Année universitaire 2024-2025
TP1 - M2

Introduction à SPARK – prise en main

1. Installation de Spark (en mode local, *télécharger*, *décompresser*, *utiliser*)

⇒ Allez sur le site : https://spark.apache.org/downloads.html
⇒ **Téléchargez** la dernière version, (compatible avec la version Hadoop 3.X), dans un répertoire de votre choix. Par exemple, ous y aurez donc le fichier : spark-3.1.0-bin-hadoop2.7.tgz
⇒ **Décompressez** le fichier dans votre répertoire :
- $ tar xvzf spark-3.1.0-bin-hadoop2.7.tgz

⇒ Un répertoire spark-3.1.0-bin-hadoop2.7a été créé : placez-vous dans ce répertoire. $
- cd spark-3.1.0-bin-hadoop2.7

⇒ Lancer le terminal de Spark:
- $ ./bin/spark-shell

⇒ L'interface Spark est accessible à l'adresse : http://localhost:4040/ (inutile de lancer le master en mode local)

2. Interaction avec Spark

Le but de cette section est de nous familiariser avec l'invite de commande de Spark (spark-shell).

Nous allons dans un premier temps lancer l'invite de commende si ce n'est pas déjà fait.

Exercice 1 : Les actions RDD (Scala):

1 Créer un RDD nommé P_Counts à partir du fichier d'entrée

```
scala> val p_counts = sc.textFile("/pagecounts")
```

2 Enregistrer le RDD créé dans un fichier
   a. Analyser le résultat.
   b. Combien de blocs avez-vous obtenus ?
   c. Visualiser via l'interface web l'exécution du job

3 Donner le nombre d'items (44)
   Tester les différentes fonctions.

4 Donner le premier item
5 Donner les 10 premiers items. Que constatez-vous ?
6 Donner les les 10 premiers items sous forme d'une liste.

Exercice 2 : Les transformations RDD (Scala):

1 Créer un RDD nommé « Filter_word » avec les items contenant le mot 'word'. Le RDD sera réutilisable par la suite, qu'est ce qu'il est nécessaire de faire ?
2 Retourner les 10 premiers items contenant le mot 'word'
3 Retourner le nombre de caractères de chaque ligne
4 Retourner le nombre de caractères du fichier
5 Retourner la liste des mots en utilisant les deux méthodes *Map* et *FlatMap*.
   a. Quelle est la différence entre ces deux fonctions ?
   b. Le résultat retourné doit être stocké dans la mémoire. Quelle fonction utiliser ?
6 Retourner le nombre de mot pour les 5 premiers items.
7 Retourner le nombre d'occurrences de chaque mot contenu dans le RDD. Le résultat doit être trié par ordre alphabétique et sauvegardé dans un fichier.

Exercice 3 : Les DataFrame (avec Spark SQL via Spark-shell)

1 Créer un *Dataframe* à partir du fichier *Personne.*
2 Afficher le Dataframe nouvellement créé.
   a. Que constatez-vous ?
   b. Quelle est la différence avec un RDD.
3 Afficher toutes les personnes âgées de plus de 20 ans.
4 Afficher l'enregistrement contenant le nom 'Justin'.