## Plan type d’un rapport de projet Travail Pratique d’Approfondissement (TPA) Big Data et Analyse de données



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

Février 2023 – Mai 2023

Etudiants :

HAJANIRIVO Sitraka -12

RAKOTOARISOA Brian Ulrich - 21

RAKOTOSON Maminirina Fabien - 31

RANAIVOSON Fenohasina Antonio - 40

Enseignants :

GABRIEL MOPOLO MOKE – Big Data

SERGIO SIMONIAN

NICOLAS PASQUIER

Table des illustrations:

Figure 1: Architecture Datalake groupe 13 6

Table des matières

[Plan type d’un rapport de projet Travail Pratique d’Approfondissement (TPA) Big Data et Analyse de données 1](#_Toc134657559)

[1. Introduction générale (pas plus d’1 page) 4](#_Toc134657560)

[2. Présentation du projet 4](#_Toc134657561)

[3. Répartition du travail en membre du groupe 4](#_Toc134657562)

[4. Architecture du data lake 4](#_Toc134657563)

[4.1 Les sources de données 6](#_Toc134657564)

[4.1.1 MongoDb 6](#_Toc134657565)

[4.1.2 HDFS 6](#_Toc134657566)

[4.1.3 OracleNosql 6](#_Toc134657567)

[4.2 Les données du datalake 6](#_Toc134657568)

[5. Construction du data lake par étape 7](#_Toc134657569)

[5.1 Collecte des données 7](#_Toc134657570)

[6. Hadoop Map Reduce 7](#_Toc134657571)

[7. Analyse de données avec des outils de machine learning (R, …) 7](#_Toc134657572)

[7.1 Analyse préparatoire 7](#_Toc134657573)

[8. Conclusion générale 7](#_Toc134657574)

[9. Références et Bibliographie 7](#_Toc134657575)

[10. Annexes 8](#_Toc134657576)

[10.1 Vidéo de présentation de votre projet 8](#_Toc134657577)

[10.2 Dossier contenant les scripts et programmes de construction du lac de données 8](#_Toc134657578)

[10.2.1 Sources de données 8](#_Toc134657579)

[10.2.2 Script de création du datalake 9](#_Toc134657580)

[10.3 Dossier contenant les scripts et programmes Hadoop Map Reduce 9](#_Toc134657581)

[10.4 Dossier contenant les scripts et programmes d’analyse de données 10](#_Toc134657582)

# Introduction générale

De nos jours, les algorithmes font partie du quotidien, nous leurs déléguons des tâches pour nous faciliter un peu plus chaque jour notre vie.

Elles nous aident à prendre les décisions les plus influentes sur nos activités tant professionnelles que privées.

C’est donc tout naturellement qu’elles pénètrent petit à petit dans chaque métier qui existe.

# **Présentation** du projet

Il s'agit d'un projet où notre équipe fournira une solution à un concessionnaire afin de lui permettre de mieux cerner les besoins d’un client se présentant à lui.

En résumé, il voudrait savoir quelle voiture dans son catalogue correspond le mieux à un client donné. Pour ce faire, nous avons rassemblé les données provenant de ce concessionnaire dans un datalake.

Ces données sont fournies sous la forme de fichiers csv. Les fichiers csv sont répartis entre plusieurs sources de données, on accède ensuite à ces données dans le datalake.

# Répartition du travail en membre du groupe

Nous nous sommes répartis le travail comme suit :

Pour la partie import des données dans chaque source :

Sitraka s’est chargé d’écrire les scripts d’import des données de clients et co2 dans HDFS

Brian s’est chargé de l’import des données dans Mongodb et immatriculations dans HDFS

Antonio s’est chargé de l’import des données dans Oracle Nosql

Fabien s’est chargé du programme Hadoop pour traiter le fichier co2.csv

# Architecture du data lake

L’architecture du Data lake est la suivante :

