

# Rapport HADOOP MAP REDUCE

[Source Code](#)

[Archive](#)

[VPS](#)

[Explications](#)

[Hive](#)

## Source Code

Le code source du programme map/reduce utilisé pour l'adaptation du fichier CO2.csv et son intégration dans la table catalogue se trouve à l'emplacement suivant

## Archive

`/DATABASE/HDFS`

## VPS

`/root/tpa/tpa/DATABASE/HDFS`

Afin d'accéder au notebook pour executer les cellules depuis le VPS

- Ouvrir deux terminaux
  - Dans le premier terminal (connecté au VPS)

```
cd /root/tpa/tpa/DATABASE/HDFS
jupyter notebook --allow-root
```

- Dans le second terminal (non connecté au VPS, en local sur votre machine avec openssh, ne fonctionne pas sous windows)

```
ssh -L 8888:localhost:8888 root@85.31.239.246
```

## Explications

On utilisera ici Apache Spark avec l'API pyspark afin de réaliser les opérations map/reduce nécessaires.

Afin d'adapter le fichier CO2.csv on récupère d'abord le fichier depuis HDFS, on précisera ici que le fichier à déjà été ajouté au système de fichier HDFS grâce au script [init.sh](#)

```
csv_file_path = "/root/tpa/tpa/DATABASE/data/CO2.csv"
```

Une fois le fichier récupéré on transforme alors le csv en un dataframe spark afin d'en faciliter la manipulation.

```
custom_schema = StructType([
    StructField("", IntegerType(), True),
    StructField("Marque / Modele", StringType(), True),
    StructField("Bonus / Malus", StringType(), True),
    StructField("Rejets CO2 g/km", IntegerType(), True),
    StructField("Cout energie", StringType(), True),
])

df_co2 = spark.read.csv(csv_file_path, header=True, schema=custom_schema)
```

On a donc à notre disposition un data frame contenant les informations du fichier CO2.csv:

	Marque / Modele	Bonus / Malus	Rejets CO2 g/km	Cout energie
2	AUDI E-TRON SPORT...	-6 000€ 1	0	319 €
3	AUDI E-TRON SPORT...	-6 000€ 1	0	356 €
4	AUDI E-TRON 55 (4...	-6 000€ 1	0	357 €
5	AUDI E-TRON 50 (3...	-6 000€ 1	0	356 €
6	BMW i3 120 Ah	-6 000€ 1	0	204 €
7	BMW i3s 120 Ah	-6 000€ 1	0	204 €
8	CITROEN BERLINGO	-6 000€ 1	0	203 €
9	CITROEN C-ZERO	-6 000€ 1	0	491 €
10	DS DS3 CROSSBACK ...	-6 000€ 1	0	251 €
11	HYUNDAI KONA elec...	-6 000€ 1	0	205 €
12	HYUNDAI KONA elec...	-6 000€ 1	0	205 €
13	JAGUAR I-PACE EV4...	-6 000€ 1	0	271 €
14	KIA e-NIRO Moteur...	-6 000€ 1	0	212 €
15	KIA e-NIRO Moteur...	-6 000€ 1	0	203 €
16	KIA SOUL Moteur Ä...	-6 000€ 1	0	214 €
17	KIA SOUL Moteur Ä...	-6 000€ 1	0	214 €
18	MERCEDES EQC 400 ...	-6 000€ 1	0	291 €
19	MERCEDES VITO Tou...	-6 000€ 1	0	411 €
20	MERCEDES VITO Tou...	-6 000€ 1	0	411 €
21	MINI MINI Cooper ...	-6 000€ 1	0	199 €

