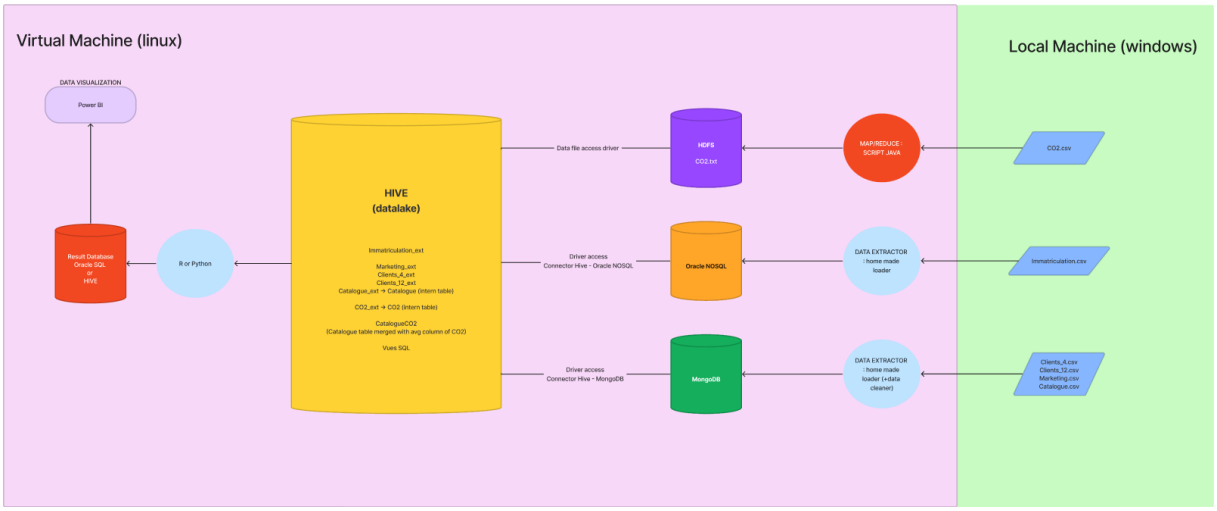
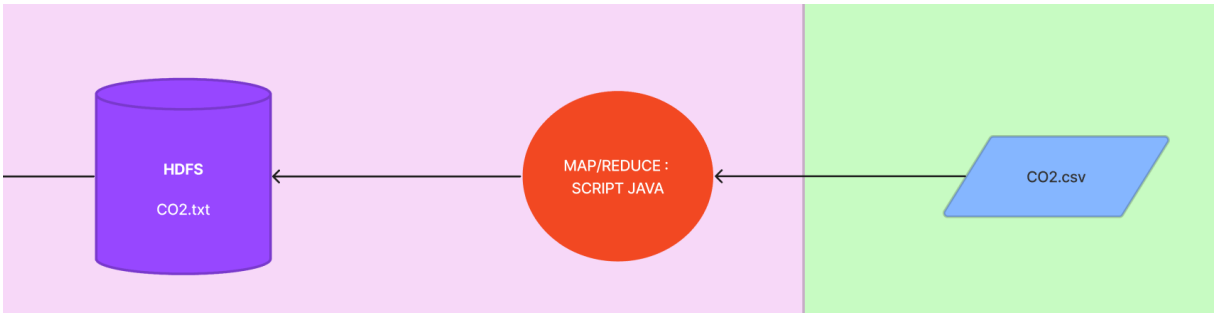
Architecture DBA / DL



Architecture de notre projet

Première étape



Lien entre CO2.csv et sa zone de stockage (HDFS)

Tout commence avec 6 fichiers csv. Nous allons implémenter ces différents fichiers dans 3 zones de stockage différentes.

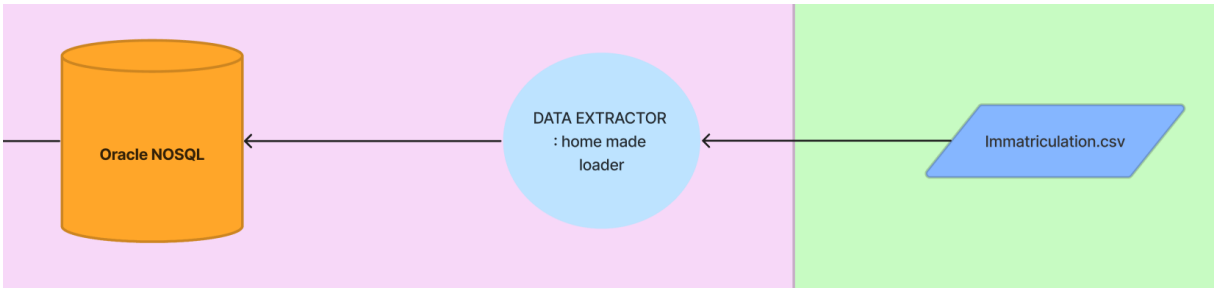
Pour le fichier CO2.csv, nous n'avions pas le choix de la zone de stockage (HDFS) sur lequel sera réalisé un programme de type Map/Reduce (qui est un script Java) pour nettoyer, formater et adapter ce dernier selon les demandes du cahier des charges.

Pour le fichier Immatriculation.csv, nous allons extraire ses données via un extracteur de données maison (expliqué dans le rapport concerné). Ces données seront ensuite stockées dans Oracle NoSQL.

