

Hadoop

IN ACTION

Chuck Lam



MANNING

Appendix A. HDFS file commands.....	1
--	----------

Appendix

HDFS file commands

This appendix lists the HDFS commands for managing files. They're in the form of

```
hadoop fs -cmd <args>
```

where `cmd` is the specific file command and `<args>` is a variable number of arguments.

You can see the command usage in the following convention. Parameters inside brackets ([]) are optional and ellipsis (...) means the optional parameter can be repeated. FILE is for filenames whereas PATH can be either filenames or directory names. SRC and DST are path names but they function specifically as source and destination, respectively. LOCALSRC and LOCALDST are further required to be on the local filesystem.

Command	Usage and description
cat	<pre>hadoop fs -cat FILE [FILE ...]</pre> <p>Displays the files' content. For reading compressed files, you should use the <code>text</code> command instead.</p>
chgrp	<pre>hadoop fs -chgrp [-R] GROUP PATH [PATH ...]</pre> <p>Changes the group association for files and directories. The <code>-R</code> option applies the change recursively. The user must be the files' owner or a superuser. See section 8.3 for more background information on the HDFS file permission system.</p>

Command	Usage and description
chmod	<pre>hadoop fs -chmod [-R] MODE[,MODE ...] PATH [PATH ...]</pre> <p>Changes the permissions of files and directories. Similar to its Unix equivalent, MODE can be a 3-digit octal mode, or {augo}+/-{rwxX}. The -R option applies the change recursively. The user must be the files' owner or a superuser. See section 8.3 for more background information on the HDFS file permission system.</p>
chown	<pre>hadoop fs -chown [-R] [OWNER][:[GROUP]] PATH [PATH ...]</pre> <p>Changes the ownership of files and directories. The -R option applies the change recursively. The user must be a superuser. See section 8.3 for more background information on the HDFS file permission system.</p>
copyFromLocal	<pre>hadoop fs -copyFromLocal LOCALSRC [LOCALSRC ...] DST</pre> <p>Identical to put (copy files from the local file system).</p>
copyToLocal	<pre>hadoop fs -copyToLocal [-ignorecrc] [-crc] SRC [SRC ...] LOCALDST</pre> <p>Identical to get (copy files to the local file system).</p>
count	<pre>hadoop fs -count [-q] PATH [PATH ...]</pre> <p>Displays the number of subdirectories, number of files, number of bytes used, and name for all files/directories identified by PATH. The -q option displays quota information.</p>
cp	<pre>hadoop fs -cp SRC [SRC ...] DST</pre> <p>Copies files from source to destination. If multiple source files are specified, destination has to be a directory.</p>
du	<pre>hadoop fs -du PATH [PATH ...]</pre> <p>Displays file sizes. If PATH is a directory, the size of each file in the directory is reported. Filenames are stated with the full URI protocol prefix. Note that although du stands for <i>disk usage</i>, it should not be taken literally, as disk usage depends on block size and replica factors.</p>
dus	<pre>hadoop fs -dus PATH [PATH ...]</pre> <p>Similar to du, but for a directory, dus reports the sum of file sizes in aggregate rather than individually.</p>
expunge	<pre>hadoop fs -expunge</pre> <p>Empties the trash. If the trash feature is enabled, when a file is deleted, it is first moved into the temporary .Trash/ folder. The file will be permanently deleted from the .Trash/ folder only after a user-configurable delay. The expunge command forcefully deletes all files from the .Trash/ folder. Note that as long as a file is in the .Trash/ folder, it can be restored by moving it back to its original location.</p>
get	<pre>hadoop fs -get [-ignorecrc] [-crc] SRC [SRC ...] LOCALDST</pre> <p>Copies files to the local filesystem. If multiple source files are specified, local destination has to be a directory. If LOCALDST is -, the files are copied to stdout.</p> <p>HDFS computes a checksum for each block of each file. The checksums for a file are stored separately in a hidden file. When a file is read from HDFS, the checksums in that hidden file are used to verify the file's integrity. For the get command, the -crc option will copy that hidden checksum file. The -ignorecrc option will skip the checksum checking when copying.</p>

