How Streaming Works

In the above example, both the mapper and the reducer are python scripts that read the input from standard input and emit the output to standard output. The utility will create a Map/Reduce job, submit the job to an appropriate cluster, and monitor the progress of the job until it completes.

When a script is specified for mappers, each mapper task will launch the script as a separate process when the mapper is initialized. As the mapper task runs, it converts its inputs into lines and feed the lines to the standard input (STDIN) of the process. In the meantime, the mapper collects the line-oriented outputs from the standard output (STDOUT) of the process and converts each line into a key/value pair, which is collected as the output of the mapper. By default, the prefix of a line up to the first tab character is the key and the rest of the line (excluding the tab character) will be the value. If there is no tab character in the line, then the entire line is considered as the key and the value is null. However, this can be customized, as per one need.

When a script is specified for reducers, each reducer task will launch the script as a separate process, then the reducer is initialized. As the reducer task runs, it converts its input key/values pairs into lines and feeds the lines to the standard input (STDIN) of the process. In the meantime, the reducer collects the line-oriented outputs from the standard output (STDOUT) of the process, converts each line into a key/value pair, which is collected as the output of the reducer. By default, the prefix of a line up to the first tab character is the key and the rest of the line (excluding the tab character) is the value. However, this can be customized as per specific requirements.

Important Commands

Parameters	Options	Description
-input directory/file-name	Required	Input location for mapper.
-output directory-name	Required	Output location for reducer.

-mapper executable or script or JavaClassName	Required	Mapper executable.
-reducer executable or script or JavaClassName	Required	Reducer executable.
-file file-name	Optional	Makes the mapper, reducer, or combiner executable available locally on the compute nodes.
-inputformat JavaClassName	Optional	Class you supply should return key/value pairs of Text class. If not specified, TextInputFormat is used as the default.
-outputformat JavaClassName	Optional	Class you supply should take key/value pairs of Text class. If not specified, TextOutputformat is used as the default.
-partitioner JavaClassName	Optional	Class that determines which reduce a key is sent to.
-combiner streamingCommand or JavaClassName	Optional	Combiner executable for map output.
-cmdenv name=value	Optional	Passes the environment variable to

		streaming commands.
-inputreader	Optional	For backwards-compatibility: specifies a record reader class (instead of an input format class).
-verbose	Optional	Verbose output.
-lazyOutput	Optional	Creates output lazily. For example, if the output format is based on FileOutputFormat, the output file is created only on the first call to output.collect (or Context.write).
-numReduceTasks	Optional	Specifies the number of reducers.
-mapdebug	Optional	Script to call when map task fails.
-reducedebug	Optional	Script to call when reduce task fails.