

HBase-Writer 0.19.1 Released

Contributed by Ryan Smith
Monday, 16 February 2009

HBase-Writer version 0.19.1 has been released and is available for download now. This version has been tested on a few Heritrix2 crawls on Hadoop 0.19.0 and HBase -0.19.0 and runs better now. This version fixes the previous new feature to work properly. In 0.19.0, if "only_new_records" is set to "true" and duplicate url records were in the hbase table, Heritrix would not download the content. Which is fine except, then you cant crawl any new records because you have to download the page to get all the links to follow. So this issue would better be solved by Heritrix itself by overriding extractor classes in Heritrix or taking snapshots during the crawl so you can pick up where you left off. So now in hbase-writer version 0.19.1, when "only_new_records" is set to "true", Heritrix will always download the content associated with the crawled urls, but its content will only be written to the given HBase table once. The next version of hbase-writer will have the option to not download the content if the record in hbase already exists (0.19.0 functionality).

Also important to note, Hadoop uses Java 1.6 now , and so HBase-Writer does as well. Happy crawling & enjoy!