Afin de réaliser les opérations sur les colonnes Bonus / Malus, Rejets CO2 et Cout energie, il nous faut d'abord nettoyer les données, en effet pour réaliser un calcul de moyenne nous allons transformer toute les valeurs en `int` pour la colonne cout energie, ainsi que pour la colonne Bonus / Malus, de plus pour cette dernière il nous faut retirer les informations en trop tel les 1 après la valeur du Bonus / Malus

```
# Ici on crée une colonne Marque afin d'extraire la marque du m
# Cela sera utile plus tard
df_co2 = df_co2.withColumn("Marque", F.split(df_co2["Marque / M

# Convert to integer
df_co2 = df_co2.withColumn("Cout energie", F.regexp_replace("Cout
df_co2 = df_co2.withColumn("Cout energie", df_co2["Cout energie"]

# Remove the extra characters after the euro sign and remove the
# Then convert to integer
pattern = r'([+-]?\\d+)\\s*€.*'
```

```
df_co2 = df_co2.withColumn("Bonus / Malus", F.regexp_replace(F.col("Bonus / Malus"), " ", ""))
df_co2 = df_co2.withColumn("Bonus / Malus", F.regexp_replace(F.col("Bonus / Malus"), " ", ""))

df_co2.show()
```

	Marque / Modele	Bonus / Malus	Rejets CO2 g/km	Cout energie	Marque
2	AUDI E-TRON SPORT...	-6000	0	319	AUDI
3	AUDI E-TRON SPORT...	-6000	0	356	AUDI
4	AUDI E-TRON 55 (4...	-6000	0	357	AUDI
5	AUDI E-TRON 50 (3...	-6000	0	356	AUDI
6	BMW i3 120 Ah	-6000	0	204	BMW
7	BMW i3s 120 Ah	-6000	0	204	BMW
8	CITROEN BERLINGO	-6000	0	203	CITROEN
9	CITROEN C-ZERO	-6000	0	491	CITROEN
10	DS DS3 CROSSBACK ...	-6000	0	251	DS
11	HYUNDAI KONA elec...	-6000	0	205	HYUNDAI
12	HYUNDAI KONA elec...	-6000	0	205	HYUNDAI
13	JAGUAR I-PACE EV4...	-6000	0	271	JAGUAR
14	KIA e-NIRO Moteur...	-6000	0	212	KIA
15	KIA e-NIRO Moteur...	-6000	0	203	KIA
16	KIA SOUL Moteur A...	-6000	0	214	KIA
17	KIA SOUL Moteur A...	-6000	0	214	KIA
18	MERCEDES EQC 400 ...	-6000	0	291	MERCEDES
19	MERCEDES VITO Tou...	-6000	0	411	MERCEDES
20	MERCEDES VITO Tou...	-6000	0	411	MERCEDES
21	MINI MINI Cooper ...	-6000	0	199	MINI

Pour obtenir les valeurs moyennes on va grouper les lignes par Marque et ainsi faire une moyenne de toute les valeurs

Enfin on retire la colonne Marque / Modele, seulement la marque du véhicule nous intéresse.

```
result_df = df_co2.groupBy("Marque").agg(
    F.round(F.avg("Bonus / Malus"), 2).alias("Avg_Bonus_Malus"),
    F.round(F.avg("Rejets CO2 g/km"), 2).alias("Avg_Rejets_CO2"),
    F.round(F.avg("Cout energie"), 2).alias("Avg_Cout_energie")
).withColumn("Marque", F.lower(F.col("Marque")))
```

Marque	Avg_Bonus_Malus	Avg_Rejets_CO2	Avg_Cout_energie
mercedes	8237.36	187.63	749.98
porsche	NULL	69.86	89.71
hyundai	-6000.0	8.67	151.0
toyota	NULL	32.0	43.0
skoda	-6000.0	27.56	98.89
nissan	5802.4	160.0	681.2
land	NULL	69.0	78.0
citroen	-6000.0	0.0	347.0
bentley	NULL	84.0	102.0
audi	-6000.0	26.1	191.6
mini	-6000.0	21.5	126.0
peugeot	-6000.0	15.83	144.17
jaguar	-6000.0	0.0	271.0
volvo	NULL	42.45	72.73
tesla	-6000.0	0.0	245.89
bmw	-6000.0	39.26	80.53
volkswagen	-6000.0	13.33	149.33
kia	-6000.0	10.33	157.67
smart	-6000.0	0.0	191.36
renault	-6000.0	0.0	206.0

