```
(base) fitec@fitec-HP-ProBook-450-G5:~/Téléchargements/TOS/TOS_DI-20211109_1610-V8.0.1$ cd ~ (base) fitec@fitec-HP-ProBook-450-G5:~$ cd Documents/ (base) fitec@fitec-HP-ProBook-450-G5:~/Documents$ mkdir Spark_Shell_MapReduce (base) fitec@fitec-HP-ProBook-450-G5:~/Documents$ cd Spark_Shell_MapReduce/ (base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$ git init Dépôt Git vide initialisé dans /home/fitec/Documents/Spark_Shell_MapReduce/.git/ (base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$
```

1- Création du répertoire projet, initialisation de git

```
(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$ code
(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$ ls
input.txt
(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$ git add .
(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$ git commit -m 'initial commit with text f
[master (commit racine) e438093] initial commit with text file
1 file changed, 26 insertions(+)
create mode 100644 input.txt
(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark_Shell_MapReduce$
```

- 2- Création fichier texte à passer dans MapReduce
- 3- On copie/charge un exemplaire du fichier .txt vers le volume de notre cluster.

J'ai copié/collé à la main vers le volume monté à l'adresse:

var/lib/docker/volumes/sparkcluster\_shared\_data/\_data

J'aurais pu me servir des lignes de commande docker, comme utilisé plus tard pour rapatrier les données du volume docker.

selon le modèle:

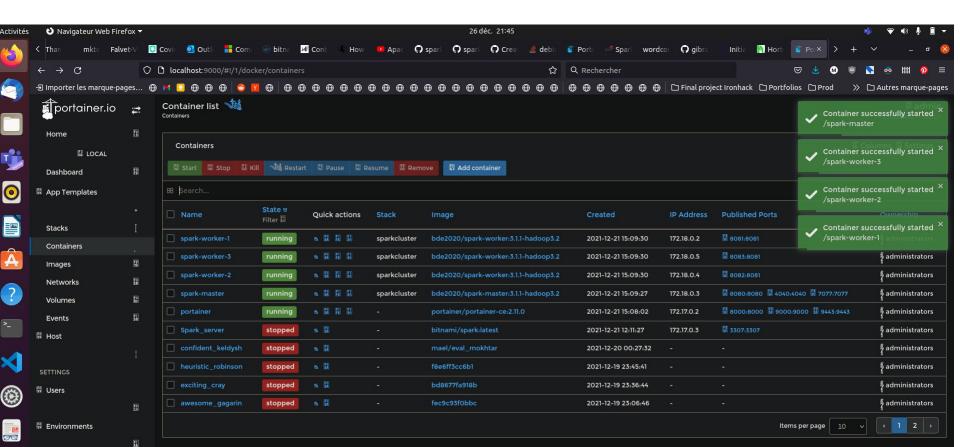
docker cp dummy:/path/to/file /dest/to/file

```
home > fitec > Documents > Spark Shell MapReduce > ≡ input.txt
      Barney Mayerson se réveilla la tête comme prise
      dans un étau, pour découvrir autour de lui une
      chambre inconnue, dans un immeuble de conapts qui
      ne lui disait absolument rien. À côté de lui, les cou-
      vertures remontées jusqu'à ses épaules nues, dormait
      une fille qu'il ne connaissait pas, la bouche entrou-
      verte et la tête auréolée d'une cascade de cheveux
      d'un blanc cotonneux.
      Je sens que je vais être en retard au boulot . Après
      s'être glissé hors du lit, il tangua un peu pour se redres-
      ser sur ses jambes, les yeux fermés, le cœur au bord
     des lèvres. Pour ce qu'il en savait, il pouvait fort bien
      être à plusieurs heures de route de son bureau ; peut-
 14 être même ne se trouvait-il plus aux États-Unis. Au
 15 moins n'avait-il pas quitté la Terre ; la pesanteur qui
 16 le faisait tituber était normale, familière.
      Et dans la pièce voisine, juste à côté du sofa, se
 18 trouvait la valise – tout aussi familière – de son psy-
      chiatre, le Docteur Smile.
      Pieds nus, il marcha à pas feutrés jusqu'au séjour
 21 et s'assit près de la valise. Après l'avoir ouverte, il
```

manipula les commandes censées mettre en route le Docteur Smile. Des compteurs commencèrent à s'ani-

mer, le mécanisme à ronronner. « Où suis-je ? lui demanda Barney. Et à quelle distance de New York ? » C'était le principal. Il remarqua alors l'horloge s

