CES Data Scientist, Télécom ParisTech

Installation et configuration des clients Hadoop

Pierre Senellart (pierre.senellart@telecom-paristech.fr)

15 décembre 2014

Le but de ce TP est de configurer votre environnement pour accéder aux services de Hadoop (HDFS, HBase) à partir de Python, et de tester votre installation. Vous pouvez suivre les instructions soit sur les ordinateurs de la salle de TP, soit sur votre propre ordinateur, à condition qu'il soit connecté au réseau de Télécom ParisTech. Si votre ordinateur fonctionne sous Windows, un prérequis est l'installation de Maven `http://maven.apache.org/`. Sous Linux et sous Mac OS X, ce n'est pas nécessaire (mais une machine virtuelle Java est nécessaire dans tous les cas, ainsi qu'un interpréteur Python).

1 Hadoop côté serveur : Paramètres

Deux installations de Hadoop et de HBase vous sont accessibles ; la première est un vrai cluster Hadoop formé de plusieurs dizaines de serveurs – la serveur maître est `lame10` ; la seconde est une installation pseudo-distribuée sur une simple serveur – la serveur maître est `tiresias`. Essayez d'utiliser `lame10` ; en cas de soucis, passez sur `tiresias`. Ces deux serveurs ne sont joignables que depuis le réseau de Télécom ParisTech. Dans tout le TP, *`serveur`* désignera le nom de l'une ou l'autre de ces machines.

1. Accédez à l'interface Web HDFS de chacun de ces deux clusters à `http://`*`serveur`*`:50070/` et explorez ; comparez, en particulier, l'espace de stockage disponible et le nombre de nœuds actifs et inactifs.
2. Accédez à l'interface Web HBase de chacun de ces deux clusters à `http://`*`serveur`*`:60010/` et explorez ; comparez, en particulier, la liste des serveurs de régions.

Sur les deux serveurs, vous avez accès à une table HBase du nom de `simplewiki` contenant les articles de Simple English Wikipedia `http://simple.wikipedia.org/` (une version de Wikipedia écrite en anglais simplifié). Cette table a une ligne par article (l'identifiant ou clef de chaque ligne est le titre de l'article), une seule famille de colonnes (`wiki`) et plusieurs colonnes contenant les dernier auteur, estampille temporelle, texte, etc., de chaque article.

2 Hadoop côté client : Installation

Si vous êtes sous Windows, suivez les instructions sur `https://wiki.apache.org/hadoop/Hadoop2OnWindows` en parallèle de celles données ici.

1. Téléchargez l'archive complète de Hadoop disponible sur le site officiel du projet Apache Hadoop, dans sa dernière version (2.6.0).
2. Extrayez-la dans un répertoire `hadoop/` de votre dossier personnel (ou similaire).

3. Éditez le fichier `etc/hadoop/hadoop-env.sh` de ce répertoire pour y ajouter les lignes (par exemple à la fin) :

   ```
   export HADOOP_PREFIX=$HOME/hadoop/
   export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64/
   ```

   en adaptant à l'endroit dans lequel vous avez extrait l'archive Hadoop, et à l'endroit dans lequel est installé la machine virtuelle Java.

4. Ajoutez au fichier `etc/hadoop/core-site.xml`, à l'intérieur de la balise `<configuration>`, la propriété suivante :

   ```
   <property>
       <name>fs.defaultFS</name>
       <value>hdfs://serveur:8020</value>
   </property>
   ```

   (Ajustez le nom du serveur.)

5. Ajoutez `hadoop/bin` à votre variable d'environnement `PATH` ; par exemple, si vous utilisez le shell `zsh`, vous pouvez ajouter au `.zshrc` de votre répertoire personnel la ligne

   ```
   PATH=$HOME/hadoop/bin:$PATH
   ```

   en adaptant en fonction de l'endroit dans lequel vous avez extrait l'archive Hadoop.

6. Testez votre installation en démarrant un nouveau terminal et en tapant :

   ```
   hadoop fs -ls /
   ```

   Vous devriez voir le contenu du répertoire racine du système de fichiers HDFS.

3 Clients Python pour Hadoop

1. Si vous êtes administrateur de la machine sur laquelle vous êtes, passez cette étape. Sinon, nous allons créer un environnement virtuel Python dans lequel vous pourrez installer de nouveaux modules sans être administrateur :
   a) Téléchargez l'archive `https://pypi.python.org/packages/source/v/virtualenv/virtualenv-1.11.6.tar.gz`
   b) Extrayez-la dans un répertoire temporaire
   c) Placez-vous dans ce répertoire temporaire et créez un environnement Python virtuel avec la commande :

      ```
      python virtualenv.py ~/virtualenv
      ```

   d) Ajoutez dans votre script de login (par exemple `.zshrc` si votre shell de login est `zsh`) la ligne :

      ```
      source $HOME/virtualenv/bin/activate
      ```

      afin de positionner les variables d'environnement permettant d'utiliser votre environnement virtuel Python à la place de l'environnement système.
   e) Testez l'installation de votre environnement virtuel en ouvrant un nouveau terminal et en tapant :

      ```
      which python
      ```

      Si votre environnement virtuel est bien configuré, vous devriez voir une référence à la commande `python` de votre environnement virtuel, pas à celle du système.

2. Installez le module Python Hadoopy permettant d'accéder à Hadoop depuis Python avec la commande :

   ```
   pip install -e git+https://github.com/bwhite/hadoopy#egg=hadoopy
   ```

   (Si vous êtes administrateur, cette commande doit être exécutée en tant qu'administrateur.)

3. Installez le module Python Happybase permettant d'accéder à HBase depuis Python avec la commande :

   ```
   pip install happybase
   ```

   (Si vous êtes administrateur, cette commande doit être exécutée en tant qu'administrateur.)

4. En vous appuyant sur la documentation du module Hadoopy disponible à l'URL `http://www.hadoopy.com/en/latest/tutorial.html` (en particulier, de `hadoopy.ls()`), testez votre installation Hadoopy en écrivant un programme Python listant les entrées du répertoire racine du système de fichiers HDFS.

5. En vous appuyant sur la documentation du module HBase disponible à l'URL `http://happybase.readthedocs.org/en/latest/`, testez votre installation HBase en écrivant un programme Python qui affiche le contenu d'une page de Simple English Wikipedia (au format Wiki) dont le titre est donné en argument de ligne de commande. Demandez à l'enseignant si vous rencontrez des problèmes de droits lors de l'accès à la table.