PROJET BIG DATA ANALYTICS

Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèles de Véhicules

MBDS ESTIA

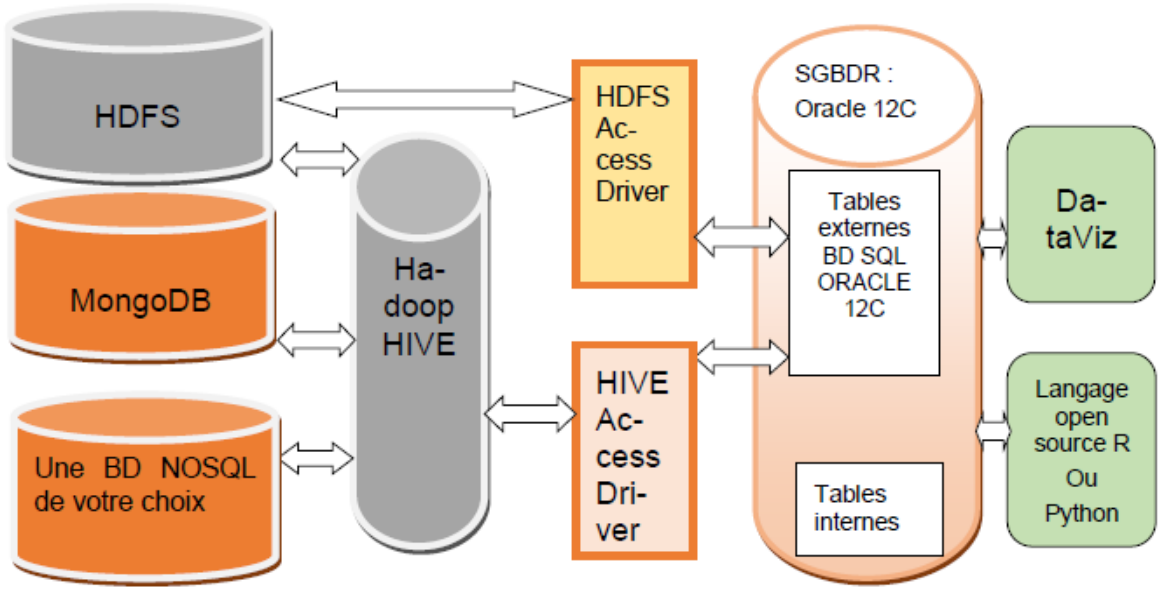
Lucie CARRON – Inès KERGALE – Lilian BOURGEOIS – Augustin AUBERT

Table des matières

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 (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 :

- Le fichier MARKETING devra être chargé sur la base NOSQL Oracle
- Les fichiers CATALOGUE et IMMATRICULATIONS doivent être des fichiers hadoop HDFS
- Notre fichier CLIENTS sera chargé sur la base Oracle SQL comme table interne.