Enfin, pour les fichiers Clients_4.csv, Clients_12.csv, Marketing.csv, Catalogue.csv, nous allons utiliser de la même façon que pour Immatriculation.csv un extracteur de données maison. En outre, nous

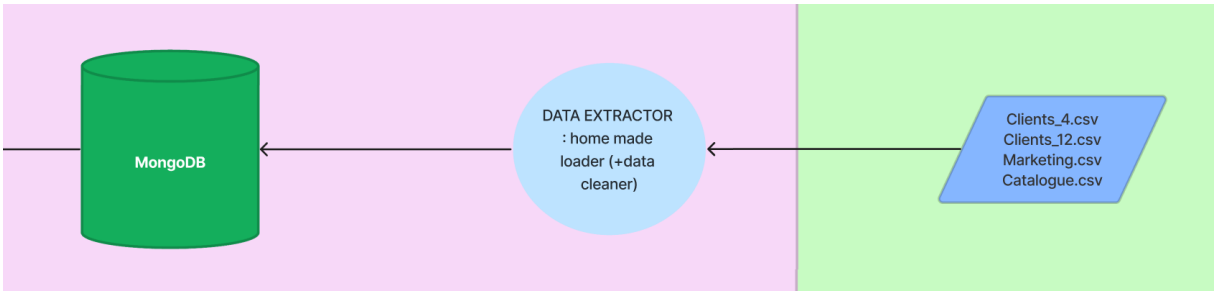
utilisons un nettoyeur de données pour les fichiers Clients, afin de pouvoir travailler avec les données les plus « propres » possibles. Toutes ces données seront ensuite stockées dans des tables MongoDB.

Deuxième étape



Lien entre Immatriculation.csv et sa zone de stockage (OracleNoSQL)

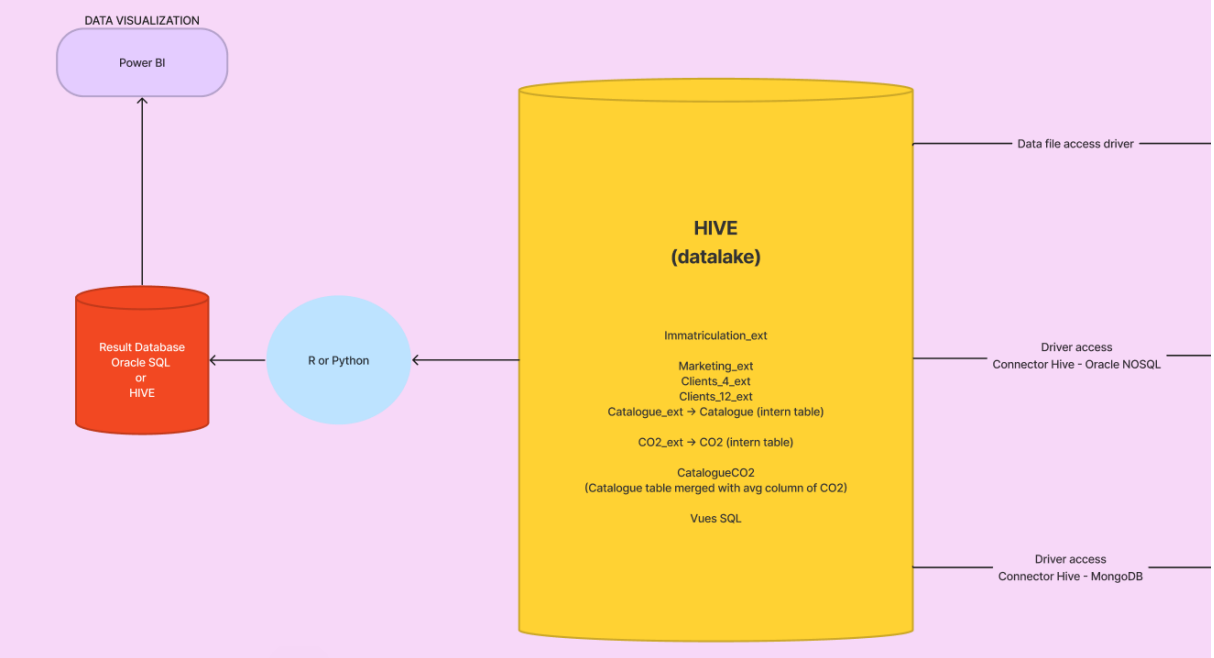
Une fois toutes ces données triées, nettoyées et insérées dans des tables, nous allons tout récupérer dans Hive via des Driver Access, Hive représentant ici notre DataLake. Ainsi Hive récupère les données de HDFS avec un DataFile Access Driver, dans Oracle NoSQL avec un Access Driver et dans MongoDB également avec un Access Driver.



Lien entre les quatre csv et leurs zones de stockage (MongoDB)

Nous avons donc dans notre DataLake Hive les tables Immatriculation_ext, Marketing_ext, Clients_4_ext, Clients_12_ext, Catalogue_ext que l'on transforme en table interne Catalogue, CO2_ext que l'on transforme en table interne CO2, la table fusionnée CatalogueCO2 et des vues SQL (tables virtuelles).

Troisième étape



Liens entre le datalake et les différentes zones de stockage et entre le datalake et la partie analyse

Nous allons ensuite insérer les résultats dans notre datalake via le langage R. On récupère les tables à l'aide de requêtes SQL afin de faire la partie analyse. Enfin, les données de résultats pourront être insérées puis visualisées à partir de Hive.

Selon l'avancée du projet, une partie visualisation sera réalisée grâce à Power BI.