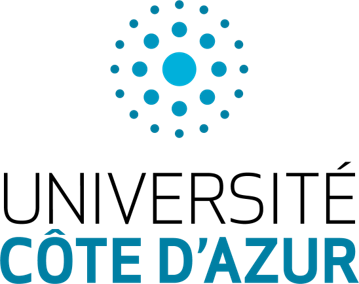
Chloe Maccarinelli - Arnaud Fernadez - Thomas Beatini - Cédric Ortega

M2 MIAGE MBDS UNIVERSITE côte d’azur

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

Architecture du Projet Big Data Voitures et Activités Attendues par G. MOPOLO et B. RENAUT (à rendre 6 Janvier 2020)

Une image contenant clipart

Description générée automatiquement

# Présentation du projet

## 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'inté-resser des clients qui se présentent dans la concession

• Qu'il puisse envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sé-lectionnés par son service marketing (voir ci-dessous)

## Informations Données par le Concessionnaire

L’interview du gestionnaire de la concession automobile nous a permis de définir le contexte et les ob-jectifs de l’application :

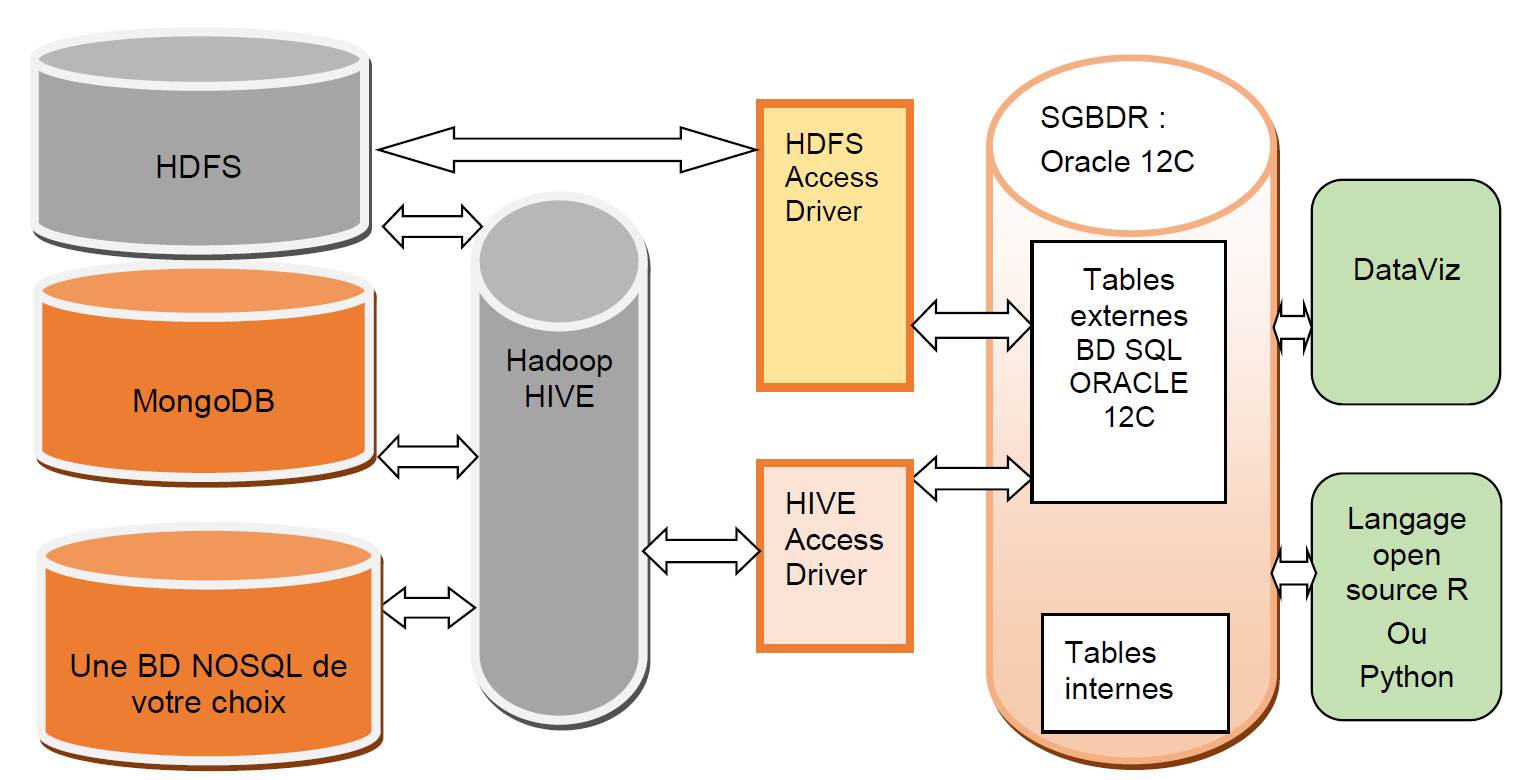
« Les différents véhicules de notre catalogue répondent à des besoins différents. Certains sont petits afin de mieux circuler en ville, d'autres ont de l'espace pour transporter toute une famille tandis que certains sont plus puissants et destinés à une clientèle plus fortunée. Nous souhaitons définir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

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

# Architecture 1 : DATA LAKE HADOOP avec tables externes

L’architecture DATA LAKE HADOOP qui consiste à s’appuyer les tables externes pour accéder aux données de sources hétérogènes (MongoDB, une 2ème BD NoSQL de votre choix, Hadoop HDFS, Oracle SQL).

L’Accès aux données pour les Data Visualization et Data Analysis with R se fera via le langage SQL interrogeant des tables externes et internes.