ORACLE NOSQL

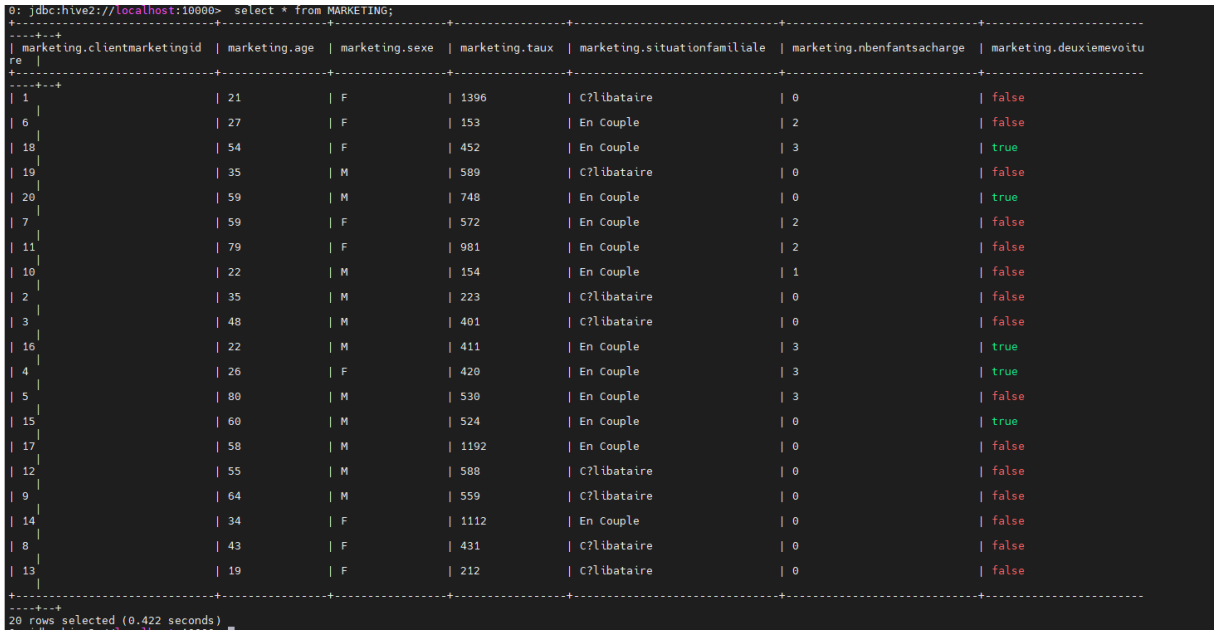
En premier lieu on commence a créer des dossiers et importer nos fichiers csv dans notre vm.
Ensuite nous lançons notre code java nommé DataImportMarketing.java qui va importer la table marketing à partir du fichier csv.
Puis nous nous connectons à oracle No sql afin de vérifier que la table marketing est bien remplie.

```
kv-> get table -name MARKETING
{"CLIENTMARKETINGID":5,"AGE":"80","SEXE":"M","TAUX":"530","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"3","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":15,"AGE":"60","SEXE":"M","TAUX":"524","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"true"}
{"CLIENTMARKETINGID":17,"AGE":"58","SEXE":"M","TAUX":"1192","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":9,"AGE":"64","SEXE":"M","TAUX":"559","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":14,"AGE":"34","SEXE":"F","TAUX":"1112","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":8,"AGE":"43","SEXE":"F","TAUX":"431","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":13,"AGE":"19","SEXE":"F","TAUX":"212","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":7,"AGE":"59","SEXE":"F","TAUX":"572","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"2","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":11,"AGE":"79","SEXE":"F","TAUX":"981","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"2","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":1,"AGE":"21","SEXE":"F","TAUX":"1396","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":6,"AGE":"27","SEXE":"F","TAUX":"153","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"2","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":18,"AGE":"54","SEXE":"F","TAUX":"452","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"3","DEUXIEMEVOITURE":"true"}
{"CLIENTMARKETINGID":19,"AGE":"35","SEXE":"M","TAUX":"589","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":20,"AGE":"59","SEXE":"M","TAUX":"748","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"true"}
{"CLIENTMARKETINGID":10,"AGE":"22","SEXE":"M","TAUX":"154","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"1","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":2,"AGE":"35","SEXE":"M","TAUX":"223","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":3,"AGE":"48","SEXE":"M","TAUX":"401","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}
{"CLIENTMARKETINGID":16,"AGE":"22","SEXE":"M","TAUX":"411","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"3","DEUXIEMEVOITURE":"true"}
{"CLIENTMARKETINGID":4,"AGE":"26","SEXE":"F","TAUX":"420","SITUATIONFAMILIALE":"En Couple","NBENFANTSACHARGE":"3","DEUXIEMEVOITURE":"true"}
{"CLIENTMARKETINGID":12,"AGE":"55","SEXE":"M","TAUX":"588","SITUATIONFAMILIALE":"C@libataire","NBENFANTSACHARGE":"0","DEUXIEMEVOITURE":"false"}

20 rows returned
```

Pour la partie NoSQL c'est ok.
Nous avons fait la première partie du chemin et devons donc maintenant créer une table externe pointant vers la table marketing d'oracle nosql.
Pour cela nous nous connectons à HIVE :

```
[oracle@bigdatalite ~]$ beeline
Beeline version 1.1.0-cdh5.4.0 by Apache Hive

-- Se connecter à HIVE
beeline> !connect jdbc:hive2://localhost:10000

Enter username for jdbc:hive2://localhost:10000: oracle
Enter password for jdbc:hive2://localhost:10000: ********
(password : welcome1)

-- Supprimer la table MARKETING si elle existe déjà
jdbc:hive2://localhost:10000> drop table MARKETING;

 /*

No rows affected (0.101 seconds)

*/

-- Création de la table externe MARKETING pointant vers la table MARKETING de ORACLE NOSQL
jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE MARKETING(
    CLIENTMARKETINGID int,
    AGE string ,
    SEXE string,
    TAUX string,
    SITUATIONFAMILIALE string,
    NBENFANTSACHARGE string,
    DEUXIEMEVOITURE string
)
STORED BY 'oracle.kv.hadoop.hive.table.TableStorageHandler'
TBLPROPERTIES (
"oracle.kv.kvstore" = "kvstore",
"oracle.kv.hosts" = "bigdatalite.localdomain:5000",
"oracle.kv.hadoop.hosts" = "bigdatalite.localdomain/127.0.0.1",
"oracle.kv.tableName" = "MARKETING");
```

