**TP : Lancer le ‘’wordcount’’ suivant les 2 modes d’exécutions de Hadoop**

Le Tp consiste à exécuter un job Map reduce en python (en JAVA, si vous le souhaiter) dans 2 environnements différents :

* **Mode local (standalone)** : dans ce mode, tout s’exécute au sein d’une seule JVM, en local. C'est le mode recommandé en phase de développement.
* **Mode local pseudo-distribué (pseudo-distribué)** : dans ce mode, le fonctionnement en mode cluster est simulé par le lancement des tâches dans différentes JVM exécutées localement.

I) Map/Reduce (sans Hadoop)

1) Télécharger le répertoire **TP 1 Hadoop** ainsi que les script mapper et reducer

2)Télécharger les fichiers txt (et les copier dans hadoop/wc/input) qui serviront de corpus pour les job Word Count MapReduce

3) S'assurer de donner les droits d'exécutions sur les fichiers mapper.py et reducer.py depuis répertoire de hadoop

4) Assurer vous que le mapper donne le résultat escompté en exécutant en ligne de commande depuis le dossier hadoop la commande suivante :  **cat input/5000.txt | ./mapper0.py**

Que signifie le symbole ‘’|’’

5) Assurer vous que le mapper donne le résultat escompté en exécutant depuis le dossier hadoop la commande suivante :  **cat input/5000.txt | ./mapper.py**

Que signifie le symbole du pipe ‘’|’’

6) Exécuter maintenant : **cat wc/input/5000.txt | ./mapper.py | sort**   
Qu’observez-vous ?

7) Exécutez maintenant la commande complète  
**cat wc/input/5000.txt | ./mapper.py | sort | ./reducer0.py**

Et vérifiez que le programme produit le résultat attendu. Si ça fonctionne,  
félicitations! Vous venez d’exécuter votre premier programme Map/Reduce en  
Utilisant propre implémentation de Map/Reduce (en ligne de commande)

II) Ex´ecution de WordCount en Map/Reduce avec Hadoop

Maintenant nous allons exécuter le Job sur dossier de fichier input contenant **5000.txt**  et **20417-8.txt** avec une exécution en streaming

Lien utile :

* https://phoenixnap.com/kb/install-hadoop-ubuntu

1. Assurez-vous d’avoir le dossier hadoop (Le path étant signifié dans les commandes exécuter plus haut) et que hadoop marche en exécutant la commande **hadoop –version**

A/ La chaine d'exécutions du job MapReduce

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \

-input /user/hadoop/wc/input \

-output /user/hadoop/wc/output \

-file /home/hadoop/mapper.py \

-mapper /home/hadoop/mapper.py \

-file /home/hadoop/reducer.py \

-reducer /home/hadoop/reducer.py

Que signifie chaque ligne de code ?