Hadoop Distributed File System

Data science
Master 2 ISiDIS
2017 / 2018

Le but est d'installer le framework hadoop sur différents ordinateurs en particulier sur un cluster de Raspberry pi et d'utiliser les commandes de base du système HDFS.

1 Installation de Hadoop

La meilleure documentation est celle originale fournie par hadoop :
`http://hadoop.apache.org` ou `http://hadoop.apache.org/docs/current/`.

N'hésitez pas à la consulter pendant l'installation. Par ailleurs, comme d'habitude, vous trouverez beaucoup d'informations complémentaires sur le web.

Avant de commencer à proprement parler l'installation, jetez un coup d'oeil sur le tutoriel de yahoo qui sera réaliser à la fin de ce tp :
`https://developer.yahoo.com/hadoop/tutorial/module2.html`

1.1 Sur machine locale

Afin d'installer la dernière version du framework hadoop (2.7 ou 3.0) suivez les instructions du site web de Bilel Derbel : `https://sites.google.com/site/bilelderbelpro/home/teaching/ppdg5khadoop`

Vous pouvez également suivre les instructions du tutoriel de Yahoo qui utilise une machine virtuelle :
`https://developer.yahoo.com/hadoop/tutorial/module3.html`

1.2 Sur un cluster de raspberry

Un tutoriel pour l'installation de hfs sur un cluster de raspberry pi est donné par Alan Verdugo :
`https://developer.ibm.com/recipes/tutorials/building-a-hadoop-cluster-with-raspberry-pi/`

1. Prendre des notes sur les manipulations que vous faites pendant l'installation.
2. Suivre les instructions données par le tutoriel signalé plus haut et consulter les tutoriels/aides citées plus bas.
3. Il faut d'abord installer l'OS sur la carte sd de chaque raspberry pi :
   `https://www.raspberrypi.org/`
4. Ensuite, vous pouvez mettre des noms de machine normalisée (rpXX) à l'aide du fichier `/etc/hostname` et affecter des ip statiques à l'aide du fichier `/etc/network/interfaces`.

5. Vous pouvez suivre globalement les instructions données sur le blog de Widriksson : `http://www.widriksson.com/raspberry-pi-hadoop-cluster/`.
   Seulement il faudra surement mettre à jour certaine partie pour la dernière version de hadoop. Par exemple, en suivant ce blog : `http://scn.sap.com/community/bi-platform/blog/2015/04/25/a-hadoop-data-lab-project-on-raspberry-pi--part-14`
6. Pour le passage au cluster de Rpy, un autre blog de Carsten Mönning donne des indications pour la version 2.6 :
   `http://scn.sap.com/community/bi-platform/blog/2015/07/10/a-hadoop-data-lab-project-on-ras`

2 Découvrir par la pratique

Maintenant que votre système HDFS est en place, vous pouvez faire "intelligemment" (en adaptant) le tutoriel proposé par Yahoo : `https://developer.yahoo.com/hadoop/tutorial/module2.html`

Vous pouvez réaliser ce tutoriel soit sur votre propre machine, ou encore mieux sur le cluster de raspberry ce qui permet de comprendre la configuration du système.