On récupère ensuite nos informations provenant du catalogue, on se retrouve alors avec un second dataframe de la forme suivante:

marque	nom	puissance	longueur	nbPlaces	nbPortes	couleur	occasion	prix
volvo	S80 T6	272	très longue	5	5	blanc	false	50500
volvo	S80 T6	272	très longue	5	5	noir	false	50500
volvo	S80 T6	272	très longue	5	5	rouge	false	50500
volvo	S80 T6	272	très longue	5	5	gris	true	35350
volvo	S80 T6	272	très longue	5	5	bleu	true	35350
volvo	S80 T6	272	très longue	5	5	gris	false	50500
volvo	S80 T6	272	très longue	5	5	bleu	false	50500
volvo	S80 T6	272	très longue	5	5	rouge	true	35350
volvo	S80 T6	272	très longue	5	5	blanc	true	35350
volvo	S80 T6	272	très longue	5	5	noir	true	35350
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	false	27340
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	true	19138
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	true	19138
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	false	27340
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	false	27340
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	true	19138
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	true	19138
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	true	19138
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	false	27340
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	false	27340

Il ne reste plus qu'à joindre les deux dataframes afin d'ajouter les valeurs moyennes selon la marque de chaque véhicule

```
joined_df = df_catalogue.join(result_df, df_catalogue.marque ==
```

marque	nom	puissance	longueur	nbPlaces	nbPortes	couleur	occasion	prix	Avg_Bonus_Malus	Avg_Rejets_CO2	Avg_Cout_energie
volvo	S80 T6	272	très longue	5	5	blanc	false	50500	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	noir	false	50500	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	rouge	false	50500	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	gris	true	35350	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	bleu	true	35350	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	gris	false	50500	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	bleu	false	50500	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	rouge	true	35350	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	blanc	true	35350	NULL	42.45	72.73
volvo	S80 T6	272	très longue	5	5	noir	true	35350	NULL	42.45	72.73
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	false	27340	-6000.0	13.33	149.33

Pour ce qui est des valeurs NULL, elles sont ici car il n'y avait pas de valeurs pour les voitures de la marque volvo ou seat par exemple.

marque	nom	puissance	longueur	nbPlaces	nbPortes	couleur	occasion	prix	Avg_Bonus_Malus	Avg_Rejets_CO2	Avg_Cout_energie
seat	Toledo 1.6	102	longue	5	5	blanc	false	18880	NULL	NULL	NULL
seat	Toledo 1.6	102	longue	5	5	noir	false	18880	NULL	NULL	NULL
seat	Toledo 1.6	102	longue	5	5	rouge	false	18880	NULL	NULL	NULL
seat	Toledo 1.6	102	longue	5	5	gris	false	18880	NULL	NULL	NULL
seat	Toledo 1.6	102	longue	5	5	bleu	false	18880	NULL	NULL	NULL

On désire remplacer ces valeurs nulles par la moyenne des véhicules.

On crée alors un nouveau dataframe contenant les valeurs moyenne de chaque colonne.

```
global_avg_df = joined_df.agg(
    F.round(F.avg('Avg_Bonus_Malus'), 2).alias('Avg_Bonus_Malus'),
    F.round(F.avg('Avg_Rejets_CO2'), 2).alias('Avg_Rejets_CO2'),
    F.round(F.avg('Avg_Cout_energie'), 2).alias('Avg_Cout_energie')
)
```

```
global_avg_df.show()
```

```

+-----+-----+-----+
|Avg_Bonus_Malus|Avg_Rejets_CO2|Avg_Cout_energie|
+-----+-----+-----+
|      -3631.88|       43.81|       255.1|
+-----+-----+-----+

