Plateforme Hidoop

Projet de données réparties

2020 - 2021

LAPLAGNE Chloé
RAZAFIMANANTSOA Nathan
GUILLAUD Thomas
GRUNIG Axel

Ce document présente le fonctionnement global de la plateforme.

Le répertoire *hidoop* correspond à l'arborescence de fichiers suivante :

- le répertoire **config** contient les fichiers d'initialisation pouvant être utiles lors du lancement de la plateforme
- le répertoire **data** accueille les fichiers de données de l'application
- le répertoire **doc** accueille les rapports attendus
- le répertoire **src** contient les codes sources. Ce répertoire contient lui-même les sous-répertoires suivants :
- application, pour le code des applications
- config, pour les utilitaires de configuration
- formats, pour la spécification et la réalisation des formats
- hdfs, pour la mise en œuvre de hdfs
- ordo pour l'ordonnancement et le contrôle des tâches Map/Reduce

Configuration de Hidoop

La variable système **HIDOOP_HOME** est nécessaire au fonctionnement de Hidoop. Il s'agit de la localisation du répertoire *hidoop*, utilisé dans config.Projet.PATH .

La configuration se fait via un fichier *conf.xml* placé dans le répertoire **HIDOOP_HOME/config/**. Un fichier d'exemple est donné ci-dessous :

```
<?xml version="1.0" encoding="UTF-8"?>
<config metadata="meta">
    <default-chunk-size value="64" unit="bytes" />
    <servers>
        <node ip="127.0.0.1"/>
        <node ip="chewie"/>
        <node ip="yoda"/>
    </servers>
</config>
```

- ***config***: l'attribut obligatoire *metadata* est le nom du fichier de métadonnées situé dans **HIDOOP_HOME/data**.
    - ***default-chunk-size***: élément optionnel précisant la taille de chunk à utiliser par défaut dans HDFS. Si absent, cette taille est de 64MB.

      Note : les unités de tailles supportées sont bytes, kB, MB, GB (non sensible à la casse). Par simplicité, une unité inconnue a le même effet que bytes.

    - ***servers*** *:* liste des serveurs. L'attribut *ip* d'un *node* correspond en réalité soit à l'adresse ip de la machine soit à son nom (dns). Cette liste ne peut pas être vide.

Ce fichier est utilisé par la classe ***config.AppData*** permettant de charger la configuration de l'application.

Compilation

```
$ export HIDOOP_HOME=/path/to/hidoop
$ cd $HIDOOP_HOME/src
$ javac application/MyMapReduce.java application/Count.java
ordo/WorkerImpl.java hdfs/HdfsClient.java hdfs/HdfsServer.java
```

Ou, pour spécifier un répertoire particulier pour les classes :

```
$ export HIDOOP_HOME=/path/to/hidoop
$ export HIDOOP_CLASSES=/path/to/class/files
$ cd $HIDOOP_HOME/src
$ javac -d $HIDOOP_CLASSES  application/MyMapReduce.java
application/Count.java ordo/WorkerImpl.java hdfs/HdfsClient.java
hdfs/HdfsServer.java
```

Script de lancement

Le script ***hidoop.sh*** ouvre un shell permettant de lancer / arrêter automatiquement les serveurs indiqués dans le fichier de configuration *conf.xml* via ssh (machines N7). Le répertoire où les commandes sont exécutées est ***HIDOOP_HOME***/*src* par défaut, mais il est possible de définir une variable système **HIDOOP_CLASSES** qui indique la localisation des classes java de l'application.

Il s'agit d'un shell bash supportant les commandes spécifiques suivante :

- ***start*** pour lancer les serveurs
- ***stop*** pour arrêter les serveurs
- ***hdfs*** raccourci pour java hdfs.HdfsClient, suivi des mêmes arguments
- ***mmr*** raccourci pour java application.MyMapReduce, suivi des mêmes arguments

Les sorties de chaque serveur sont redirigées vers les fichiers *<ip_serveur>.log* dans dans **HIDOOP_HOME.**

Il est conseillé d'utiliser la plateforme via ce script.

Utilisation

La plateforme Hidoop supporte les opérations suivantes :

- Écriture d'un fichier dans HDFS :

  ```
  hidoop> hdfs -w <nom_fichier_local> options
  ```

  Le fichier écrit dans HDFS aura le nom du fichier local.
  Options :
  - `-f ln|kv` : format du fichier (ln par défaut)
  - `--chunks-size=<taille>` : taille des chunks en octets (si non strictement positif, la taille par défaut est utilisée)
  - `--rep=<facteur>` : facteur de réplication (entier positif), non supporté et toujours égal à 1 dans cette version

- Lecture d'un fichier dans HDFS

  ```
  hidoop> hdfs -r <nom_fichier> options
  ```

  Options :
  - ➢ `<fichierLocal>` : fichier local de destination. Lecture du fichier dans '`r_<nom_fichier>`' si ce paramètre n'est pas spécifié.

- Liste des fichiers dans HDFS

  ```
  hidoop> hdfs -l
  ```

- Suppression d'un fichier de HDFS

  ```
  hidoop> hdfs -d <nom_fichier>
  ```

- Exécution d'une application en Map/Reduce

  ```
  hidoop> mmr <nom_fichier>
  ```

  Le fichier créé contenant les résultats finaux est `<nom_fichier>-tot`