Sort

Sort Example

The **Sort** example simply uses the map/reduce framework to sort the input directory into the output directory. The inputs and outputs must be Sequence files where the keys and values are BytesWritable.

The mapper is the predefined IdentityMapper and the reducer is the predefined IdentityReducer, both of which just pass their inputs directly to the output.

To run the program:

bin/hadoop jar hadoop-*-examples.jar sort [-m <#maps>] [-r <#reduces>] <in-dir> <out-dir>

Running Sort Benchmark

To use the sort example as a benchmark, generate 10GB/node of random data using RandomWriter. Then sort the data using the sort example. This provides a sort benchmark that scales depending on the size of the cluster. By default, the sort example uses 1.0 * capacity for the number of reduces and depending on your cluster you may see better results at 1.75 * capacity.

The commands are:
% bin/hadoop jar hadoop-*-examples.jar randomwriter rand

% bin/hadoop jar hadoop-*-examples.jar sort rand rand-sort The first command will generate the unsorted data in the *rand* directory. The second command will read that data, sort it, and write into the *rand-sort* directory.

Sort supports generic options : see DevelopmentCommandLineOptions