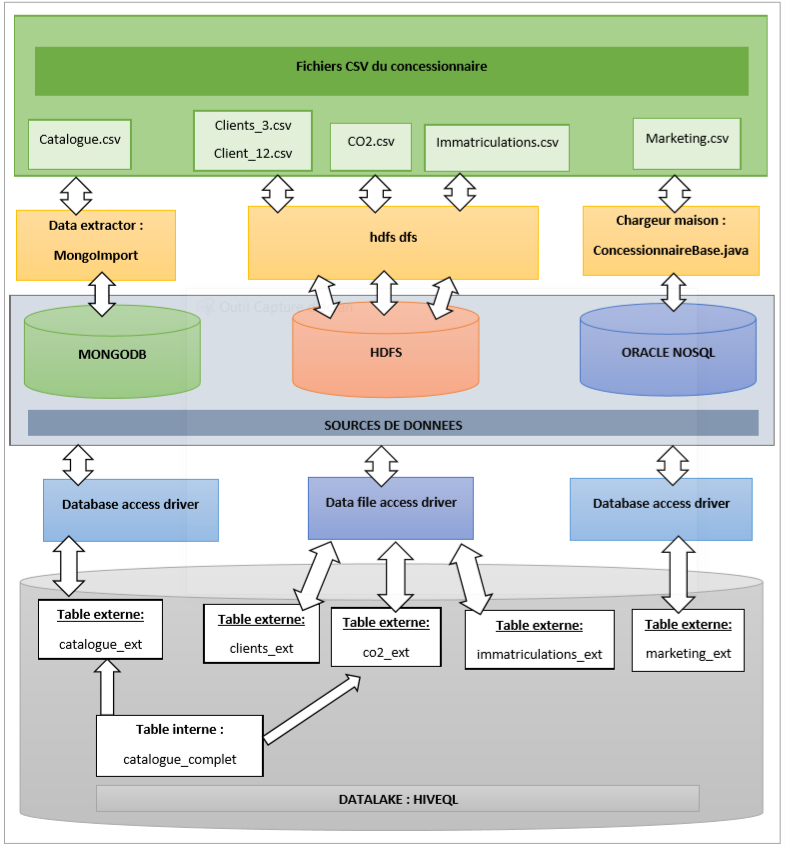


Figure : Architecture du Datalake - groupe 1

## Les sources de données

Nous avons décidé d’utiliser les sources de données suivantes :

### MongoDb

1. Nous avons décidé que les données du Catalogue seront dans la source de données MongoDb. Ces données sont enregistrées dans la collection « catalogue » de la base concessionnaire\_g13.
2. Alimentation :

L’alimentation de la source se fait avec l’extracteur de données MongoImport de MongoDb.

### HDFS

1. Les différents fichiers qu’on a placés dans cette source :

* Le fichier co2.csv : ce fichier étant à traiter par hadoop, le résultat de sortie doit donc se trouver dans dans HDFS. Donc naturellement, le fichier co2 après traitement a donc été placé dans hdfs.
* Les fichiers Clients\_3.csv et Clients\_12.csv
* Le fichier Immatriculations.csv

1. Alimentation :

L’alimentation de cette source se fait via l’utilitaire hadoop fs.

### OracleNosql

1. Le fichier contenant les données des clients séléctionnés par le service Marketing y a été placé.
2. Alimentation :

L’alimentation de cette source se fait avec un chargeur maison, plus précisément un programme Java.

## Les données du datalake

Le datalake est implémenté avec HIVEQL, les données y sont rendues accessibles avec les tables externes pointant vers les différentes sources de données.

Les données venant de mongoDb et OracleNosql sont rendues accessibles avec l’utilisation des Database access driver lors de la création des tables externes tandisque l’accès aux données de HDFS se fait via le data file access driver.

La table interne catalogue\_complet est la table catalogue\_ext à laquelle on a ajouté les colonnes d’informations sur les données de CO2.

# Construction du datalake par étape

## Collecte des données

En ce qui concerne la collecte des données, nous avons décidé dans un premier temps de placer le fichier Immatriculations.csv dans la source Oracle Nosql, mais le fichier étant trop volumineux, l’extraction et l’enregistrement des donnés se faisait entre 15 et 20 minutes contre 30 secondes dans HDFS, c’est pourquoi nous l’avons déplacé dans HDFS.

# Hadoop Map Reduce

# Analyse de données avec des outils de machine learning (R, …)

## Analyse préparatoire

# Conclusion générale

# Références et Bibliographie

[1] GABRIEL MOPOLO MOKE, 4Concepts\_Des\_Lacs\_Donnees\_Et\_Mise\_en\_Pratique.pdf, Edition, 2022 / 2023

[2] MongoDB, Using MongoDb with Hadoop & Spark : Part 2 – Hive Example, [www.mongodb.com](http://www.mongodb.com), 2023, <https://www.mongodb.com/blog/post/using-mongodb-hadoop-spark-part-2-hive-example>

[3] MongoDB, How to import Data into MongoDb with mongoimport, [www.mongodb.com](http://www.mongodb.com), 2023, https://www.mongodb.com/developer/products/mongodb/mongoimport-guide

[4] Benjamin Renaut mis à jour par Sergio Simonian, HadoopSparkMapReduce\_1.pdf, 2023

