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

TP1 Apache Spark : installation standalone

--Machine virtuelle Ubuntu--

1) Télécharger Apache Spark à partir du lien suivant : https://spark.apache.org/downloads.html.
2) Installer java avec sudo apt-get install default-jdk.
3) Créer un dossier spark dans /usr/local avec : mkdir /usr/local/spark (en cas de problème de permission vous pouvez utiliser : sudo chmod -R 777 /usr/local).
4) Accéder à Téléchargements et extraire le dossier d'installation de spark avec : tar -zxvf spark-2.3.1-bin-hadoop2.6.tgz.
5) Déplacer le dossier de Téléchargements à /usr/local/spark avec : mv spark-2.3.1-bin-hadoop2.6 /usr/local/spark.
6) Ouvrir le bashrc avec : sudo gedit .bashrc et ajouter les lignes suivantes :
   #spark
   export SPARK_HOME=/usr/local/spark/spark-2.3.1-bin-hadoop2.6
   export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
   #end spark
7) Lancer spark en mode python avec : pyspark.
8) Utiliser exit() pour quitter et revenir au shell.

--Installation Windows--

1) Télécharger spark sur windows à partir du lien : https://spark.apache.org/downloads.html.
2) Télécharger et installer java jdk 8 à partir du http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html.
3) Vérifier l'installation en accédant à cmd (en mode administrateur) en tapant java -version.
4) Télécharger et installer anaconda avec python 3.6 https://www.anaconda.com/download/.
5) Déplacer le dossier d'installation dans un dossier (le nommer spark par exemple) sur le lecteur C: ou de préférence D:.
6) Puisque nous utilisons une version spark pré-construite pour fonctionner avec hadoop, on a besoin d'un autre fichier pour la faire fonctionner. Pour ce faire, créer un nouveau dossier nommé WinUtils sous n'importe quel lecteur (D: par exemple), télécharger le dossier compressé de l'adresse suivante : https://github.com/steveloughran/winutils, extraire et copier tout le contenu du dossier bin (relatif à la version hadoop pré-construite installer avec spark) dans le dossier WinUtils créer précédemment.
7) Créer les variables d'environnements utilisateurs suivantes :
   HADOOP_HOME D:\WinUtils
   JAVA_HOME C:\Program Files\Java\jdk1.8.0_181
   SPARK_HOME D:\Spark\spark-2.3.1-bin-hadoop2.6
8) Modifier la variable d'environnement système path en ajoutant : D:\Spark\spark-2.3.1-bin-hadoop2.6\bin