NutchHadoopSingleNodeTutorial

Running Nutch in (pseudo) distributed-mode

(This tutorial is based on a linux operating system)

---

1. Step: Download and install Hadoop in pseudo-distributed mode, as explained here:

Hadoop Single Node Setup.

Here, it's important to set up *HADOOP_HOME* to point to the root of the hadoop installation, similar to *JAVA_HOME* it has to be set globally, so the hadoop start-up script can be called from anywhere.

(Check this by running: ' *echo $HADOOP_HOME* ' in the console, which should return the path to the root of your hadoop installation.)

N.B. Make sure your Hadoop installation is working correctly by running the examples as mentioned in the link above before trying to integrate Nutch!

E.g. try to connect to the jobtracker at: http://localhost:8088/.

---

2. Step: Download and install Nutch 1.x:

Download a stable source version e.g. apache-nutch-1.15-src.zip from http://nutch.apache.org/downloads.html.

For installation of apache-nutch-1.15-src.zip:

- Unzip and over the terminal cd into the freshly exracted folder *apache-nutch-1.15*
- Run 'ant runtime' in this folder

This command builds the runtime environment, where *runtime/local* stores all configuration files, libraries etc. but does not use the hadoop version, which has been set up here (pseudo-distributed mode), but the local (standalone) non-distributed version, that is often used for debugging and described in more detail here:
Hadoop Standalone Setup.

However, the nutch-job jar used for hadoop in pseudo-distributed mode lives in *runtime/deploy/*. As a consequence, any modification to the configuration files in *$NUTCH/conf* (the configuration directory at the root) require a re-build with 'ant' to make sure the changes become part of the nutch-job jar as well.

N.B. Make sure that the property mapreduce.framework.name in etc/hadoop/mapred-site.xml is set as mentioned in the hadoop documentation above.

See: NutchTutorial on how to set up a specific configuration and run a crawl.

Running Nutch on Hadoop, even on a single node, has distinct advantages, namely monitoring jobs in the jobtracker, containing detailed information about the individual jobs, such as counters, showing input/output, job configurations, which are invaluable especially in cases of failure.

Note: The Hadoop-related part of the setup is also relevant for setting up Nutch 2.x in pseudo-distributed mode.