

TP 1: Commandes HDFS

Le but de ce TP est de manipuler les principales commandes shell de HDFS et de tester les transferts des fichiers entre un client et les machines du cluster Hadoop.

Attention ! il ne faut pas faire des copy/paste des commandes mais tapez directement les commandes car vous risquez d'introduire des caractères erronés.

Découvrir les différentes commandes HDFS et options avec :

```
hadoop fs -help
```

On crée d'abord un dossier input sur HDFS:

```
hadoop fs -mkdir -p input
```

On copie ensuite un fichier existant, exemple file1.txt dans le dossier input sur HDFS :

```
hadoop fs -put /Data/file1.txt input
```

Vérifier que le fichier file1.txt est bien dans input

```
hadoop fs -ls input
```

On obtient :

```
Found 1 items
```

```
-rw-r--r-- 2 root supergroup 338 2020-04-10 08:44 input/file1.txt
```

Le fichier a été stocké dans le datanode sur HDFS.

Vous pouvez vous connecter sur le datanode pour vérifier qu'un bloc a été créé.

Sur datanode, il suffit de consulter le contenu en navigant sous /usr/local/hadoop/data/datanode/current.

Dans mon cas, c'est dans /usr/local/hadoop/data/datanode/current/BP-1344118684-172.19.0.4-1664787033810/current/finalized/subdir0/subdir0 que le bloc a été créé.

Il suffit de visualiser le contenu avec ls -l ou lire le block avec cat blk_1073741857

Pour superviser l'état du datanode, se connecter sur le site : <http://localhost:9870/> et remarquez qu'il y a bien votre fichier avec la bonne taille crée sur HDFS.

Voici quelques commandes à tester avec HDFS :

```
hadoop fs -put /data/file.txt input
```

```
hadoop fs -get file.txt
```

```
hadoop fs -tail file.txt
```

```
hadoop fs -cat file.txt  
hadoop fs -mv file.txt newfile.txt  
hadoop fs -rm newfile.txt  
hadoop fs -mkdir -p myinput
```

Maintenant que vous êtes familiarisés avec les commandes de base, on peut utiliser un grand fichier volumineux que vous avez sur vos machines.

On va générer un fichier de 0 de 1M appelé zeros.dat, il suffit de taper la commande sur la machine namenode.

```
dd if=/dev/zero of=zeros.dat bs=1M count=150
```

Le mettre sur HDFS et refaire les manipulations testées avec file.txt.

Vérifier que le fichier est sur HDFS :

```
hadoop fs -ls input
```

Taper la commande suivante (qui prendra un peu de temps cette fois-ci vue la taille du fichier)

```
hadoop fs -put /Data/zeros.dat input
```

Sur combien de blocks est stocké le fichier zeros.dat ?

Utiliser la commande suivante :

```
hadoop fsck /path/file.txt -files -blocks
```

Remarques :

Au lieu d'utiliser le fichier 'fake' zero.dat, vous pouvez télécharger un des fichiers CSV du site suivant pour tester un fichier Big data: <https://www.stats.govt.nz/large-datasets/csv-files-for-download/>