**SequenceFile in Hadoop**

The **Hadoop Distributed File System (HDFS)** is designed to store and process large (terabytes) data sets. However, storing a large number of small files in HDFS is inefficient.

Files and blocks are name objects in HDFS and they occupy namespace. The namespace capacity of the system is naturally limited by the physical memory in the NameNode.

When there are large numbers of small files stored in the system, metadata occupies large portion of the namespace.

In case of MapReduce, when file size is very small, the input for each process is very little and there are large number of map tasks. For example, a 10GB file broken up into files of size 100KB each, use a map of their own. Thus the time taken to finish the job considerably increases.

Hadoop is not restricted to processing plain text data. For user custom binary data type, one can use the SequenceFile. SequenceFile is a flat file consisting of binary key/value pairs (although you can always use Text data type for the key). In fact, internally Hadoop uses SequenceFile to store the temporary outputs of maps.   
  
The other objective of using SequenceFile is to 'pack' many small files into a single large SequenceFile for the MapReduce computation since the design of Hadoop prefers large files (Remember that Hadoop default block size for data is 64MB).

**Problem Statement:**

Let’s assume we have lots and lots of small images files and need to remove duplicate files from the available data.

As most binary formats—particularly those that are compressed or encrypted—cannot be split and must be read as a single linear stream of data. Using such files as input to a MapReduce job means that a single mapper will be used to process the entire file, causing a potentially large performance hit.

In such a situation, it is preferable to either use a splitable format such as SequenceFile, or, if you cannot avoid receiving the file in the other format, do a preprocessing step that converts it into a splitable format.

Step 1:

For Demo here we show how to create SequenceFile from lot of image files.

1. Move files to hdfs
2. Prepare list of files which has to be given as input to our mapper
3. Use BinaryFilesToHadoopSequenceFile to convert

Step 2:

Write Map Reduce program to remove duplicate files from the given SequenceFile.

1. Run ImageDriver with ImageDuplicatesMapper and ImageDupsReducer on the created SequenceFile.
2. Check the list of Non duplicated files.