Machine virtuelle Hadoop pour les TP de BigData

Pierre Nerzic

Abstract

Explications pour installer Hadoop sur Qemu ou VirtualBox en mode pseudo-distribué.

IMPORTANT : vous ne pourrez faire ces manipulations que si votre machine contient au moins 8 Go de RAM et 20 Go d'espace disque disponible. Une machine ayant moins de RAM plantera sûrement.

Préparations

Téléchargement de l'archive

Vous devez télécharger une archive, `hadoop-amd64.zip`. Son URL vous sera donné en cours. Elle fait environ 2Go. Décompressez-la. Elle contient ce fichier d'explications, deux scripts de lancement, ainsi qu'un fichier `hadoop-amd64.iso`. Ce dernier est un DVD d'installation en mode *preseed*, c'est à dire préconfiguré pour une installation automatique. Elle est comme les images ISO qu'on télécharge pour installer un système sur un PC. Elle contient tous les logiciels nécessaires pour obtenir une machine Hadoop complète. Cette image pourrait être installée sur un vrai PC, mais il y a plusieurs options qui sont trop restreintes. Surtout, **ne pas l'installer sur votre PC réel** car ça écraserait totalement votre système et bien suivre les étapes suivantes !

Installation des logiciels sur votre PC

En premier, vous devez installer Qemu, également appelé KVM. QEmu est le logiciel utilisé par AndroidStudio pour faire fonctionner les tablettes virtuelles AVD. Il est plus simple d'emploi que VirtualBox, par contre, il est entièrement en ligne de commande. Sur Windows, il y a un bug avec QEmu qui le rend extrêmement lent et donc il faut utiliser VirtualBox.

- Sur Debian ou Ubuntu, il suffit de taper `sudo apt-get install qemu-kvm libvirt-bin qemu-utils`. Une fois les paquets installés, vous devez faire `adduser $LOGNAME kvm` puis `adduser $LOGNAME libvirt`. Consulter https://debian-facile.org/doc:systeme:kvm pour des explications détaillées.
- Sur Windows, vous devrez utiliser VirtualBox de Oracle. Allez sur la page VirtualBox, et cliquez sur le gros bouton de téléchargement. Choisissez Windows Hosts dans la liste (109Mo) et installez-le. Ensuite, il suffit de configurer une machine pour système Linux Debian 64 bits, ayant entre 4 et 8 Go de RAM et 10 Go ou plus de disque dur virtuel (type VDI, dynamiquement alloué). Il faut également monter l'image `hadoop-amd64.iso` sur le lecteur de DVD virtuel (onglet `Stockage` du panneau de configuration) et démarrer dessus au premier lancement.
- Sur Mac, installez-le par `brew install qemu`. Je ne peux pas vous aider davantage, ça sera à vous de consulter les documentations. Ou alors faites avec VirtualBox.

Sur toutes les machines, vous devrez vérifier que le BIOS active les options de virtualisation : `Intel VT-x enabled`.

Configuration de la machine virtuelle

Selon votre PC, il peut être nécessaire de modifier la mémoire et la taille du disque virtuel à utiliser. La mémoire RAM minimale est 4 Go et le disque minimal est 8 Go. Avec aussi peu, on arrive tout juste à travailler. Vous ne pourrez pas télécharger la totalité des données prévues pour faire les TP. Mettez un peu plus si vous pouvez, mais gardez au moins 2 Go de RAM pour votre système hôte, sinon tout sera bloqué. Par défaut, il y a 6 Go de RAM et 10 Go de disque.

- Sur Linux, éditez le script `launch_hadoop.sh`. Changez les lignes `RAM=6G` et `HDISK=10G`, par exemple mettez `RAM=8G` et `HDISK=12G`.
- Sur Windows, modifiez les préférences de VirtualBox.

Premier lancement

- Sur Linux, c'est très simple : lancer la commande `sudo ./launch_hadoop.sh install-single`
- Sur Windows, démarrez la machine virtuelle avec VirtualBox. Elle doit booter sur le DVD virtuel associé à l'image ISO.
- Ensuite, dans tous les cas, la machine virtuelle étant démarrée et montrant le menu d'installation :

  1. Choisir l'item : « Pseudo-Distributed Cluster » en tapant entrée. Ce mode configure Hadoop en mode simple machine : tous les services tournent sur une seule machine.
  2. Choisir votre clavier (français par défaut), tapez entrée.
  3. Laisser le système s'installer tout seul en automatique, ça dure 5 à 15 minutes selon la puissance de votre PC. Le système redémarre automatiquement deux fois après l'installation, il affiche `master login:` ou un écran noir, mais attendre encore avant de tenter une connexion, car il doit initialiser HDFS, Hive et HBase au premier démarrage. Il ne sera prêt que lorsque vous verrez la bannière de connexion X11.
  4. L'installation est finie quand vous voyez l'écran graphique de connexion (fond bleu/vert avec l'éléphant hadoop).