La table externe est bien remplie, vérification :



```
0: jdbc:hive2://localhost:10000>  select * from MARKETING;
+-----------------------------+---------------+----------------+----------------+-----------------------------+---------------------------+-----------------------
----+--+
| marketing.clientmarketingid  | marketing.age  | marketing.sexe  | marketing.taux  | marketing.situationfamiliale  | marketing.nbenfantsacharge  | marketing.deuxiemevoitu
re  |
+-----------------------------+---------------+----------------+----------------+-----------------------------+---------------------------+-----------------------
----+--+
| 1                           | 21            | F              | 1396           | C?libataire                 | 0                         | false
   |
| 6                           | 27            | F              | 153            | En Couple                   | 2                         | false
   |
| 18                          | 54            | F              | 452            | En Couple                   | 3                         | true
   |
| 19                          | 35            | M              | 589            | C?libataire                 | 0                         | false
   |
| 20                          | 59            | M              | 748            | En Couple                   | 0                         | true
   |
| 7                           | 59            | F              | 572            | En Couple                   | 2                         | false
   |
| 11                          | 79            | F              | 981            | En Couple                   | 2                         | false
   |
| 10                          | 22            | M              | 154            | En Couple                   | 1                         | false
   |
| 2                           | 35            | M              | 223            | C?libataire                 | 0                         | false
   |
| 3                           | 48            | M              | 401            | C?libataire                 | 0                         | false
   |
| 16                          | 22            | M              | 411            | En Couple                   | 3                         | true
   |
| 4                           | 26            | F              | 420            | En Couple                   | 3                         | true
   |
| 5                           | 80            | M              | 530            | En Couple                   | 3                         | false
   |
| 15                          | 60            | M              | 524            | En Couple                   | 0                         | true
   |
| 17                          | 58            | M              | 1192           | En Couple                   | 0                         | false
   |
| 12                          | 55            | M              | 588            | C?libataire                 | 0                         | false
   |
| 9                           | 64            | M              | 559            | C?libataire                 | 0                         | false
   |
| 14                          | 34            | F              | 1112           | En Couple                   | 0                         | false
   |
| 8                           | 43            | F              | 431            | C?libataire                 | 0                         | false
   |
| 13                          | 19            | F              | 212            | C?libataire                 | 0                         | false
   |
+-----------------------------+---------------+----------------+----------------+-----------------------------+---------------------------+-----------------------
----+--+
20 rows selected (0.422 seconds)
```

Dernière partie du chemin : la création de tables externes Oracle SQL pointant vers les tables externes HIVE.

```
--Supprimer les tables si elles existent

drop table MARKETING_O_EXT;

--créer la table MARKETING_O_EXT

CREATE TABLE MARKETING_O_EXT(
CLIENTMARKETINGID number,
AGE varchar2(40),
SEXE varchar2(40),
TAUX varchar2(40),
SITUATIONFAMILIALE varchar2(40),
NBENFANTSACHARGE varchar2(40),
DEUXIEMEVOITURE varchar2(40)
)
ORGANIZATION EXTERNAL (
TYPE ORACLE_HIVE
DEFAULT DIRECTORY ORACLE_BIGDATA_CONFIG
ACCESS PARAMETERS
(
com.oracle.bigdata.tablename=default.MARKETING
)
)
REJECT LIMIT UNLIMITED;
```

