Année universitaire : 2021-2022
Resp. module : M.-L. Messai

TD 2- TCSD

1 Objectifs

- Mettre en place l'environnement Hadoop sous la distribution Cloudera en VM.
- Prise en main de HDFS dans d'un cluster simple nœud.
- Apprendre les commandes HDFS afin de bien gérer les fichiers sur HDFS.

2 Activité 1 : Mettre en place la VM

Ce premier TP consiste le premier pas vers l'utilisation du *framework* Hadoop. Nous allons utiliser une machine virtuelle basé sur la distribution CentOS de Linux avec l'outil Cloudera configuré. Cloudera est l'une des distribution Hadoop les plus utilisée.

- Télécharger Cloudera QuickStarts VM à partir du lien suivant :

https://downloads.cloudera.com/demo_vm/vmware/cloudera-quickstart-vm-5.13.0-0-vmware.zip

- Décompresser le fichier *cloudera-quickstart-vm-5.13.0-0-vmware.zip*.
- Importer dans VMware la machine virtuelle de Cloudera que vous avez décompresser au préalable.
- Attribuerez 8 Giga de RAM et 4 processeurs à la VM et lancez-la en mode NAT.

3 Activité 2 : Commandes HDFS

- Ouvrez un terminal et tapez la commande suivante pour avoir un clavier en azerty :

```
setxkbmap fr
```

- Assurez-vous que vous êtres dans le répertoire /home/cloudera puis lancez Cloudera Manager par la commande :

```
sudo ./cloudera-manager –express –force
```

- Si le lancement se passe bien, le message : « Success ! » s'affiche avec les informations de connexion sur le manager (url, username, password).
- `hadoop fs` : cette commande affiche la liste des commandes supportées par HDFS (Vous pouvez utiliser la commande `hdfs dfs`, les deux commandes sont équivalente).
- Pour connaitre la version de hadoop, la commande est :

```
hadoop version (ou hdfs version)
```

- Toutes les commande ont le format : `hdfs dfs -COMMANDE` (`hadoop fs -COMMANDE`).
- Les noms de commandes et leurs fonctionnalités ressemblent à celles du shell Uinx.
- Pour afficher de l'aide d'une commande donnée : `hadoop fs -help COMMANDE`

3.1 Importer et exporter des données

- `hdfs dfs -ls` : liste l'ensemble des fichiers du répertoire utilisateur HDFS. Elle n'affiche rien pour l'instant car le dossier HDFS est vide.
- `hdfs dfs -ls /` : affiche ce qu'il y a à la racine HDFS. Vous pouvez lister le contenu des répertoires racine, par exemple `hdfs dfs -ls /user`.
- Créez localement un fichier texte *monfichier.txt*, modifiez son contenu, sauvegardez et quittez.

- Copiez ce fichier sur HDFS par hdfs dfs -put monfichier.txt. Utilisez hdfs dfs -ls -R pour vérifier.
- Une autre commande permet aussi d'envoyer une copie de vos données sur HDFS est :

hdfs dfs -copyFromLocal monfichier.txt

- Si vous voulez envoyer vos données vers HDFS sans garder une copie en local :

hdfs dfs - moveFromLocal monfichier.txt

- Affichez son contenu sur HDFS : hdfs dfs -cat monfichier.txt
- Pour les fichiers longs, vous pouvez faire hdfs dfs -cat monfichier.txt | less ou hdfs dfs -cat bonjour.txt | more
- Pour supprimer un fichier depuis le système de fichiers HDFS :

hdfs dfs -rm monfichier.txt

- hdfs dfs -mkdir CHEMIN1 CHEMIN2 … : cette commande crée les répertoires du chemin 1 puis chemin2, … etc.
- Créez localement un dossier nommé data et envoyez-le sur HDFS.
- Copiez le fichier monfichier.txt dans le répertoire data à l'aide de la commande -cp (vérifiez).
- Créez un dossier datasets dans le dossier data, puis déplacez monfichier.txt dans datasets à l'aide de la commande -mv (vérifiez).
- Créer une copie de monfichier.txt dans le répertoire data sous le nom copiedemonfichier.txt.
- Pour transférer un fichier de HDFS vers le local, la commande est :

hdfs dfs -get *monfichier.txt*

hdfs dfs -copyToLocal *monfichier.txt*

- Avant de lancer cette commande, il faut vérifier que l'espace local disponible est suffisant pour recevoir les données HDFS.
- Si vous voulez envoyer des données de HDFS vers le système local sans garder une copie en HDFS :

hdfs dfs - moveToLocal *mondossier*

- Pour supprimer un répertoire depuis le système de fichiers HDFS :

hdfs dfs -rm -r *mondossier*

- Essayez la commande -chmod sur un fichier / dossier pour modifier ses droits. Vérifiez.
- Une commande qui vous permet de voir « l'état de santé » de votre HDFS (elle vérifie les incohérences : blocks manquants, nom de réplicas insufusants,…) :

hdfs fsck /user/cloudera

- La liste de toutes les commandes est sur la page :

https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

4 Activité 3

- A partir de la VM, téléchargez les données disponibles sur le site :

http://grouplens.org/datasets/movielens/1m/

- Décompressez le fichier zip.
- Le fichier *rating.dat* contient plus d'un million d'évaluations anonymes d'environ 3 900 films réalisé par 6 040 utilisateurs de MovieLens (voir le README pour plus de détails).
- Déroulez les étapes de création de deux dossier /datasets/movies et copiez le fichier *rating.dat* à partir du système local vers HDFS (dans movies).
- Pour voir la décomposition d'un fichier en plusieurs blocs, récupérez le fichier zip gutenberg-200M.txt.gz. Décompressez-le puis copiez-le dans HDFS dans un dossier books dans le dossier précédemment créée datasets.
- Affichez combien de blocs occupe le fichier avec la commande :

```
hdfs fsck /user/cloudera/books -files -blocks
```

5 Activité 4 : Fichiers de configuration Hadoop

Tous les fichiers de configuration d'Hadoop sont disponibles dans le répertoire /etc/hadoop/conf.

Le fichier /etc/hadoop/conf/hdfs-site.xml contient les paramètres spécifiques au système de fichiers HDFS.

- Consultez le contenu de ce fichier. Quelle est la valeur du paramètre dfs.replication. Ce dernier permet de préciser le nombre de réplication d'un block sur les nœuds d'un cluster. Justifiez !
- Vous pouvez afficher la valeur de réplication directement par la commande :

```
hdfs getconf -confkey dfs.replication
```

- La taille du bloc : HDFS stocke les fichiers dans le cluster en les décomposant en blocs de taille fixe. Quelle est la taille du bloc sur votre HDFS ?

```
hdfs getconf -confkey dfs.blocksize
```

- Vous pouvez changer la taille du bloc pour un fichier par la commande :

```
hdfs dfs -D dfs.blocksize=67108864 -put Monfichier
```

- Créer un fichier text_64.txt puis envoyez-le sur HDFS en fixant la décomposition en blocs de 64 Mo. Vérifiez.