Dr. Heni Bouhamed
Data Scientist, Maître Assistant à la Fseg-Université de Sfax
Formation Data Science
Module 7 : Apache Spark

TP4 Apache Spark : installation avec Yarn de Hadoop et première application de streaming

---Installation avec Yarn---

1) Ajouter au bachrc les deux lignes suivantes :
   export HADOOP_CONF_DIR=$HADOOP_INSTALL/etc/hadoop
   export YARN_CONF_DIR=$HADOOP_INSTALL/etc/Hadoop
2) Mettre à jour le bashrc avec source .bashrc.
3) Accéder au dossier conf de spark avec cd $SPARK_HOME/conf.
4) Copier le fichier spark-defaults.conf.template pour avoir un autre fichier avec le nom spark-defaults.conf (avec cp spark-defaults.conf.template spark-defaults.conf).
5) Modifier le fichier spark-defaults.conf en activant spark.master avec yarn :
   Spark.master yarn (commande gedit spark-defaults.conf).
6) Dans le dossier conf de spark, copier le fichier spark-env.sh.template pour avoir un autre fichier avec le nom spark-env.sh (avec cp spark-env.sh.template spark-env.sh).
7) Modifier le fichier spark-env.sh en ajoutant :
   export HADOOP_CONF_DIR=$HADOOP_INSTALL/etc/hadoop
   export YARN_CONF_DIR=$HADOOP_INSTALL/etc/Hadoop
8) Formatter le namenode avec hdfs namenode –format.
9) Démarrer les services de Hadoop avec start-all.sh.
10) Sortir du mode safemode avec hdfs dfsadmin -safemode leave
    Remarque : le dossier par défaut : hdfs://localhost :9000/user/<nom utilisateur>
    Pour lire à partir d'un fichier local : file:///home/....
11) En mode yarn, on ne peut que lancer des scripts au niveau terminal avec la commande « spark-submit ».
12) Importer le dossier test.txt dans hdfs avec : hdfs dfs -put test.txt /user/…
13) Créer un script python ayatn le nom sparksubmityarn.py contenant les lignes de codes suivantes :

```
#!/usr/bin/env python
from pyspark import SparkContext
sc=SparkContext("local[*]",appName="app")
data = sc.textFile("test.txt")
print(data.collect())
from numpy import array
parsedData = data.map(lambda line:array([float(x) for x in line.split(' ')]))
print(parsedData.collect())
```

14) Sur le terminal exécuter : spark-submit sparksubmityarn.py

---Exemple de Streaming avec Netcat---

1) Créer un script python ayant le nom sparks.py dans le dossier home contenant les lignes de codes suivantes :

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

# Create a local StreamingContext with 2 working thread and batch interval of 1 second
sc = SparkContext("local[2]", "NetworkWordCount")
ssc = StreamingContext(sc, 20)
# Create a DStream that will connect to hostname:port, like localhost:9999
lines = ssc.socketTextStream("localhost", 9999)
# Split each line into words
words = lines.flatMap(lambda line: line.split(" "))
# Count each word in each batch
pairs = words.map(lambda word: (word, 1))
wordCounts = pairs.reduceByKey(lambda x, y: x + y)

# Print the first ten elements of each RDD generated in this DStream to the console
wordCounts.pprint()
ssc.start() # Start the computation
ssc.awaitTermination() # Wait for the computation to terminate
```

2) Lancer netcat dans un terminal avec nc –lk 9999.
3) Passer à un autre terminal et lancer le sparks.py avec spark-submit sparks.py localhost 9999.
4) Ecrire sur le terminal de netcat et observer le résultat sur le terminal de lancement de sparks.py.