Table créée.

```
SQL> select * from MARKETING_O_EXT
  2  ;

CLIENTMARKETINGID AGE                  SEXE                 TAUX                 SITUATIONFAMILIALE   NBENFANTSACHARGE     DEUXIEMEVOITURE
----------------- -------------------- -------------------- -------------------- -------------------- -------------------- --------------------
      ########## 59                   F                    572                  En Couple            2                    false
      ########## 79                   F                    981                  En Couple            2                    false
      ########## 21                   F                    1396                 C?libataire          0                    false
      ########## 27                   F                    153                  En Couple            2                    false
      ########## 54                   F                    452                  En Couple            3                    true
      ########## 35                   M                    589                  C?libataire          0                    false
      ########## 59                   M                    748                  En Couple            0                    true
      ########## 22                   M                    154                  En Couple            1                    false
      ########## 35                   M                    223                  C?libataire          0                    false
      ########## 48                   M                    401                  C?libataire          0                    false
      ########## 22                   M                    411                  En Couple            3                    true

CLIENTMARKETINGID AGE                  SEXE                 TAUX                 SITUATIONFAMILIALE   NBENFANTSACHARGE     DEUXIEMEVOITURE
----------------- -------------------- -------------------- -------------------- -------------------- -------------------- --------------------
      ########## 26                   F                    420                  En Couple            3                    true
      ########## 80                   M                    530                  En Couple            3                    false
      ########## 60                   M                    524                  En Couple            0                    true
      ########## 58                   M                    1192                 En Couple            0                    false
      ########## 55                   M                    588                  C?libataire          0                    false
      ########## 43                   F                    431                  C?libataire          0                    false
      ########## 19                   F                    212                  C?libataire          0                    false
      ########## 64                   M                    559                  C?libataire          0                    false
      ########## 34                   F                    1112                 En Couple            0                    false

20 rows selected.
```

Ça y est nous avons notre table Marketing_O_EXT créée sur Oracle SQL.

HDFS

Nous allons maintenant créer les tables Oracle SQL à partir des fichiers Immatriculations.csv et catalogue.csv.

Nous avons choisi ces 2 fichiers là car ce sont les plus volumineux.

Nous avons créé sur hdfs des dossiers afin de stocker les fichiers csv.

```
--Créer un répertoire pour y mettre les fichiers:
hdfs dfs -mkdir /groupe5

--Pour l'architecture 1:
hdfs dfs -mkdir /groupe5/archi1
hdfs dfs -mkdir /groupe5/archi1/dossier_catalogue
hdfs dfs -mkdir /groupe5/archi1/dossier_immatriculations
```

Puis, de la même manière que pour Oracle NOSQL nous avons créé des tables externes sur HIVE pointant vers HDFS puis en dernière étape, des tables externes Oracle SQL pointant vers les tables externes HIVE.

```
CREATE EXTERNAL TABLE  IMMATRICULATIONS_HDFS_EXT  (IMMATRICULATION string, MARQUE string, NOM string,
PUISSANCE int, LONGUEUR string, NBPLACES int, NBPORTES int, COULEUR string, OCCASION string, PRIX int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION 'hdfs:/groupe5/archi1/dossier_immatriculations';
```

```
CREATE EXTERNAL TABLE  CATALOGUE_HDFS_EXT  (MARQUE string, NOM string, PUISSANCE int,
 LONGUEUR string, NBPLACES int, NBPORTES int, COULEUR string, OCCASION string, PRIX int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION 'hdfs:/groupe5/archi1/dossier_catalogue';
```

```
0: jdbc:hive2://localhost:10000> desc IMMATRICULATIONS_HDFS_EXT;
+-----------------+------------+----------+--+
|     col_name    | data_type  | comment  |
+-----------------+------------+----------+--+
| immatriculation | string     |          |
| marque          | string     |          |
| nom             | string     |          |
| puissance       | int        |          |
| longueur        | string     |          |
| nbplaces        | int        |          |
| nbportes        | int        |          |
| couleur         | string     |          |
| occasion        | string     |          |
| prix            | int        |          |
+-----------------+------------+----------+--+
10 rows selected (0.062 seconds)
```

```
0: jdbc:hive2://localhost:10000> desc CATALOGUE_HDFS_EXT;
+------------+------------+----------+--+
|  col_name  | data_type  | comment  |
+------------+------------+----------+--+
| marque     | string     |          |
| nom        | string     |          |
| puissance  | int        |          |
| longueur   | string     |          |
| nbplaces   | int        |          |
| nbportes   | int        |          |
| couleur    | string     |          |
| occasion   | string     |          |
| prix       | int        |          |
+------------+------------+----------+--+
9 rows selected (0.06 seconds)
```

