

# HDFS Terminal Commands

## Logging into Master Node:

```
ssh -i "<Path to PEM file>" ubuntu@<public IP address>
```

## Hadoop Paths:

```
/usr/local/hadoop // Main Hadoop directory  
/usr/local/hadoop/hadoop_data/hdfs/namenode // Namenode directory
```

## HDFS Terminal Commands:

```
$ $HADOOP_HOME/bin/hadoop jar // Runs the MapReduce python scripts (input/output dir may change)  
    $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar  
    -input /user/purchases.txt -output /output/ -file mapper.py  
    -file reducer.py -mapper mapper.py -reducer reducer.py  
$ hdfs dfs -mkdir /<folderName> // Creates a directory called users on HDFS  
$ hdfs dfs -copyFromLocal ~/<fileName> / // Copies input file to HDFS directory  
$ hdfs dfs -tail /user/purchases.txt // Displays last few lines of file  
$ hdfs dfs -getmerge /output/ results.txt // Merges MapReduce results from HDFS into local directory  
$ hdfs dfs -rm -r /output // Removes the output directory  
$ hdfs dfs -ls / // Lists the directory (users) in the HDFS
```

## Hadoop General Notes:

- Make sure that python files have their EOL styles changed in Notepad++ before transfer (Edit → EOL Conversion → Unix)
- All python files have to have #!/usr/bin/python at the top.
- The map and reduce files should not have a mapper(): or reducer(): method
- Using Hadoop streaming with python, a space between emitted items will allow those items to act as the key
  - 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 entire line is considered as key and the value is null.

## Terminal Commands:

```
$ ls -lsa // List all files in directory  
$ touch <filename> // Create new file  
$ mv <oldName> <newName> // Rename file  
$ mv <filename> <directory> // Move file  
$ mv -t <directory> <filename1> <filename2> // Move multiple files  
$ rm <filename> // Delete file  
$ wget <downloadLink> // Download file  
$ gunzip <filename>.gz // Unzip file  
$ python <filename>.py // Run python file  
$ scp ~/Lab0/profile r2:~/ // Copying files between hostname nodes  
$ chmod 755 ~/<filename> // Changes the file permissions to default read/write  
$ head -50 shakespeare.txt // Prints out the first 50 lines  
$ echo -e "a b c d e f" | python mapper.py // Pipes input into the python file, to be read as stdin  
$ python RandomMapper.py // Prompts for input, which can be multiline (hit Ctrl-D after)
```

## Other:

```
Alt-Shift-T // Cuts from beginning of file in nano
```

## Powershell Commands:

```
gc -Head 10 .\shakespeare.txt | python // Pipes a subset of input data through the mapper and reducer files  
.\RandomMapper.py | Sort-Object | python .\RandomReducer.py
```