Working with HDFS on the Cluster

HDFS is somewhat mis-named: it's not a “filesystem” in the traditional sense. HDFS doesn't get mounted as a filesystem. You can't use the traditional Unix commands like `ls` or `cp` or `less` on the contents, because they aren't files in the computer's filesystems.

Instead, there are `hdfs dfs` equivalents that will go out to the cluster and do similar operations.

To list the contents of a directory: (respectively, your home directory, an `output` directory in your home directory, the directory of data sets for this course)

```
hdfs dfs -ls
hdfs dfs -ls output
hdfs dfs -ls /courses/353/
```

To examine the contents of HDFS files (in particular, a job's output), you can pipe the files out of HDFS and into a shell command: (for uncompressed and GZIP-compressed output)

```
hdfs dfs -cat output/part* | less
hdfs dfs -cat output/part* | gunzip | less
```

Assuming the files are small enough to be reasonably copied to the gateway's filesystem, you can copy them from HDFS to a local file like this:

```
hdfs dfs -copyToLocal output
```

Please occasionally clean up your output files on the cluster, just so they aren't taking up room:

```
hdfs dfs -rm -r out*
```

For more information, you can have a look at the HDFS filesystem commands *(https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-common/FileSystemShell.html)* . (Note: `hdfs dfs` and `hadoop fs` commands are synonyms. For some reason, their own docs use the older style.)

Updated Fri Aug. 28 2020, 11:21 by ggbaker.