# Hadoop and Spark Fundamentals
# The Linux Command Line/HDFS Cheat Sheet

For those new to the Linux command line. Version date: December 15, 2017

## Text Terminal Access

To access a Linux based Hadoop using the command line you need a text terminal connection. This includes connecting to a virtual machine on a laptop (i.e. there needs to be a way to connect to the virtual machine.) Depending on your computer or laptop a text terminal my already be available. In every case, the access method is secure shell or "ssh" that provides an encrypted pathways to and from the remote machine (or local VM)

| Operating System | Windows (free) | Mac OS | Linux |
|---|---|---|---|
| Text Terminal | Putty: *http://www.putty.org*<br>MobaXterm: *http://mobaxterm.mobatek.net** | Native "Terminal" under Utilities | Native |
| Access Method | ssh | ssh | ssh |

** recommended, mobaxterm also provides an X-windows client.*

## HOW TO CONNECT TO A HADOOP CLUSTER OR HOST

You will need an account on the cluster or host. This includes a USERNAME and PASSWORD. These should be assigned by a system administrator (or part of default user). The account may be on a "login node" of the cluster or some other host that has access to the cluster. To access the command line using an internet connected system use ssh to connect to the host:

```
ssh USERNAME@xxx.yyy.zzz.aaa  (IP address)
```

or

```
ssh USERNAME@somename.something  (fully qualified domain name, FQDN)
```

You will be asked for your password. If the login and connection were successful, you should see the following in you terminal window:

```
[USERNAME]$
```

The USERNAME may contain other information depending on your system. The $ prompt is where your typing will start. From this point on, only the $ prompt will be used to signify user input.

## HOW TO LIST FILES AND MOVE AROUND IN LOCAL LINUX DIRECTORIES

The Linux file system is like most standard directory based filesystems. You can see what is in your current directory by using the `ls` command. For example, after you login you will be in what is known as your *home* directory. The `ls` command will show the files in your current directory

```
$ ls
Hadoop_Fundamentals_Code_Notes-V3 README
```

In this directly there are two "files" one is a directory called Hadoop_Fundamentals_Code_Notes-V3 and one file called README. (In Linux a directory. looks like a file when using ls. A "long" listing is given if the "`-l`" option is given with `ls`. For example:

```
[deadline@limulus ~]$ ls -l
total 8
drwxr-xr-x 9 deadline deadline 4096 Dec  8  2014 Hadoop_Fundamentals_Code_Notes-V3
-rw-rw-r-- 1 deadline deadline   28 Jul 31 17:35 README
```

The long listing gives, the permissions, the owner, group, size, modification date and name. The "d" in front of Hadoop_Fundamentals_Code_Notes-V3 indicates that it is a directory. (depending on your systems, directories are often color coded for identification) `ls` can list what is in the directory. (the `-l` option can be used)

```
$ls Hadoop_Fundamentals_Code_Notes-V3/
Lesson-2  Lesson-3  Lesson-4  Lesson-5  Lesson-6  Lesson-7  Lesson-8  README  README.copyright
```

You can move the  Hadoop_Fundamentals_Code_Notes-V3 directory. Using the `cd` command (change directory)

```
$ cd Hadoop_Fundamentals_Code_Notes-V3/
$ ls
Lesson-2  Lesson-3  Lesson-4  Lesson-5  Lesson-6  Lesson-7  Lesson-8  README  README.copyright
```

To check what directory you are in the `pwd` command can be used to show your directory. path:

```
$ pwd
/home/deadline/Hadoop_Fundamentals_Code_Notes-V3
```

To move up a directory in the path use the "`..`" notation

```
cd ..
$ pwd
/home/deadline
```

Files may be copied using the "`cp`" command.

```
$ cp file1 file2
```

Files can be removed using the "rm" command.

```
$ rm file1
```

To rename a file, copy then remove. There is much more and this should be enough to get around in a Linux files system.

## HOW TO VIEW A TEXT FILE

A text file can be viewed in many ways. There are two simple ways to view a file. The first uses the "cat" command that will just print the file to the screen. for long files use the pipe "|" command and send the output to "more." Hit the space key to move through the file.

```
cat NOTES.txt |more
```

You can also use the vi (or vim) editor to view a text file.

```
vi NOTES.txt
```

To read and view the file there are three simple commands to use:

  CTRL-F move forward in the file

  CTRL-B move backward in the file

  :q quit the vi editor and return to the command line.

If the file is gibberish then it is not a text (plain ASCII) text file. You can check the type of file by using the "file" command.

