

HBase-Writer 0.18.1 Released

Contributed by Ryan Smith
Tuesday, 28 October 2008

HBase-Writer 0.18.1 has been initially released. This is a processor plugin following the Heritrix2 processor API. With HBase-Writer, you can have Heritrix2 crawl and save its results directly to a table in HBase. The HBase-Writer plugin was based off the Heritrix-HDFS-Writer plugin. Thanks to Questio for the support in releasing this project.

HBase-Writer is designed to be extensible but as it is, it can be used as a powerful web crawling tool.

Out of the box, HBase-Writer is ready to write the crawled result of individual urls from Heritrix2, including http headers and the url content, into a given HBase table. The row key is the url itself, and the content and headers are stored in 2 separate column families.

As of today, Heritrix2 still needs a little bit of work to get the Web UI 100% operational. If you use it, you will probably find yourself having to password authenticate multiple times and getting a few stacktraces, NullPointerExceptions, etc. These issues seem to be limited to the web ui and the crawler itself works as advertised. These web ui issues should be resolved shortly as the Heritrix2 project stated they are going to take some time to do some more releases. So for now, to get around this, just edit the sheet files and seed file from the jobs directory under heritrix2: `vi heritrix/jobs/profile-xyz/sheets/global.sheet`

One of the key advantages to extending HBase-Writer yourself is the ability to parse and process the raw content on the fly, thus avoiding a MapReduce job or 2 to index your data in the format you would ultimately like for searching. Currently, the only way to extend HBase-Writer is to extend its 3 classes: HBaseWriter, HBaseWriterPool & HBaseWriterProcessor; but you only need to implement a method or 2 in each class. When you extend HBaseWriter, you will want to override the processContent(BatchUpdate bu) method to process the data and assign new data to new columns using the BatchUpdate object. In future releases of HBase, BatchUpdate will be deprecated for another design: RowUpdate.

HBase-Writer will stay up-to-date with the latest version of HBase and will match HBase's versioning for the time being. HBase-Writer uses Hudson for continuous integration builds, Nexus for its archive and repository manager and Maven for its build framework. All these resources can be viewed from the HBase-Writer project website.

To contribute or help in the development of HBase-Writer, please create an Issue on the project website and upload any patches for review.

- HBase-Writer -Heritrix2 Processor plugin for writing web crawl output to hbase tables.
- Heritrix-HDFS-Writer -Heritrix2 Processor plugin for writing web crawl output to the hdfs filesystem.
- Heritrix2 - The Internet Archiver's very own crawler.
- HBase - A distributed 'BigTable' storage engine.

- Hadoop - HBase runs on top of the Hadoop distributed filesystem.