Puis depuis sqlplus :

```
--créer la table IMMATRICULATIONS

CREATE TABLE  IMMATRICULATIONS_O_EXT(
IMMATRICULATION varchar2(10),
MARQUE varchar2(40),
NOM varchar2(40),
PUISSANCE number,
LONGUEUR varchar2(40),
NBPLACES number,
NBPORTES number,
COULEUR varchar2(40),
OCCASION varchar2(40),
PRIX number
)
ORGANIZATION EXTERNAL (
TYPE ORACLE_HIVE
DEFAULT DIRECTORY ORACLE_BIGDATA_CONFIG
ACCESS PARAMETERS
(
com.oracle.bigdata.tablename=default.IMMATRICULATIONS_HDFS_EXT
)
)
REJECT LIMIT UNLIMITED;

--créer la table CATALOGUE_O_EXT

CREATE TABLE CATALOGUE_O_EXT(
MARQUE varchar2(40),
NOM varchar2(40),
PUISSANCE number,
LONGUEUR varchar2(40),
NBPLACES number,
NBPORTES number,
COULEUR varchar2(40),
OCCASION varchar2(40),
PRIX number
)
ORGANIZATION EXTERNAL (
TYPE ORACLE_HIVE
DEFAULT DIRECTORY ORACLE_BIGDATA_CONFIG
ACCESS PARAMETERS
(
com.oracle.bigdata.tablename=default.CATALOGUE_HDFS_EXT
)
)
REJECT LIMIT UNLIMITED;
```

Vérifications :

```
SQL> SELECT count(*) from IMMATRICULATIONS_O_EXT;

  COUNT(*)
----------
   2000001

SQL> SELECT count(*) from CATALOGUE_O_EXT;

  COUNT(*)
----------
       271
```

Cela correspond bien aux données des fichiers CSV immatriculations et catalogue.

ORACLE SQL

Nous créons directement la table CLIENT sur sqlplus :

CREATE TABLE CLIENT(AGE varchar2(30), SEXE varchar2(30), TAUX varchar2(30), SITUATIONFAMILIALE varchar2(30), NBENFANTSACHARGE varchar2(30), XVOITURE varchar2(30), IMMATRICULATION varchar2(30));

Puis via sqlloader nous chargeons les données du fichier csv.

```
sqlldr  userid=CARONBZ2021@ORCL/CARONBZ202101
control=$MYPROJECTHOME/sqlloader/control/control_client.txt
log=$MYPROJECTHOME/sqlloader/log/client_log.log errors=0 skip=1
```

```
Commit point reached - logical record count 99501
Commit point reached - logical record count 99565
Commit point reached - logical record count 99629
Commit point reached - logical record count 99693
Commit point reached - logical record count 99757
Commit point reached - logical record count 99821
Commit point reached - logical record count 99885
Commit point reached - logical record count 99949
Commit point reached - logical record count 100000

Table CLIENT:
  100000 Rows successfully loaded.

Check the log file:
  /home/CARRON/Projet3AMBDS/sqlloader/log/client_log.log
for more information about the load.
```

On vérifie ensuite :

```
SQL> SELECT * FROM CLIENT OFFSET 0 ROWS FETCH NEXT 10 ROWS ONLY;

AGE        SEXE               TAUX               SITUATIONFAMILIALE   NBENFANTSA XVOITURE           IMMATRICULATION
---------- ------------------ ------------------ -------------------- ---------- ------------------ --------------------
62         M                  1262               En Couple            1          false              6290 DM 24
68         M                  514                En Couple            2          false              7530 VH 52
26         F                  181                En Couple            4          true               7168 HX 32
34         M                  829                C?libataire          0          false              1539 UR 49
50         M                  1169               En Couple            4          false              4738 YG 76
53         M                  156                En Couple            2          false              5176 NX 17
32         F                  566                C?libataire          0          false              8958 OW 94
30         M                  465                En Couple            2          true               1268 FW 51
20         M                  450                En Couple            0          true               3531 MD 12
74         M                  426                En Couple            2          false              8777 ZT 53

10 rows selected.
```

La table est bien remplie.

