Upgrading Hadoop

The purpose of this document is to show how to upgrade the version of hadoop within Nutch.

- Download the latest release version of Hadoop. It is preferred to download a release version instead of building from source because the release contains the Hadoop binaries and are build by Hadoop QA. If you download from source you will need to build the C libs. Also remember that the name used to build hadoop will appear in the hadoop admin screens. If you are upgrading Hadoop for the Nutch release, it is preferred to download the latest binary release for Hadoop.
- Unzip the release and copy the lib/native/* directories into your clean Nutch trunk workspace under trunk/lib/native where trunk is the root of the Nutch Trunk. You will also want to copy the hadoop-core jar from the root of the hadoop release into the trunk/lib directory.
- Remove the *.la files from the trunk/lib/native/OS directories (ex. trunk/lib/native/Linux-i386-32/libhadoop.la). These are just script files and are not needed for the release. You will also want to remove any older versions of the hadoop-core jar from the trunk/lib directory.
- If there are any errors or code that needs to be changed because of Hadoop API upgrades, that would need to happen here.
- Do a full clean and build of Nutch through the ant clean and package targets.
- Run the full test suite for Nutch using the ant test target.
- It is best to run a few full fetches and indexes using the new Hadoop versions. If this is not possible, see if you can build a drop and allow others to run some fetches. It is best to do this using Nutch in a distributed mode.
- Once all tests have passed and a few fetch cycles have been run, post a patch with the relevant changes. Then following the standard commit rules for wait time before commit, you can commit into the nutch repository. Make sure to change the trunk/CHANGES.txt file to relect the hadoop upgrade and any significant hadoop API changes that may have occurred.