Nutch2Tutorial

Nutch 2.X Tutorial

http://gora.apache.org/resources/img/gora-logo.png http://hbase.apache.org/images/hbase_logo.png

Introduction

This document describes how to get Nutch 2.X to use HBase as a storage backend for Gora. It is assumed that you have *a working* knowledge of configuring Nutch 1.X, as currently configuration in 2.X is more complex. It is important to take this in to consideration before progressing any further. We therefore **strongly advise** that you check out the Nutch 1.X tutorial.

Obtaining Software and Configuration

- Grab the latest distribution of Nutch 2.X from here. **Do NOT build the source yet**. From now on we will refer to the directory where the Nutch code resides as $NUTCH_HOME.
- Download and configure HBase 0.98.8-hadoop2. You can get it here (**N.B.** Each version of Gora is tied to a particular version of HBase, we therefore suggest you use this version if possible. If you decide to use another version of HBase please do not be surprised if the stack does not work. You should also obtain current documentation for HBase however please again take into consideration that the version of HBase we recommend you use may not correlate to the current documentation. Please keep this in mind and use your initiative.
- Specify the GORA backend in $NUTCH_HOME/conf/nutch-site.xml along with all of the other Configuration options suggested within the Nutch 1.x tutorial.

```
<property>
 <name>storage.data.store.class</name>
 <value>org.apache.gora.hbase.store.HBaseStore</value>
 <description>Default class for storing data</description>
</property>
```

- Ensure the HBase gora-hbase dependency is available in $NUTCH_HOME/ivy/ivy.xml

```
    <!-- Uncomment this to use HBase as Gora backend. -->

    <dependency org="org.apache.gora" name="gora-hbase" rev="0.6.1" conf="*->default" />
```

- In addition add the missing hbase-common-0.98.8-hadoop2.jar transitive dependency, this is a bug in gora-hbase 0.6.1 as described here. This bug is removed in current Gora development.

```
    <dependency org="org.apache.hbase" name="hbase-common" rev="0.98.8-hadoop2" conf="*->default" />
```

- Ensure that HBaseStore is set as the default datastore in $NUTCH_HOME/conf/gora.properties. Other documentation for HBaseStore can be found here.

```
    gora.datastore.default=org.apache.gora.hbase.store.HBaseStore
```

- **N.B.** It's probably worth checking and setting all your usual configuration settings within $NUTCH_HOME/conf/nutch-site.xml etc. before progressing.
- Compile Nutch -> via

```
ant runtime
```

- Make sure HBase is started and working properly as per the quick start tutorial.
- Create a list of URLs as you would do within the Nutch 1.X tutorial.

Invoke Nutch

You should then be able to inject URLs into HBase. Try going to *$NUTCH_HOME/runtime/local/bin* and do :

```
  nutch inject /someseedDir
  nutch readdb
```

Whats Next

You may want to check out the documentation for the Nutch Web Application and then the Nutch REST API as this gives a comprehensive overview of ongoing work with making Nutch 2.X easier to use.

Extra/Important Notes

N.B. The crawl command in the bin/nutch script is deprecated. You should use individual commands or alternatively use the bin/crawl script... which effectively chains together individual commands.

You should find more details in the logs on *$NUTCH_HOME/runtime/local/logs/hadoop.log*.

N.B. It's possible to encounter the following exception: java.lang.NoClassDefFoundError: org/apache/hadoop/hbase/HBaseConfiguration; this is caused by the fact that sometimes the hbase TEST jar is deployed in the lib dir. To resolve this just copy the lib over from your installed HBase dir into the build lib dir. (This issue is currently in progress).

N.B. The process of using the other datastore implementations offered within Gora e.g. Apache Cassandra, Accumulo, can be achieved simply by tweaking the above settings prior to compiling the Nutch code.

N.B. As of Apache Gora release 0.3, the gora-sql 0.1.1-incubating artifact is deprecated. The choice is to downgrade to Nutch 2.1 if you wish to use MySQL or HSQLDB as a Gora backend.

For more details of the command line interface options, please see here, or of course run ./bin/nutch which will print usage to std out. Finally, for a more detailed Nutch (1.X) tutorial, please see here

back to FrontPage