FITEC - Data analyst 3
Solène DRUAIS

Word Count use cluster Spark with Docker and Portainer

1 - Création du répertoire projet, initialisation de git et création du fichier texte « lePetitPrince.txt »

```
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark$ cd Spark_Shell_MapReduce
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ git init
Dépôt Git vide initialisé dans /home/fitec/Spark/Spark_Shell_MapReduce/.git/
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ code
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ ls
lePetitPrince.txt
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ git add .
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ git commit -m "initial commit with text file : lePetitPrince.txt"
[master (commit racine) 3499f59] initial commit with text file : lePetitPrince.txt
 1 file changed, 24 insertions(+)
 create mode 100644 lePetitPrince.txt
```

home > fitec > Spark > Spark_Shell_MapReduce > lePetitPrince.txt

```
1  Un aviateur, le narrateur du conte, se bloque avec son avion au milieu du désert du Sahara à la
2  suite d'une panne de moteur. Alors qu'il tente de réparer son avion, un petit garçon apparaît
3  et lui demande de dessiner un mouton : « S'il vous plaît… dessine-moi un mouton ! ».
4  Jour après jour, le narrateur découvre l'histoire du Petit Prince. Il lui raconte qu'il vient
5  d'une autre planète : "l'astéroïde B 612», une planète très petite à peine plus grande qu'une
6  maison où il a laissé derrière lui trois volcans et une rose, une fleur unique dont il est
7  amoureux. Le petit prince confie à l'aviateur avoir peur que le mouton qu'il lui a dessiné
8  fasse du mal à sa rose.
9  Le petit prince lui raconte aussi qu'il a visité d'autres planètes avant d'arriver sur la
10 Terre. D'une planète à une autre, il a rencontré des gens bizarres: un roi qui prétend régner
11 sur tout avec le pouvoir absolu, un vaniteux qui se voit comme l'homme le plus beau et le plus
12 intelligent alors qu'il est seul sur sa minuscule planète, un homme d'affaires propriétaire
13 d'étoiles qui passe son temps à les compter, un ivrogne qui boit pour oublier qu'il boit,
14 l'allumeur de réverbères qui effectue un travail absurde et ininterrompu et un vieux monsieur
15 géographe qui écrit, dans des livres énormes les informations portées à lui par les explorateurs.
16 Sur la Terre, le Petit Prince a rencontré un renard, il lui a appris qu'il est important de se
17 faire des amis qu'on doit les apprivoiser et les considérer comme des êtres uniques.
18 Chaque jour l'aviateur apprend de nouvelles choses sur le petit prince, sur ses sentiments, ses
19 peurs, ses doutes, son départ, son voyage et sur sa planète.
20 Huit jours après l'atterrissage dans le désert, l'heure de la séparation des deux amis est
21 venue. Afin de retourner sur sa planète, le petit prince a recours au serpent qui résout toutes
22 les énigmes. Le petit prince repart vers sa planète en laissant le narrateur tout seul. Enfin,
23 l'aviateur réussi à réparer son avion et quitte lui aussi le désert en espérant revoir le petit
24 prince un jour.
```

2 - Copie du fichier source vers le volume de data partagé, chargement du fichier vers spark-master puis vérification de l'existence du fichier sur ce cluster

```
(base) fitec@fitec-HP-ProBook-450-G5:~/Spark/Spark_Shell_MapReduce$ sudo docker cp lePetitPrince.txt spark-master:/root
bash-5.0# ls
bin   data  dev   etc   execute-st
bash-5.0# cd root
bash-5.0# ls
lePetitPrince.txt
bash-5.0#
```

3 - Lancement du shell sur Portainer (Scala)

```
bash-5.0# cd root
bash-5.0# ls
lePetitPrince.txt
bash-5.0# /spark/bin/spark-shell --master spark://spark-master:7077
21/12/29 17:38:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your pla
tform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel
).
21/12/29 17:38:26 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting po
rt 4041.
21/12/29 17:38:26 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting po
rt 4042.
Spark context Web UI available at http://2ae26c095809:4042
Spark context available as 'sc' (master = spark://spark-master:7077, app id = app-20211229
173826-0002).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_275)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

4 - Création du RDD avec le fichier .txt et exécution de Map Reduce.
Vérification des opérations et sauvegarde du fichier Résultat vers output (« resLePetitPrince.txt »)
Utilisation de la commande counts.take(10) pour vérifier le contenu

```
scala> val textFile = sc.textFile("/data/lePetitPrince.txt")
textFile: org.apache.spark.rdd.RDD[String] = /data/lePetitPrince.txt MapPartitionsRDD[10] at textFile at <console>:24

scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_+_)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[13] at reduceByKey at <console>:25
```

```
scala> counts.toDebugString
res0: String =
(2) ShuffledRDD[4] at reduceByKey at <console>:25 []
 +-(2) MapPartitionsRDD[3] at map at <console>:25 []
    |  MapPartitionsRDD[2] at flatMap at <console>:25 []
    |  /data/lePetitPrince.txt MapPartitionsRDD[1] at textFile at <console>:24 []
    |  /data/lePetitPrince.txt HadoopRDD[0] at textFile at <console>:24 []

scala> counts.cache()
res1: counts.type = ShuffledRDD[4] at reduceByKey at <console>:25

scala> counts.saveAsTextFile("/data/output/reslePetitPrince.txt")

scala> counts.take(10)
res3: Array[(String, Int)] = Array((planète,4), (pouvoir,1), (êtres,1), (repart,1), (volcans,1), (apprend,1), (vers,1), (B,1), (les,6), (d'affaires,1))

scala> for (Mot <- counts.take(10)) println(Mot)
(planète,4)
(pouvoir,1)
(êtres,1)
(repart,1)
(volcans,1)
(apprend,1)
(vers,1)
(B,1)
(les,6)
(d'affaires,1)
```

5 - Déplacement des données du volumes vers le dossier du projet
Dépôt du travail sur GitHub