Command	Usage and description
getmerge	<pre>hadoop fs -getmerge SRC [SRC ...] LOCALDST [addnl]</pre> <p>Retrieves all files identified by SRC, merges them, and writes the single merged file to LOCALDST in the local filesystem. The option addnl will add a newline character to the end of each file.</p>
help	<pre>hadoop fs -help [CMD]</pre> <p>Displays usage information for the command CMD. If CMD is omitted, it displays usage information for all commands.</p>
ls	<pre>hadoop fs -ls PATH [PATH ...]</pre> <p>Lists files and directories. Each entry shows name, permissions, owner, group, size, and modification date. File entries also show their replication factor.</p>
lsr	<pre>hadoop fs -lsr PATH [PATH ...]</pre> <p>Recursive version of ls.</p>
mkdir	<pre>hadoop fs -mkdir PATH [PATH ...]</pre> <p>Creates directories. Any missing parent directories are also created (like Unix mkdir -p).</p>
moveFromLocal	<pre>hadoop fs -moveFromLocal LOCALSRC [LOCALSRC ...] DST</pre> <p>Similar to put, except the local source is deleted after it's been successfully copied to HDFS.</p>
moveToLocal	<pre>hadoop fs -moveToLocal [-crc] SRC [SRC ...] LOCALDST</pre> <p>Displays a “not implemented yet” message.</p>
mv	<pre>hadoop fs -mv SRC [SRC ...] DST</pre> <p>Moves files from source(s) to destination. If multiple source files are specified, destination has to be a directory. Moving across filesystems is not permitted.</p>
put	<pre>hadoop fs -put LOCALSRC [LOCALSRC ...] DST</pre> <p>Copies files or directories from local system to destination filesystem. If LOCALSRC is set to -, input is set to stdin and DST must be a file.</p>
rm	<pre>hadoop fs -rm PATH [PATH ...]</pre> <p>Deletes files and empty directories.</p>
rmr	<pre>hadoop fs -rmr PATH [PATH ...]</pre> <p>Recursive version of rm.</p>
setrep	<pre>hadoop fs -setrep [-R] [-w] REP PATH [PATH ...]</pre> <p>Sets the target replication factor to REP for given files. The -R option will recursively apply the target replication factor to files in directories identified by PATH. The replication factor will take some time to get to the target. The -w option will wait for the replication factor to match the target.</p>
stat	<pre>hadoop fs -stat [FORMAT] PATH [PATH ...]</pre> <p>Displays “statistical” information on files. The FORMAT string is printed exactly but with the following format specifiers replaced.</p> <ul style="list-style-type: none"> %b Size of file in blocks %F The string “directory” or “regular file” depending on file type %n Filename %o Block size %r Replication %Y UTC date in yyyy-MM-dd HH:mm:ss format %Y Milliseconds since January 1, 1970 UTC

Command	Usage and description
tail	<pre>hadoop fs -tail [-f] FILE</pre> Displays the last one kilobyte of FILE.
test	<pre>hadoop fs -test [-ezd] PATH</pre> Performs one of the following type checks on PATH: <ul style="list-style-type: none"> -e PATH existence. Returns 0 if PATH exists. -z Empty file. Returns 0 if file length is 0. -d Returns 0 if PATH is a directory.
text	<pre>hadoop fs -text FILE [FILE ...]</pre> Displays the textual content of files. Identical to <code>cat</code> if files are text files. Files in known compressed format (gzip and Hadoop's binary sequence file format) are uncompressed first.
touchz	<pre>hadoop fs -touchz FILE [FILE ...]</pre> Creates files of length 0. Fails if files already exist and have nonzero length.