```

Il ne reste plus qu'à remplacer les valeurs null par les valeurs de ce dataframe.

```
final_df = joined_df.na.fill(global_avg_df.first().asDict())
```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|marque|nom|puissance|longueur|nbPlaces|nbPortes|couleur|occasion|prix|Avg_Bonus_Malus|Avg_Rejets_CO2|Avg_Cout_energie|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|volvo|S80 T6|272|très longue|5|5|blanc|false|50500|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|noir|false|50500|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|rouge|false|50500|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|gris|true|35350|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|bleu|true|35350|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|gris|false|50500|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|bleu|false|50500|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|rouge|true|35350|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|blanc|true|35350|-3631.88|42.45|72.73|
|volvo|S80 T6|272|très longue|5|5|noir|true|35350|-3631.88|42.45|72.73|
|volswagen|Touran 2.0 FSI|150|longue|7|5|rouge|false|27340|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|gris|true|19138|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|bleu|true|19138|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|gris|false|27340|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|bleu|false|27340|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|blanc|true|19138|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|noir|true|19138|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|rouge|true|19138|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|blanc|false|27340|-6000.0|13.33|149.33|
|volswagen|Touran 2.0 FSI|150|longue|7|5|noir|false|27340|-6000.0|13.33|149.33|
only showing top 20 rows

```

Si on reprend l'exemple de la marque SEAT

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|marque|nom|puissance|longueur|nbPlaces|nbPortes|couleur|occasion|prix|Avg_Bonus_Malus|Avg_Rejets_CO2|Avg_Cout_energie|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|seat|Toledo 1.6|102|longue|5|5|blanc|false|18880|-3631.88|43.88|264.45|
|seat|Toledo 1.6|102|longue|5|5|noir|false|18880|-3631.88|43.88|264.45|
|seat|Toledo 1.6|102|longue|5|5|rouge|false|18880|-3631.88|43.88|264.45|
|seat|Toledo 1.6|102|longue|5|5|gris|false|18880|-3631.88|43.88|264.45|
|seat|Toledo 1.6|102|longue|5|5|bleu|false|18880|-3631.88|43.88|264.45|

```

Enfin on extrait le dataframe résultat dans un fichier CSV afin de provisionner hive

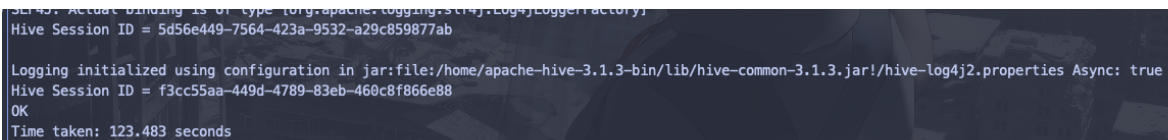
```
output_path = "/root/tpa/tpa/DATABASE/HDFS/final_df.csv"
```

```
final_df.write.csv(output_path, mode="overwrite", header=True)
```

## Hive

Afin d'ajouter la nouvelle table dans notre data lake hive voici les commandes à exécuter:

```
cd ~/tpa/tpa/DATABASE/HDFS/  
  
# cette opération peut prendre du temps (patience)  
hive -f co2_base.hql
```



```
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]  
Hive Session ID = 5d56e449-7564-423a-9532-a29c859877ab  
  
Logging initialized using configuration in jar:file:/home/apache-hive-3.1.3-bin/lib/hive-common-3.1.3.jar!/hive-log4j2.properties Async: true  
Hive Session ID = f3cc55aa-449d-4789-83eb-460c8f866e88  
OK  
Time taken: 123.483 seconds
```

Pour vérifier si tout a bien fonctionné

```
hive  
# cette opération peut prendre du temps (patience)  
hive> select * from catalogue;
```



volvo	S80 T6	272	très longue	5	5	blanc	false	50500	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	noir	false	50500	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	rouge	false	50500	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	gris	true	35350	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	bleu	true	35350	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	gris	false	50500	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	bleu	false	50500	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	rouge	true	35350	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	blanc	true	35350	-3631.88	42.45	72.73
volvo	S80 T6	272	très longue	5	5	noir	true	35350	-3631.88	42.45	72.73
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	gris	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	bleu	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	rouge	true	19138	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	blanc	false	27340	-6000.0	13.33	149.33
volkswagen	Touran 2.0 FSI	150	longue	7	5	noir	false	27340	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	blanc	true	8540	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	blanc	false	12200	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	noir	false	12200	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	noir	true	8540	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	bleu	true	8540	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	bleu	false	12200	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	rouge	true	8540	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	rouge	false	12200	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	gris	false	12200	-6000.0	13.33	149.33
volkswagen	Polo 1.2 6V	55	courte	5	3	gris	true	8540	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	blanc	true	18641	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	noir	true	18641	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	rouge	true	18641	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	gris	true	18641	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	bleu	true	18641	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	rouge	false	26630	-6000.0	13.33	149.33
volkswagen	New Beatle 1.8	110	moyenne	5	5	gris	false	26630	-6000.0	13.33	149.33