HADOOP ARCHIVES

Overview

Hadoop archives are special format archives.

A Hadoop archive maps to a file system directory.

A Hadoop archive always has a *.har extension.

A Hadoop archive directory contains metadata (in the form of _index and _masterindex) and data (part-*) files.

The _index file contains the name of the files that are part of the archive and the location within the part files.

How to Create an Archive

Usage: hadoop archive -archiveName name -p <parent> <src>* <dest>

Archives Examples

Creating an Archive

hadoop archive -archiveName **foo.har** -p **/user/hadoop** dir1 dir2 **/user/zoo**

The above example is creating an archive using /user/hadoop as the relative archive directory. The directories /user/hadoop/dir1 and /user/hadoop/dir2 will be archived in the following file system directory -- /user/zoo/foo.har. Archiving does not delete the input files.

Looking Up Files

Looking up files in hadoop archives is as easy as doing an ls on the filesystem. After you have archived the directories /user/hadoop/dir1 and /user/hadoop/dir2 as in the example above, to see all the files in the archives you can just run:

hadoop dfs -lsr har:///user/zoo/foo.har/

Hadoop Archives and MapReduce

Using Hadoop Archives in MapReduce is as easy as specifying a different input filesystem than the default file system. If you have a hadoop archive stored in HDFS in /user/zoo/foo.har then for using this archive for MapReduce input, all you need to specify the input directory as har:///user/zoo/foo.har. Since Hadoop Archives is exposed as a file system MapReduce will be able to use all the logical input files in Hadoop Archives as input.