TABLES ORACLE SQL

```
set linesize 200
col OWNER format A30
col TABLE_NAME format A30
SELECT owner, table_name FROM dba_tables where owner='CARONBZ2021'
```

```
OWNER                          TABLE_NAME
------------------------------ ------------------------------
CARONBZ2021                    CATALOGUE_0_EXT
CARONBZ2021                    CLIENT
CARONBZ2021                    IMMATRICULATIONS_0_EXT
CARONBZ2021                    MARKETING_0_EXT
```

Les 4 tables ont bien été créée

Connexion à R

Nous pensions pouvoir utiliser R directement avec la VM mais nous ne savons pas l'utiliser comme ça sans interface donc nous tentons une connexion avec RStudio mais nous n'y arrivons malheureusement pas :

Nous avons créé un user afin d'assurer la connexion

CREATE USER &MYDBUSER IDENTIFIED BY &MYDBUSERPASS default tablespace users temporary tablespace temp;

old 1: CREATE USER &MYDBUSER IDENTIFIED BY &MYDBUSERPASS default tablespace users temporary tablespace temp

new 1: CREATE USER PROJET6 IDENTIFIED BY 123 default tablespace users temporary tablespace temp

User created.

SQL> grant dba to &MYDBUSER;

old 1: grant dba to &MYDBUSER

new 1: grant dba to PROJET6

Grant succeeded.

SQL> alter user &MYDBUSER quota unlimited on users;

old 1: alter user &MYDBUSER quota unlimited on users

new 1: alter user PROJET6 quota unlimited on users

User altered.

Puis on a tenté une connexion via RStudio

```
install.packages("RODBC")
library(RODBC)
#Connexion
connexion <- odbcConnect("ORCLPROJETDB6_DNS", uid="PROJET6", pwd="welcome1", believeNRows=FALSE)
```

```
> connexion <- odbcConnect("ORCLPROJETDB6_DNS", uid="PROJET6", pwd="welcome1", believeNRows=FALSE)
Warning messages:
1: In RODBC::odbcDriverConnect("DSN=ORCLPROJETDB6_DNS;UID=PROJET6;PWD=welcome1",  :
  [RODBC] ERROR: state IM002, code 0, message [Microsoft][Gestionnaire de pilotes ODBC] Source de données introuvable et nom de pilote non spécifié
2: In RODBC::odbcDriverConnect("DSN=ORCLPROJETDB6_DNS;UID=PROJET6;PWD=welcome1",  :
  ODBC connection failed
```

Connection failed

Nous décidons alors de lancer le projet rendu à Alison Temin en connectant avec RStudio notre base Oracle du début d'année (pas celle liée à la vm)

```
#charger les fichiers
#direction
setwd("C:/Users/l.carron/Documents/3a_S1/MBDS/COURS/6.DataScience/Projet/DATA")

#création de immatriculations car on ne peut pas utiliser une table Oracle : le fichier est trop volumineux
immatriculations <- read.csv("Immatriculations.csv" , header = TRUE, sep = ",", dec = ".")

install.packages("RJDBC")
library(RJDBC)


##classPath : add path to drivers jdbc

drv <- RJDBC::JDBC(driverClass = "oracle.jdbc.OracleDriver", classPath =  Sys.glob("C:/drivers/*"))


#Connexion
conn <- dbConnect(drv, "jdbc:oracle:thin:@(DESCRIPTION=(ADDRESS=(PROTOCOL=TCP)
                  (HOST=144.21.67.201)(PORT=1521))(CONNECT_DATA=
                  (SERVICE_NAME=pdbest21.631174089.oraclecloud.internal)))",
                  "CARON2B20", "CARON2B2001")

#Enregistrement de la table Marketing  dans Oracle
marketing <- read.csv("C:/Users/l.carron/Documents/3a_S1/MBDS/COURS/6.DataScience/Projet/DATA/Marketing.csv", header = TRUE,
                      sep = ",", dec = ".")

names(marketing)[6] = ("deuxiemeVoiture")

dbWriteTable(conn,"marketing",marketing,
             rownames=FALSE, overwrite = TRUE, append = FALSE)
```

```
#Enregistrement de la table Catalogue dans  Oracle
catalogue <- read.csv("C:/Users/l.carron/Documents/3a_S1/MBDS/COURS/6.DataScience/Projet/DATA/Catalogue.csv", header = TRUE,
                      sep = ",", dec = ".")


dbWriteTable(conn,"catalogue",catalogue,
             rownames=FALSE, overwrite = TRUE, append = FALSE)

#Enregistrement de la table Client dans  Oracle
client <- read.csv("C:/Users/l.carron/Documents/3a_S1/MBDS/COURS/6.DataScience/Projet/DATA/Clients_5.csv", header = TRUE,
                   sep = ",", dec = ".")

dbWriteTable(conn,"client5",client,
             rownames=FALSE, overwrite = TRUE, append = FALSE)


#Visualisation des tables
allTables <- dbGetQuery(conn, "SELECT owner, table_name FROM all_tables where
                        owner = 'CARON2B20'")

#A partir des tables créées sur oracle nous créons les dataframe :

marketing <- dbGetQuery(conn, "select * from marketing")
catalogue <- dbGetQuery(conn, "select * from catalogue")
client <- dbGetQuery(conn,"select * from client5")

#immatriculations est déjà créé
```

