A propos

Pour ce projet de traitement de données pour les systèmes distribués, nous avons décidé de ne pas faire de scripts d'installation (pour la partie distribuée). Nous avons opté pour une explication générique sur l'installation et la configuration des différents outils (Zookeeper, Spark). De cette manière, une équipe qui souhaite mettre en place l'application, pourra l'intégrer sur n'importe quel environnement UNIX tout en ayant la main sur les configurations.

Pour l'exemple d'installation et de configuration nous utiliserons 3 machines. Pour la mise à l'échelle, il suffira d'augmenter le nombre de machines. Nous prendrons les machines suivantes pour illustrer l'intégration :

Machine 1 :

host : HOST1

port : PORT1

Machine 2 :

host : HOST2

port : PORT2

Machine 3 :

host : HOST3

port : PORT3

Nous supposons que sur chacune de ces machines il y aura en distribué : un Zookeeper, un Spark Master, un Spark Slave. Nous allons voir ci-dessous comment installer et configurer ces outils.

Installation de Zookeeper

Zookeeper est l'outil qui nous permet de gérer le Spark Master actif grâce à un algorithme de consensus. Pour chaque Spark Master que nous voulons créer pour augmenter la robustesse et la disponibilité, il faut un Zookeeper. Dans notre cas, nous allons créer 3 Zookeepers.

Lien de téléchargement : https://zookeeper.apache.org/releases.html

Dans un premier temps, il faut installer Zookeeper sur chaque machine. Une fois l'installation faite, il faut créer un fichier vide myid sur la racine de chaque Zookeeper. Puis dans ce fichier il faut mettre un numéro correspondant à une identification de la machine. Par exemple :

- myid du HOST1 : 1
- myid du HOST2 : 2
- myid du HOST3 : 3

Pour que les Zookeeper puissent se connecter entre eux il faut créer un cluster grâce à un fichier de configuration. Dans chaque Zookeeper, il faut créer le fichier de configuration zoo.cfg contenant les lignes suivantes (en veillant à remplacer *FOLDER_ZOOKEEPER_MYID* par le dossier où myid est situé et les *HOST* correspondants):

clientPort=2181
dataDir=FOLDER_ZOOKEEPER_MYID
syncLimit=2
initLimit=5
tickTime=2000
server.1=HOST1:2888:3888
server.2=HOST2:2888:3888
server.3=HOST2:2888:3888

Pour lancer Zookeeper : $./zkStart.sh start

Pour arrêter Zookeeper : $./zkStart.sh stop

Il ne manque plus qu'à connecter les sparks master à ces Zookeeper.

Installation de Spark

Spark est l'outil qui nous permet de gérer le flux des tâches pour effectuer des calculs distribués sur des données de notre base de données distribuée MongoDB.

Lien de téléchargement : http://spark.apache.org/downloads.html

Dans un premier temps, il faut installer Spark sur chaque machine. Une fois l'installation faite, il faut créer un fichier vide ha.conf dans le dossier conf de chaque Spark. Puis dans ce fichier il faut spécifier l'adresse du Zookeeper que l'on veut connecter au spark master de la machine. Par exemple dans le spark de la machine HOST1 il faut mettre dans le fichier :

spark.deploy.recoveryMode=ZOOKEEPER
spark.deploy.zookeeper.url=HOST1:2181
spark.deploy.zookeeper.dir=/TMP_FOLDER

Pour lancer un spark master : $./sbin/start-master.sh --webui-port 8080 --properties-file conf/ha.conf

Pour arrêter un spark master : $./sbin/stop-master.sh --webui-port 8080 --properties-file conf/ha.conf

Pour lancer un spark slave : $./sbin/start-slave.sh spark://HOST1:7077,HOST2:7077,HOST3:7077

Pour arrêter un spark slave : $./sbin/stop-slave.sh spark://HOST1:7077,HOST2:7077,HOST3:7077