Travail avec le système

Démarrage

- Sur Linux, c'est très simple : lancer la commande `sudo ./launch_hadoop.sh single` (attention, le paramètre est différent du premier lancement, sinon vous repartez pour une installation).
- Sur Windows, c'est très simple : démarrez la machine virtuelle sans le DVD d'installation, voir l'onglet *Storage* du panneau de configuration.

Attendre de voir l'écran de connexion graphique.

Connexion

Le compte pour travailler est `uti`, le mot de passe est `=uti=`. C'est un système Debian 9 (stable, nom de code *stretch*), avec un environnement de bureau minimaliste appelé OpenBox.

Pour éteindre la machine, il faut toujours utiliser le menu « Éteindre » du fond d'écran, ou taper la commande `sudo shutdown -h now` dans un terminal. Surtout ne jamais fermer la fenêtre abruptement. Malheureusement, QEmu n'affiche aucune boîte de confirmation « Voulez-vous vraiment éteindre la machine ? »

Connexion shell par SSH

Sur Linux, on peut aussi se connecter à la machine virtuelle à partir d'un shell réel, par : `ssh -o UserKnownHostsFile=/dev/null -o StrictHostKeyChecking=no -p 20022 uti@localhost`. Ça permet d'éviter d'utiliser l'écran virtuel et de faire des copier-coller directement. Il est possible de copier des fichiers par `scp -p 20022 FICHIER uti@localhost:/home/uti/` ou dans l'autre sens : `scp -p 20022 uti@localhost:/home/uti/FICHIER ..`

Sur Windows, on doit pouvoir faire la même chose avec PuTTY, en se connectant sur le port 20022. Les extensions VirtualBox n'ont pas été installées, alors pas de copier-coller direct.

Il est possible de naviguer sur les sites de la machine virtuelle à partir du navigateur de la machine réelle. Il vous suffit d'ouvrir le lien http://localhost:20080 sur votre navigateur. Certains URL devront être modifiés à la main, par exemple `http://master.cluster.virt:8042/` à remplacer par `http://localhost:8042/` pour

accéder au NodeManager, ou alors il vous faudrait modifier votre fichier `/etc/hosts` pour y rajouter cette ligne : `127.0.0.1  master.cluster.virt`

C'est rendu possible parce que les ports de la machine virtuelle ont été redirigés : son port 22 est relié au port 20022, le port 80 au 20080, etc.

Applications disponibles

Les applications sont disponibles soit avec le menu de fond d'écran, soit avec la barre des tâches en bas. Seules les applications essentielles ont été installées : java8, eclipse, firefox. D'autres logiciels peuvent être installés avec `sudo apt-get install nom-du-paquet`. Surveillez l'occupation du disque virtuel avec `pydf`.

Firefox

Ouvrir le navigateur (bouton dans la barre en bas ou clic droit sur le fond d'écran). La page d'accueil fournit des liens vers différents composants de Hadoop. Vous pouvez aussi faire apparaître la barre personnelle pour avoir des raccourcis sur les composants. C'est dans le menu des marque-pages, à côté de l'étoile à droite de la zone où on saisit l'URL ; descendre dans Barre personnelle et cocher « Afficher la barre personnelle ». Pour faire une recherche sur internet, il suffit d'ouvrir un nouvel onglet, avec `CTRL-T` ou le bouton `+`, et taper la recherche dans la zone de texte qui apparaît.

Services

Au départ, seuls les services Hadoop sont actifs : HDFS, Yarn et HBase. La raison est que chaque service tente de consommer toute la mémoire disponible. Pour activer l'un des autres : Spark, Cassandra, ElasticSearch, il faut employer la commande `sudo SelectService` en passant le nom du service voulu en paramètre :

- `sudo SelectService spark` pour activer Spark, cela arrête tous les autres services sauf HDFS,
- `sudo SelectService cassandra` pour activer Cassandra, cela arrête tous les autres services,
- `sudo SelectService elasticsearch` pour activer ElasticSearch et Kibana, cela arrête tous les autres services,
- `sudo SelectService hadoop` pour activer HDFS, Yarn et HBase, et arrêter les autres services.

Certains services sont très longs à arrêter, HBase notamment, et d'autres très lents à activer, Cassandra par exemple ; il faut être patient. Les services choisis restent actifs jusqu'à un nouveau changement.

Programmation

Pour refaire les TP des cours IUT, il faut installer les données dans HDFS. Pour cela, dans un shell de la machine virtuelle, il faut lancer le script `~/InitHDFSShare.sh`. Par exemple, il télécharge les livres, les parcmètres de Paris et les relevés météo. Il y en a pour 3 Go environ et un très long moment. Si vous avez créé un disque virtuel de 8Go seulement, les horodateurs de Paris ne seront pas téléchargés.

Le workspace Eclipse s'appelle `/home/uti/Eclipse`. Eclipse est parfait pour éditer les sources. Par contre, il ne peut pas lancer les jobs MapReduce lui-même. Il faut ouvrir un terminal dans le dossier du projet considéré et taper `make`.