HDFS Archive (HAR) Compactor Introduction

The HDFS Archive compactor is designed to be used in conjunction with the binary file writer. It is a service that runs periodically to combine files on an HDFS file system in to HAR (Hadoop Archive) file to prevent the system running out of blocks. The compactor can be configured to monitor one or more content sources and will the look at files in the lower level directories of the output produced by the binary file writer. If the number of files in these directories exceeds a given threshold, the files will be added to a HAR file and then deleted. Only one HAR file will exist per lower level directory and will be added to and updated as required*

Each content source directory will be monitored and compacted in its own thread to allow multiple directories to be compacted in parallel. As each directory is compacted, it will be “locked” using a lock file to prevent multiple processes attempting to compact the same directory.

The compactor may also be configured with HDFS resource files and have security enabled (Kerberos) if required

* The Hadoop standards define HAR files as immutable. Updates are made by opening a new HAR file, and transferring files across from the old as required. New content will be added and eventually the next HAR file will replace the old which is deleted

On this page

Related pages

- Prerequisites
- How to Configure
- FAQ & Troubleshooting

Features

Some of the features of the HDFS Archive compactor include:

- Scans directories created by the HDFS Binary Writer and compacts these files produced in to HAR files
- Scans one or more content source directories
- Supports updates and deletes, updating or removing content from HAR files (by opening up an existing HAR, and transferring unchanged content to a new HAR file)
- Supports Kerberos security
  - via a key tab file
- Allows addition of Hadoop or HDFS resource files to simplify configuration

Limitations

HDFS HAR Compactor has been tested again Cloudera v5.10.1

Anything we should add?

Please let us know.