HadoopMapReduceSequenceFileFormat

Sequence File Format

A complex project using Hadoop often requires multiple map-reduces to happen in series. While the input data may be textual, it is extremely helpful to maintain intermediate data in the SequenceFile format.

SequenceFile's allow you to skip avoid parsing lines of input data into <key, value> pairs. Instead, the mapper will receive the exact <key, value> pairs that were emitted by the reducer who created the data.

This format is easily used by setting the output format of a job to be SequenceFileOutputFormat: JobConf.setOutputFormat(SequenceFileOutputFormat. class), and setting all successive jobs to use SequenceFileInputFormat: JobConf.setInputFormat(SequenceFileInputFormat.class).

While the files are not exactly human readable, their use greatly eases the implementation of map reduce sequences.