Travaux pratiques Big Data

Hadoop

Présentation des travaux

Durant cette séance, vous allez manipuler des données volumineuses dans un environnement Hadoop. Vous mettrez notamment en œuvre un montage réseau afin d'accéder aux données depuis votre console.

Toutes les opérations s'effectuent depuis le serveur qui est mis à votre disposition depuis le répertoire /usr/local/hadoop. Pour rappel, cd /usr/local/hadoop permet de se positionner dans ce répertoire.

Rappel de l'arborescence :

bin => Exécutables utilisateurs
etc => Fichiers de configurations
include => Fichiers d'entêtes pour la programmation
lib => Les librairies non vitales
libexec => exécutables internes à hadoop
logs => Journaux d'évènements
sbin => Exécutables système
share => Fichiers de données

1. Manipulation des données

Hadoop dispose de commandes interne afin de manipuler les données, créer un répertoire, modifier les autorisations, déplacer des données, créer des copies instantanées et les restaurer.

Vous trouverez la liste complète à l'adresse suivante : https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

2. Travail demandé

Pour chaque question, vous devez réaliser l'opération demandée et noter la commande que vous avez utilisée dans le document.

Au préalable, vous devez formater le système de fichier sur le nœud primaire et démarrer les services :

$ hadoop namenode -format

$ start-dfs.sh

2.1. Téléchargement des fichiers d'activités

Depuis votre répertoire personnel, télécharger les fichiers suivants sur http://bigdata.celeonet.fr, archive_ip.csv, domaines.csv, ip.csv, asnum.csv, liaison.csv, centos7.iso

```
$      wget http://bigdata.celeonet.fr/ip.csv
       wget  http://bigdata.celeonet.fr/domaines.csv
      wget  http://bigdata.celeonet.fr/asnum.csv
      wget  http://bigdata.celeonet.fr/liaison.csv
      wget  http://bigdata.celeonet.fr/centos7.csv
      wget  http://bigdata.celeonet.fr/centos7.iso
```

2.2. Création de son répertoire personnel (hduser)

Créez votre répertoire personnel dans hadoop.

```
$
   hadoop fs  -mkdir /user/hduser/jerometresor
```

2.3. Afficher l'arborescence

Vérifiez que votre répertoire est bien créé en listant l'arborescence

```
$  hadoop fs -ls  /user/hduser/jerometresor
```

2.4. Copier des fichiers locaux vers hadoop

Copiez dans votre répertoire personnel hadoop les fichiers précédemment téléchargés

```
$       hadoop fs -put *.csv /user/hduser/jerometresor
        hadoop fs -put *.iso /user/hduser/jerometresor
```

2.5. Copier des fichiers hadoop vers local

Copiez le fichier domaines.csv vers le fichier domaines-2020.csv

```
$
    hadoop fs  -get /user/hduser/jerometresor/domaines.csv  domaines-202.csv
```

2.6. Renommer un fichier dans hadoop

```
$
   Hadoop  fs  -mv  /user/hduser/jerometresor/ip.csv    /user/hduser/jerometresor/ip-2020.csv
```

2.7. Espace disque utilisé

Indiquez combien d'espace est actuellement occupé sur votre système de fichier hadoop. Afficher la valeur en « human ».

```
hadoop fs  -du  -h   /user/hduser/jerometresor
```

2.8. Supprimer un fichier

Supprimez le fichier centos7.iso

$

hadoop fs -rm /user/hduser/jerometresor/centos7.iso

2.9. Récupérer un fichier

Vous avez par erreur supprimé le fichier centos7.iso, restauré le.

$

hadoop fs -mv /user/hduser/.Trash/200213100000/user/hduser/jerometresor/centos7.iso /user/hduser/jerometresor/centos7.iso

2.10. Activer les instantanés

Activez les instantanés sur le répertoire /user

$ hdfs dfsadmin -allowSnapshot /user/hduser/jerometresor

2.11. Créer un instantané

Créez un instantané sur /user et nommer le à la date du jour (JJMMYYYY).

$

hdfs dfs -createSnapshot /user/hduser/jerometresor/snapshot_20200213

2.12. Lister les répertoires instantanés

Listez les répertoires pouvant faire l'objet d'un instantané.

```
$  hdfs lsSnapshottableDir
```

2.13. Lister les fichiers d'un instantané

Listez les fichiers de votre instantané JJMMYYYY.

NB : Les instantanés sont présents dans /user/.snapshot

```
$ hdfs dfs -ls /user/hduser/jerometresor/.snapshot/snapshot_20200213
```

2.14. Lire le contenu d'un fichier depuis un instantané

Vous souhaitez visualiser le contenu de ip-2020.csv sans le restaurer, affichez à présent le contenu (ctrl + c pour quitter).

```
$  hdfs dfs -cat /user/hduser/jerometresor/.snapshot/snapshot_20200213/ip-2020.csv
```

2.15. Lister les fichiers au travers du système en réseau (nfs)

Au préalable, exécuter les commandes suivantes :

```
$ sudo /usr/local/hadoop/bin/hdfs --daemon start portmap
$ sudo /usr/local/hadoop/bin/hdfs --daemon start nfs3
$ mkdir nfs
$ sudo mount -t nfs -o vers=3,proto=tcp,nolock,noacl,sync localhost:/ nfs
```

```
$ cd nfs -ls
```

2.16. Supprimer les fichiers du réseau en réseau

Supprimez l'ensemble des fichiers dans le répertoire user et afficher son contenu

$ rm*

Que constatez-vous ?

Les fichiers sont supprimés sur le réseau et dans hadoop .

2.17. Restaurer un instantané

Restaurez l'instantané créé XXMMJJJJ

$ hadoop fs -get /user/hduser/jerometresor/.snapshot/snapshot_20200213/* /user/hduser/jerometresor

Que constatez-vous dans le répertoire ~/nfs/user ?

Les fichiers sont restaurés

2.18. Supprimer un instantané

Supprimez l'instantané créé XXMMJJJJ

$ hdfs dfs -deleteSnapshot /user/hduser/jerometresor snapshot_20200213

2.19. Arrêter hadoop

Arrêtez le service hadoop

```
$ stop-dfs.sh
```