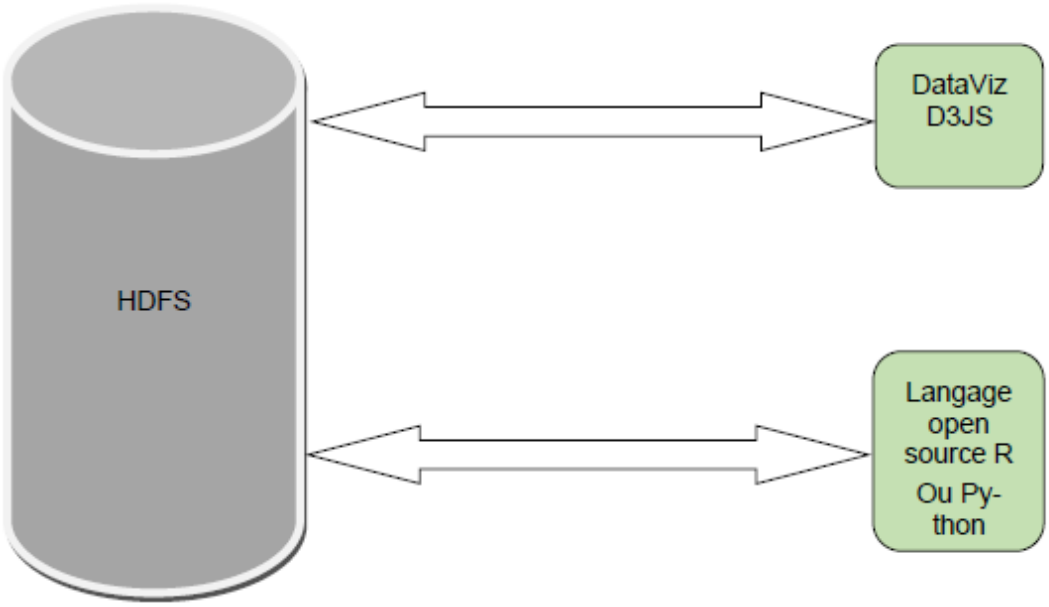
A partir de là nous nous retrouvons avec nos dataframe créés, prêts à être analysés.

Analyse exploratoire

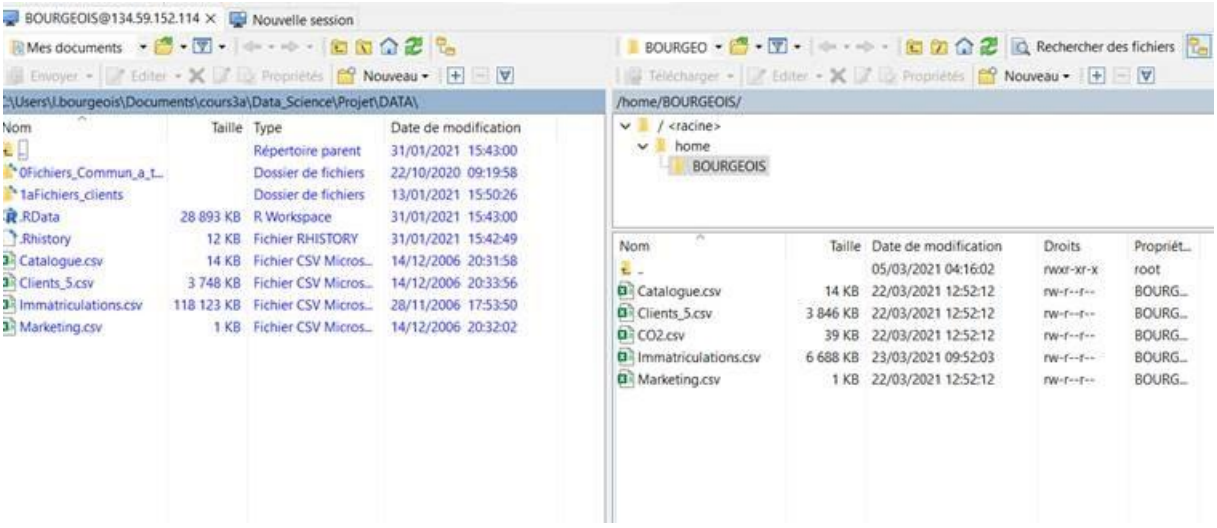
Cette partie est dans le rapport « Analyse_exploratoire » situé dans le dossier Analyse de donnees.