## 4- lancement de la stack spark standalone clsuter



```
data execute-step.sh lib
                            media
                                      proc sbin sys var
dev finish-step.sh lib64 mnt
                                       root spark tmp wait-for-step.sh
bash-5.0# cd data
bash-5.0# 1s
input.txt
bash-5.0# /spark/bin/spark-shell --master spark://spark-master:7077
21/12/27 18:59:29 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
Spark context Web UI available at http://86c4e8c0f243:4040
Spark context available as 'sc' (master = spark://spark-master:7077, app id = app-20211227
185945-0000).
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.
```

run

srv

usr

home master.sh opt

/data/input.txt

bash-5.0# 1s

etc

5- Lancement du shell portainer, localisation du chemin vers le .txt : il se situe à

bin

7- Lancement de spark sous Scala

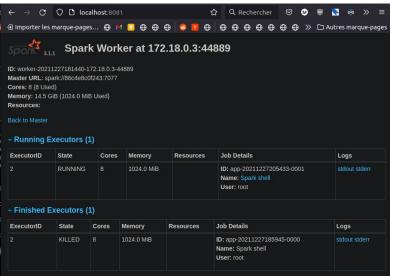
```
scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:25
  scala> counts.toDebugString
  res1: String =
  (2) ShuffledRDD[4] at reduceByKey at <console>:25 []
   +-(2) MapPartitionsRDD[3] at map at <console>:25 []
        MapPartitionsRDD[2] at flatMap at <console>:25 []
        /data/input.txt MapPartitionsRDD[1] at textFile at <console>:24 []
         /data/input.txt HadoopRDD[0] at textFile at <console>:24 []
  scala> counts.cache()
  res2: counts.type = ShuffledRDD[4] at reduceByKey at <console>:25
  scala> counts.saveAsTextFile("/data/output/input_count.txt")
  scala> counts.take(10)
  res4: Array[(String, Int)] = Array((blanc,1), (disait,1), (quelle,1), (les,3), (principal.,1), (auréolée,1), (hors,1),
  (entrou-,1), (n'avait-i1,1), (qu'i1,2))
          8- Création du RDD avec input.txt et exécution du MapReduce
```

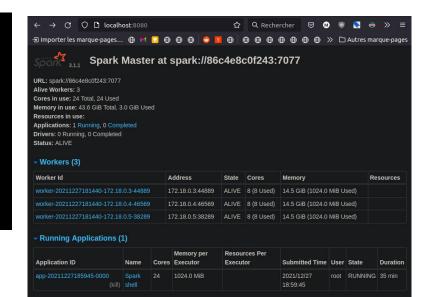
9- Vérifications des opérations à appliquer sur le RDD via toDebugString et sauvegarde

de l'output vers un dossier output (une erreur à l'écran, j'ai créé un dossier dans un dossier qui ne sert a rien). + vérification du contenu de counts avec méthode .take(10)

textFile: org.apache.spark.rdd.RDD[String] = /data/input.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> val textFile = sc.textFile("/data/input.txt")





Application ID Name Cores Memory per Executor Resources Per Executor Submitted Time User State Duration

Completed Applications (0)

(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark\_Shell\_MapReduce\$ pwd

/home/fitec/Documents/Spark Shell MapReduce

(base) fitec@fitec-HP-ProBook-450-G5:~/Documents/Spark\_Shell\_MapReduce\$ sudo docker cp 86c4e8c0f243:/data/ /home/fitec/Documents/Spark\_Shell MapRedu

10- Rapatrier les données du volume sparkcluster vers le répertoire de mon projet pour git commit et push vers github. (le code en screenshot rapatrie toutes les données du volume, je l'ai re exécuté pour ne prendre que l'output)
11- C'est fini j'écris un email à votre attention :)