CES Data Scientist, Télécom ParisTech

Index inversé en MapReduce

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

5 January 2015

Le but de ce TP est d'implémenter la construction d'un index inversé en MapReduce, donc de permettre de paralléliser et passer à l'échelle ce qui a été fait à la session précédente. Vous pourrez vous référez à la solution de référence disponible sur le site pédagogique.

Nous réutilisons les installations Hadoop côté client et serveur de la session précédente. Pour cette session, nous nous limiterons au cluster dont le maître est `tiresias` car certains logiciels ne sont pas installés sur l'autre cluster. Vous pouvez vous référer à l'énoncé de la session précédente pour l'ensemble des paramètres nécessaires.

Vous avez par ailleurs accès à l'URL : `http://tiresias.enst.fr:8088/` à une interface de visualisation des jobs MapReduce en cours.

1 Configuration préliminaire

Pour l'utilisation de la version streaming de Hadoop MapReduce, ajoutez dans le fichier `.zshrc` de votre répertoire personnel, à la fin, les lignes :

```
export HADOOP_HOME=$HOME/hadoop
export HADOOP_STREAMING=$HADOOP_PREFIX/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar
```

(en remplaçant `$HOME/hadoop` par l'emplacement de votre installation Hadoop).

Par ailleurs, veuillez supprimer les fichiers
`share/hadoop/tools/sources/hadoop-streaming-2.6.0-test-sources.jar` et
`share/hadoop/tools/sources/hadoop-streaming-2.6.0-sources.jar`
de votre installation Hadoop, qui sont sources de confusion pour l'API Hadoopy.

2 Tutoriel de Hadoopy

Lisez attentivement l'ensemble du tutoriel de l'interface Hadoopy `http://www.hadoopy.com/en/latest/tutorial.html`. Vous pourrez par la suite également utiliser la documentation de l'interface de programmation, disponible à `http://www.hadoopy.com/en/latest/api.html`.

3 De la table simplewiki à un fichier HDFS

Hadoopy gère encore mal la lecture directe depuis une table HBase (ça n'est pas un problème avec l'interface Hadoop Java). Commencez donc par écrire un script Python mettant l'intégralité de la table simplewiki (la colonne `wiki:text` uniquement) dans un fichier sur HDFS au format obtenu par l'appel `hadoopy.writedb`. Vous mettrez ce fichier sous HDFS à l'adresse suivante : `/user/login/simplewiki.tb` où login est votre nom de famille.

4 Construction de l'index inversé

Implémentez la construction de l'index inversé en MapReducem avec l'interface Python Hadoopy. Le fichier d'entrée est le fichier que vous venez de créer sur HDFS. La sortie doit aussi être placée sur HDFS, dans votre répertoire personnel. Pour faciliter le débogage :

— Commencez par utiliser `hadoopy.launch_local` plutôt que `hadoopy.launch` qui simule une exécution MapReduce au sein du même processus

— Limitez vous au 1000 premières entrées de la table simplewiki, en modifiant le script que vous avez créé à la question précédente à cet effet.

Une fois que votre programme fonctionne, enlevez ces deux restrictions et tester.

5 Chargement du résultat dans HBase

Chargez le fichier résultat dans HBase, pour produire une table au même format que celle obtenue au TP précédent. Fixez `batch_size=1000000` pour forcer des batchs de taille raisonnable.