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.



Nous avons tout d'abord récupéré les fichiers csv et les avons stockés sur la machine (copié collé) :



Ensuite il a fallu créer un répertoire sur HDFS pour y stocker nos fichiers :

hdfs dfs -mkdir /groupe5/archi2

Puis enfin nous y avons placé nos fichiers à l'aide de la commande put :

hdfs dfs -put Marketing.csv /groupe5/archi2

hdfs dfs -put Catalogue.csv /groupe5/archi2

hdfs dfs -put CO2.csv /groupe5/archi2

hdfs dfs -put Clients_5.csv /groupe5/archi2

hdfs dfs -put Immatriculations.csv /groupe5/archi2

Pour vérifier on fait la commande :

hdfs dfs -ls /groupe5/archi2

```
[BOURGEOIS@bigdatalite ~]$ hdfs dfs -ls /groupe5/archi2
Found 5 items
-rw-r--r--   1 BOURGEOIS supergroup      39354 2021-03-23 10:00 /groupe5/archi2/CO2.csv
-rw-r--r--   1 BOURGEOIS supergroup      14114 2021-03-23 10:00 /groupe5/archi2/Catalogue.csv
-rw-r--r--   1 BOURGEOIS supergroup    3937752 2021-03-23 10:00 /groupe5/archi2/Clients_5.csv
-rw-r--r--   1 CARRON    supergroup  120957648 2021-03-23 09:53 /groupe5/archi2/Immatriculations.csv
-rw-r--r--   1 BOURGEOIS supergroup        638 2021-03-23 10:00 /groupe5/archi2/Marketing.csv
```

On peut aussi obtenir des informations sur l'état du système de fichiers avec la commande fsck:

hdfs fsck /

```
.....
.....
.........Status: HEALTHY
 Total size:    9712988042 B (Total open files size: 166 B)
 Total dirs:    1982
 Total files:   4209
 Total symlinks:                0 (Files currently being written: 3)
 Total blocks (validated):      4192 (avg. block size 2317029 B) (Total open file blocks (not validated): 2)
 Minimally replicated blocks:   4192 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       2379 (56.750954 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              21362 (83.59552 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Tue Mar 30 14:44:59 EDT 2021 in 92 milliseconds


The filesystem under path '/' is HEALTHY
```

hdfs fsck /groupe5

```
Connecting to namenode via http://bigdatalite.localdomain:50070/fsck?ugi=BOURGEOIS&path=%2Fgroupe5
FSCK started by BOURGEOIS (auth:SIMPLE) from /127.0.0.1 for path /groupe5 at Tue Mar 30 14:45:51 EDT 2021
.........Status: HEALTHY
 Total size:    366893030 B
 Total dirs:    5
 Total files:   9
 Total symlinks:                0
 Total blocks (validated):      12 (avg. block size 30574419 B)
 Minimally replicated blocks:   12 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Tue Mar 30 14:45:51 EDT 2021 in 0 milliseconds


The filesystem under path '/groupe5' is HEALTHY
```

Pour ce qui est de la connexion entre HDFS et R, nous n'avons pas non plus réussi à la faire donc nous considérons que l'analyse des données à partir des fichiers CSV sur Oracle (cf plus haut) est similaire.

Comparaison des architectures

Il est intéressant de comparer l'architecture 1 DATA LAKE avec l'architecture 2 HDFS même si il nous parait évident que la numéro 1 est plus complexe et plus longue à mettre en place, elle est cependant plus « réaliste » et similaire à celles que des entreprises peuvent mettre en place. Car effectivement il est plus simple de modifier les data lakes en ajoutant des données directement via les différents intermédiaires.

MAP REDUCE

Cf le rapport dans le dossier MapReduce et la structure de cette partie dans le dossier mapReduceProjet.