# Annexes

## Vidéo de présentation de votre projet

La vidéo de présentation est disponible sur le lien suivant :

[https://drive.google.com/drive/u/0/folders/1OaYHHzKWiKqb2WJVcr-bBoqPSIKh2M8g](https://drive.google.com/drive/u/0/folders/1OaYHHzKWiKqb2WJVcr-bBoqPSIKh2M8g?fbclid=IwAR0M_vZF0SiVx3SAOc1no860HcULjfKmstTJ0hYfDDu2j3hF-cH_UerXkTE)

## Dossier contenant les scripts et programmes de construction du lac de données

### Sources de données

Les fichiers csv du concessionnaire ont été stockés dans la racine du projet dans le répertoire « csv » avec les sous répertoires suivants :

* catalogue : contenant les fichiers csv du catalogue de véhicules
* clients : contentant les fichiers csv clients concernant les achats de l’année en cours
* co2 : contennant le fichier csv d’informations sur les rejets et coûts CO2 par véhicule
* immatriculations : contenant le fichier csv des immatriculations de l’année
* marketing : contentant le fichier csv clients séléctionnés par le service Marketing

Le script pour initialiser toutes les sources de données est placé dans la racine du projet : « init\_sources.sh »

Le répertoire « alimentation\_sources » dans la racine du projet représente le répertoire de base contenant les scripts d’alimentation pour chaque source de données. Un sous répertoire a été créé pour chaque source de données.

#### Source de données HDFS

Le répertoire « concessionnaire\_g13/alimentation\_sources/hdfs » contient les scripts d’alimentation de cette source.

Dans cette source nous avons les fichiers CO2, les fichiers clients ainsi que ceux des immatriculations.

Nous avons créé un répertoire de base dans HDFS pour contenir nos fichiers : « /groupe13 »

Pour chaque fichier, un fichier de script bash a été créé permettant de le transférer vers un sous répertoire hdfs de /groupe13.

Les trois fichiers suivants y ont donc été placés :

* put\_co2\_files\_hdfs.sh : pour le transfert des fichiers csv CO2 vers le répertoire hdfs « co2 »
* put\_clients\_files\_hdfs.sh : pour le transfert des fichiers csv clients vers le répertoire hdfs « clients »
* put\_immatriculations\_fils\_hdfs.sh : pour le transfert des fichiers csv immatriculations vers le répertoire hdfs « immatriculations »

#### Source de données MongoDb

Le répertoire « concessionnaire\_g13/alimentation\_sources/mongo » contient les scripts d’alimentation de cette source.

Dans cette source nous avons les données du fichier catalogue.

Le répertoire « mongo » contient donc le fichier de script bash « import\_catalogue\_file.sh » chargé d’importer les données du fichier catalogue vers la collection « catalogue » de la base « concessionnaire\_g13 »

#### Source de données Oracle Nosql

Le répertoire « concessionnaire\_g13/alimentation\_sources/oraclenosql » contient le programme d’alimentation de cette source.

Dans cette source, nous avons le fichier de données des clients séléctionnés par le service marketing.

Le répertoire « oraclenosql » contient donc le programme Java ConcessionnaireBase.java chargé d’extraire les données marketing vers la table « MARKETING\_G13 » de kvstore.

### Script de création du datalake

Le script de création du datalake a été placé dans le répértoire « datalake\_hive » dans la racine du projet.

Ce répertoire contient donc le fichier « concessionnaire\_hive\_datalake.sql ». Le script de création de la base « concessionnaire\_dl\_g13 » dans HIVEQL et de création des tables externes pointant vers les sources de données :

* cataloguue\_ext : les données du catalogue depuis MONGODB
* clients\_ext : les données clients depuis HDFS
* co2\_ext : les données CO2 après traitement depuis HDFS
* immatriculations\_ext : les données des immatriculations depuis HDFS

## Dossier contenant les scripts et programmes Hadoop Map Reduce

Le Programme Hadoop Map Reduce a été placé dans le répertoire « hadoop » dans la racine du projet.

Ce répertoire contient :

* Le fichier « co2\_groupe13.rar » : les sources du programme hadoop
* Le fichier « co2\_groupe13-0.0.1.jar » : le programme sous forme d’executable
* Le fichier « traitement\_co2.sh » : le script permettant de lancer le traitement sur le fichier CO2.

## Dossier contenant les scripts et programmes d’analyse de données

Le répertoire contenant les scripts et programmes d’analyse de données se trouve dans le dossier R.

Ce dossier contient :

* Le fichier AnalyseExploratoire\_G13.r : L’analyse exploratoire