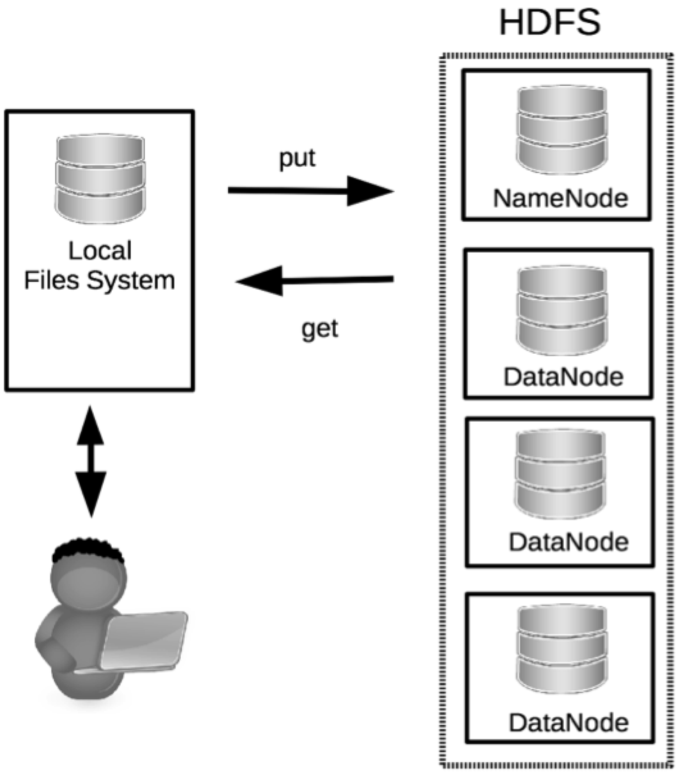
```
$ file NOTES.txt
NOTES.txt: ASCII English text
```

**HINT:** Start two two terminal windows to the Hadoop cluster. Use one window to view and read the NOTES.txt files and the other to type commands.

## HDFS Command Dereference

To interact with HDFS file system, the hdfs command must be used. All Hadoop processing happens in HDFS.

To use HDFS there are series of wrapper commands that provide a series of commands similar to those found in Linux/Unix file system.



## List Files in HDFS

To list the files in the root HDFS directory enter the following:

```
$ hdfs dfs -ls /
Found 8 items
drwxr-xr-x  - hdfs  hdfs   0 2013-02-06 21:17 /apps
drwxr-xr-x  - hdfs  hadoop 0 2014-01-01 14:17 /benchmarks
drwx------  - mapred hdfs  0 2013-04-25 16:20 /mapred
drwxr-xr-x  - hdfs  hdfs   0 2013-12-17 12:57 /system
drwxrwxr--  - hdfs  hadoop 0 2013-11-21 14:07 /tmp
drwxrwxr-x  - hdfs  hadoop 0 2013-10-31 11:13 /user
drwxr-xr-x  - doug  hdfs   0 2013-10-11 16:24 /usr
drwxr-xr-x  - hdfs  hdfs   0 2013-10-31 21:25 /yarn
```

To list files in your home directory **in HDFS** enter the following:

```
$ hdfs dfs -ls
Found 16 items
drwx------  - doug hadoop   0 2013-04-26 02:00 .Trash
drwxr-xr-x  - doug hadoop   0 2013-05-14 17:07 test
drwxr-xr-x  - doug hadoop   0 2013-05-14 17:23 test-output
drwx------  - doug hadoop   0 2013-05-15 11:21 war-and-
peace
drwxr-xr-x  - doug hadoop   0 2013-02-06 15:14 wikipedia
drwxr-xr-x  - doug hadoop   0 2013-08-27 15:54 wikipedia-
output
```

The same result can be obtained by issuing a:

```
$ hdfs dfs -ls /user/doug
```

To make a directory in HDFS use the following command. As with the –ls command, when no path is supplied the users home directory is used. (i.e. /users/doug)

```
$ hdfs dfs -mkdir stuff
```

To copy a file from your current local directory into HDFS use the following. Note again, that the absence a full path assumes your home directory. In this case the file test is placed in the directory stuff, which was created above.

```
$ hdfs dfs -put test stuff
```

The file transfer can be confirmed by using the –ls command:

```
$ hdfs dfs -ls stuff
```

```
Found 1 items
-rw-r--r--  3 doug hadoop   0 2014-01-03 17:03 stuff/test
```

Files can be copied back to your local file system using the following. In this case, the file we copied into HDFS, test, will be copied back to the current local directory with the name test-local.

```
$ hdfs dfs -get stuff/test test-local
```

The following will copy a file in within HDFS.

```
$ hdfs dfs -cp stuff/test test.hdfs
```

The following will delete the HDFS file test.dhfs that was created above (if the `skipTrash` option is not included a copy of the file will be moved to to your local .Trash directory).

```
$ hdfs dfs -rm -skipTrash test.hdfs


Deleted test.hdfs
```

The following will delete the HDFS directory stuff and all its contents.

```
$ hdfs dfs -rm -r stuff

Deleted stuff
```

## Freely Available Resources

**From DOS/Windows to Linux HOWTO** (somewhat dated but has many core concepts)

- *https://www.tldp.org/HOWTO/pdf/DOS-Win-to-Linux-HOWTO.pdf*

**Introduction to Linux: A Hands on Guide**

- *http://tille.garrels.be/training/tldp/*

**Basic Vi Commands**

- *https://www.cs.colostate.edu/helpdocs/vi.html*