

```
(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 file
[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
1 Barney Mayerson se réveilla la tête comme prise
2 dans un étau, pour découvrir autour de lui une
3 chambre inconnue, dans un immeuble de conapts qui
4 ne lui disait absolument rien. À côté de lui, les cou-
5 vertures remontées jusqu'à ses épaules nues, dormait
6 une fille qu'il ne connaissait pas, la bouche entrou-
7 verte et la tête auréolée d'une cascade de cheveux
8 d'un blanc cotonneux.
9 Je sens que je vais être en retard au boulot . Après
10 s'être glissé hors du lit, il tangua un peu pour se redres-
11 ser sur ses jambes, les yeux fermés, le cœur au bord
12 des lèvres. Pour ce qu'il en savait, il pouvait fort bien
13 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.
17 Et dans la pièce voisine, juste à côté du sofa, se
18 trouvait la valise – tout aussi familière – de son psy-
19 chiatre, le Docteur Smile.
20 Pieds nus, il marcha à pas feutrés jusqu'au séjour
21 et s'assit près de la valise. Après l'avoir ouverte, il
22 manipula les commandes censées mettre en route le
23 Docteur Smile. Des compteurs commencèrent à s'ani-
24 mer, le mécanisme à ronronner. « Où suis-je ? lui
25 demanda Barney. Et à quelle distance de New York ? »
26 C'était le principal. Il remarqua alors l'horloge s
```

4- lancement de la stack spark standalone cluster

Activités

Navigateur Web Firefox

26 déc. 21:45

localhost:9000/#/1/docker/containers

Rechercher

Importer les marque-pages...

portainer.io

Home

LOCAL

Dashboard

App Templates

Stacks

Containers

Images

Networks

Volumes

Events

Host

SETTINGS

Users

Environments

Container list

Containers

Start Stop Kill Restart Pause Resume Remove Add container

Search...

<input type="checkbox"/>	Name	State	Quick actions	Stack	Image	Created	IP Address	Published Ports	
<input type="checkbox"/>	spark-worker-1	running	a f p r	sparkcluster	bde2020/spark-worker.3.1.1-hadoop3.2	2021-12-21 15:09:30	172.18.0.2	8081:8081	
<input type="checkbox"/>	spark-worker-3	running	a f p r	sparkcluster	bde2020/spark-worker.3.1.1-hadoop3.2	2021-12-21 15:09:30	172.18.0.5	8083:8081	
<input type="checkbox"/>	spark-worker-2	running	a f p r	sparkcluster	bde2020/spark-worker.3.1.1-hadoop3.2	2021-12-21 15:09:30	172.18.0.4	8082:8081	
<input type="checkbox"/>	spark-master	running	a f p r	sparkcluster	bde2020/spark-master.3.1.1-hadoop3.2	2021-12-21 15:09:27	172.18.0.3	8080:8080 4040:4040 7077:7077	administrators
<input type="checkbox"/>	portainer	running	a f p r	-	portainer/portainer-ce:2.11.0	2021-12-21 15:08:02	172.17.0.2	8000:8000 9000:9000 9443:9443	administrators
<input type="checkbox"/>	Spark_server	stopped	a f p r	-	bitnami/spark:latest	2021-12-21 12:11:27	172.17.0.3	3307:3307	administrators
<input type="checkbox"/>	confident_keldysh	stopped	a f p r	-	mael/eval_mokhtar	2021-12-20 00:27:32	-	-	administrators
<input type="checkbox"/>	heuristic_robinson	stopped	a f p r	-	f8e6ff3cc6b1	2021-12-19 23:45:41	-	-	administrators
<input type="checkbox"/>	exciting_cray	stopped	a f p r	-	bd8677fa918b	2021-12-19 23:36:44	-	-	administrators
<input type="checkbox"/>	awesome_gagarin	stopped	a f p r	-	fec9c93f0bbc	2021-12-19 23:06:46	-	-	administrators

Items per page 10

1 2

Container successfully started /spark-master

Container successfully started /spark-worker-3

Container successfully started /spark-worker-2

Container successfully started /spark-worker-1

[illegible]

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

7- Lancement de spark sous Scala


```
scala> val textFile = sc.textFile("/data/input.txt")
textFile: org.apache.spark.rdd.RDD[String] = /data/input.txt MapPartitionsRDD[1] 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[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-il,1), (qu'il,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 à rien). + vérification du contenu de counts avec méthode .take(10)

localhost:8081

Importer les marque-pages... Autres marque-pages

Spark Worker at 172.18.0.3:44889

ID: worker-20211227181440-172.18.0.3-44889
Master URL: spark://86c4e8c0f243:7077
Cores: 8 (8 Used)
Memory: 14.5 GiB (1024.0 MiB Used)
Resources:

[Back to Master](#)

Running Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
2	RUNNING	8	1024.0 MiB		ID: app-20211227205433-0001 Name: Spark shell User: root	stdout stderr

Finished Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
2	KILLED	8	1024.0 MiB		ID: app-20211227185945-0000 Name: Spark shell User: root	stdout stderr

```
(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_MapReduce
```

- 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 :)