

HBase-Writer 0.19.0 Released

Contributed by Ryan Smith
Wednesday, 11 February 2009

HBase-Writer version 0.19.0 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 well. I was able to add a new feature: "only-new-records". This boolean option is set to "false" by default and will crawl and write all urls & their content to the given hbase table (as expected). But by setting this to "true", you ensure that only new urls(rowkeys) are written. The way it works is normally when you crawl the same site more than once, you are adding multiple cells to the various crawl columns, (i.e.: "content:raw_data", "cui:url", etc..) but each cell will have a different timestamp associated with it. So, for example, if you crawl the same site 5 times in a row, you will get one rowkey for each url crawled, but 5 occurrences of each column, each with its unique timestamp; The only exception being columns updated by the crawler in a batchUpdate will have the same timestamp. This is so you know, all cells with the same timestamp came from the same fetch. So when the Hbase-Writer option "only-new-records" is set to "true" you will get no more than one occurrence of each column per rowkey. This is useful in cases where you want to crawl a site over a long period of time and plan on starting and stopping the crawler many times. This can also be useful if you want to crawl a site and only get new urls. Future versions will implement the feature of not downloading the content from the webserver in addition to not writing it to HBase; This can greatly reduce the load on the webserver you are crawling as only the header is fetched and needed to determine if the url is already existing.

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