliser dépend de votre numéro de binôme.*

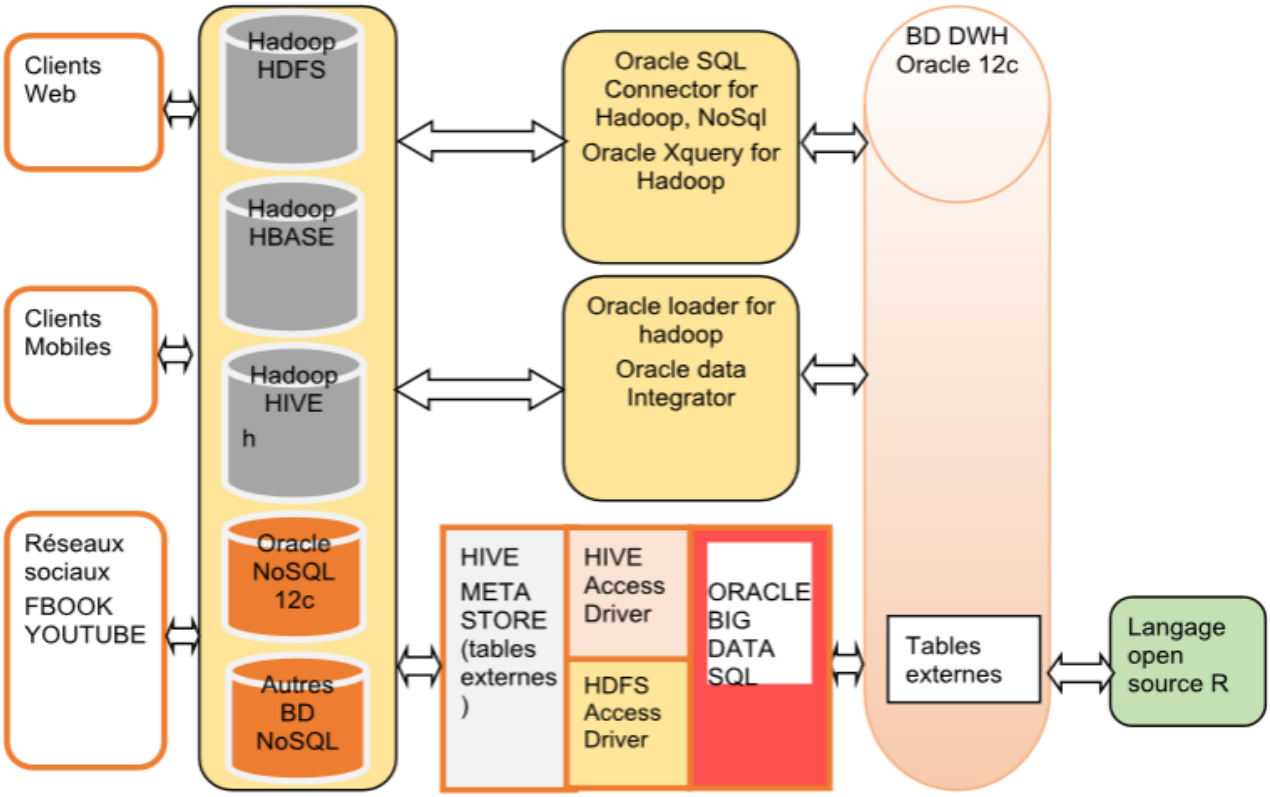
Marketing.csv : clients sélectionnés par le service marketing

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

4. Architecture Big Data à mettre en œuvre et gestion des données

L'architecture pour le stockage des données à mettre en œuvre pour la réalisation du projet est illustrée dans la figure ci-dessous.

Illustration 1: Architecture Big Data pour le stockage des données



L'organisation des données se fera comme suit :

- Une ou plusieurs de vos sources de données devront être chargés sur la base oracle NOSQL
- Une ou plusieurs de vos sources de données doivent être des fichiers Hadoop HDFS
- Une ou plusieurs de vos sources de données devront être chargés dans la base SQL

Les requis pour la mise en œuvre de cette organisation sont :

- Les données doivent être accessible au niveau de la base SQL via des tables externes.
- Le chargement des données dans la base Oracle NOSQL doit se faire à travers un programme Java. Nous restons ouverts si vous choisissez une autre solution (ODI, sqoop, …).
- Le chargement des données dans la base Oracle 12c doit se faire en utilisant l'utilitaire Oracle SQL Loader (cf. cours Big Data DBA1).
- L'accès aux données pour l'analyse avec R se fera via des requêtes SQL.

Les éléments du travail effectués pour cette partie du projet qui doivent être rendus sont :

- Un document contenant le descriptif des tâches effectuées pour le stockage des données, l'architecture du projet Big Data avec la ventilation des données par base.
- Le script de création des tables sur la base NOSQL Oracle.
- Le programme Java de chargement des données dans les tables dans la base NOSQL.

5. Quelques pistes pour la partie analytique de votre projet

- Mettre en application une méthodologie de gestion de projet et établir un plan de mise en œuvre du projet.
- Faire une analyse exploratoire initiale des données afin d'identifier d'éventuels problèmes dans les données (valeurs incohérentes, valeurs manquantes, etc.) et découvrir d'éventuelles propriétés de l'espace des données .
- Essayer de répartir les produits et/ou les clients en différentes catégories.
- Mettre au point un ou des modèles qui répondent aux objectifs donnés..

6. Documents à rendre

Vous devez déposer dans la boîte de dépôt du projet sur Jalon :

(1) 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.

(2) Les codes sources utilisés pour la gestion des données.

Il doit contenir l'ensemble des programmes et scripts que vous avez créés pour stocker et gérer vos données.

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

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