L’organisation des données se fera comme suit :

• Une ou plusieurs de vos sources de données devront être chargées sur la base NOSQL de votre choix

• Une ou plusieurs de vos sources de données devront être chargées sur la base MongoDB

• Une ou plusieurs de vos sources de données doivent être des fichiers Hadoop HDFS

• Un de vos fichiers CLIENTS sera chargé sur la base Oracle SQL comme table interne

Les données doivent être accessibles au niveau de la base SQL via des tables externes et internes. Vous répartissez vos données comme vous le souhaitez.

Le chargement des données dans les différentes bases de données doit se faire :

• Via des outils HADOOP tel que SQOOP ou des programmes java ou tout autre utilitaire (pour charger les données dans les bases NoSQL par exemple)

• Via Sqlloader ou autre (pour charger les données dans la base SQL)

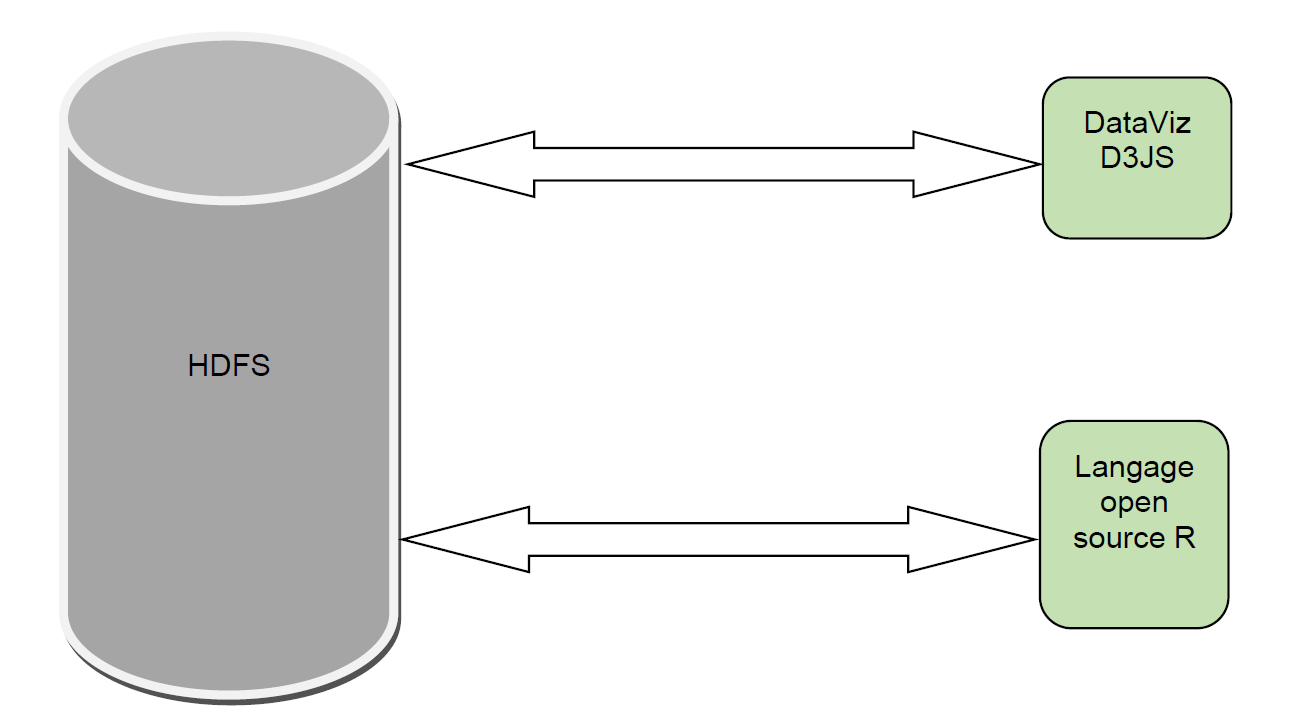
• Il faut considérer les données chargées comme si c’étaient des données brutes.

L’accès aux données pour l’analyse avec R se fera via des requêtes SQL (cela concerne l’activité demandée par Nicolas PASQUIER et Marco WINCKLER).

# Architecture 2 : HDFS HADOOP uniquement sans tables externes

L’architecture HDFS HADOOP qui consiste à accéder aux fichiers directement via le système de fichiers HADOOP HDFS.

L’Accès aux données pour la Data Visualization et la Data Analyse avec R se fera en accédant directement aux fichiers.

Vous devez charger tous vos fichiers dans hadoop hdfs et procéder à la visualisation et à la création de vos matrices en vue de l’analyse par la suite.

**Travail à rendre sur l’ingénierie des données dans les architectures 1 et 2**

* + • Scripts pour l’architecture 1 o Le script de création de documents et de chargement de données dans MongoDB, le script de création des tables externes MongoDB dans HIVE et enfin le script de création des tables externes correspondantes dans Oracle NoSQL.
  + o Le script de création de tables dans Oracle NoSQL, le script de création des tables ex-ternes Oracle NoSQL dans HIVE et enfin le script de création des tables externes corres-pondantes dans Oracle SQL
  + o Le script de création des tables externes correspondantes Hadoop HDFS dans Oracle SQL.
  + o Le programme java de chargement des données dans la base Oracle NoSQL.
  + o Le script de chargement des données dans Hadoop HDFS.
  + o Toute autre information justifiant de votre travail.
  + • Scripts pour l’architecture 2 o Le script de chargement des données dans Hadoop HDFS.
  + Scripts de comparaison des performances entre l’Architecture 1 et l’Architecture 2. Montrer les avantages et inconvénients de chacune de ces architectures.