Hadoop HDFS

April 27, 2021

1 Working with HDFS from the command line

The HDFS shell is invoked by:

```
hadoop fs <CMD>
```

Most of the commands in HDFS shell behave like corresponding Unix commands. Error information is sent to stderr and the output is sent to stdout. The command `hdfs dfs` is a synonym of `hadoop fs`.

All HDFS shell commands take *path URIs* as arguments. The URI format is `scheme://authority/path`. For HDFS the scheme is `hdfs`, and for the local filesystem the scheme is `file`. The scheme and authority are optional. If not specified, the default scheme specified in the configuration is used (typically `hdfs://<namenodehost>`). An HDFS file or directory such as `/parent/child` can be specified as `hdfs://<namenodehost>/parent/child` or simply as `/parent/child`.

Relative paths can be used. For HDFS, the current working directory is the *HDFS home directory* `/user/<username>` that often has to be created manually. The HDFS home directory can also be implicitly accessed, e.g., when using the HDFS `input` folder, the `input` directory in the home directory.

See the online help for more information. In the following we quickly recap the most useful HDFS commands with the most commonly-used options.

1.1 cat

```
hadoop fs -cat URI [URI ...]
```

Copies source paths to stdout.

1.2 copyFromLocal

```
hadoop fs -copyFromLocal <localsrc> URI
```

Similar to the `hadoop fs -put` command, except that the source is restricted to a local file reference.

1.3 copyToLocal

```
hadoop fs -copyToLocal URI <localdst>
```

Similar to the `hadoop fs -get` command, except that the destination is restricted to a local file reference.

1.4 cp

```
hadoop fs -cp URI [URI ...] <dest>
```

Copy files from source to destination. This command allows multiple sources as well in which case the destination must be a directory.

1.5 df

```
hadoop fs -df [-h] URI [URI ...]
```

Displays free space. The `-h` option will format file sizes in a "human-readable" fashion (e.g., 64.0m instead of 67108864).

1.6 du

```
hadoop fs -du [-s] [-h] URI [URI ...]
```

Displays sizes of files and directories contained in the given directory or the length of a file in case its just a file. The `-s` option will result in an aggregate summary of file lengths being displayed, rather than the individual files. Without the `-s` option, calculation is done by going 1-level deep from the given path. The `-h` option will format file sizes in a "human-readable" fashion (e.g., 64.0m instead of 67108864).

The `du` command returns three columns with the following format:

```
size    disk_space_consumed_with_all_replicas    full_path_name
```

1.7 get

```
hadoop fs -get <src> <localdst>
```

Copy files to the local file system.

1.8 help

```
hadoop fs -help
```

Return usage output.

1.9 ls

```
hadoop fs -ls [-h] [-R] <args>
```

The `-h` option formats file sizes in a human-readable fashion (eg 64.0m instead of 67108864). The `-R` option recursively lists subdirectories encountered. For a file `ls` returns stat on the file with the following format:

```
permissions number_of_replicas userid groupid filesize modification_date modification_time file
```

For a directory it returns list of its direct children as in Unix. A directory is listed as:

`permissions userid groupid modification_date modification_time dirname`

Files within a directory are order by filename by default.

1.10 mkdir

`hadoop fs -mkdir [-p] <paths>` Takes path URIs as argument and creates directories. The `-p` option behavior is much like Unix `mkdir -p`, creating parent directories along the path.

1.11 put

`hadoop fs -put [-f] [ - | <localsrc1> .. ]. <dst>`

Copy single source, or multiple sources, from local file system to the destination file system. Also reads input from stdin and writes to destination file system if the source is set to "`-`". Copying fails if the file already exists, unless the `-f` flag is given.

1.12 rm

`hadoop fs -rm [-r |-R] URI [URI ...]`

Delete files specified as args. The `-R` option deletes the directory and any content under it recursively. The `-r` option is equivalent to `-R`.

1.13 rmdir

`hadoop fs -rmdir URI [URI ...]`

Delete a directory.

1.14 usage

`hadoop fs -usage <command>`

Return the help for an individual `<command>`.

[ ]: