

UNIX practice

On the UNIX command line. Go into your bigdata folder. If you have not used the cluster before then you will be in the `gen220` project. Or you may have your own lab bigdata folder.

See [Text Editors in UNIX](#) While it takes a few steps to install and setup, [VisualStudio](#) is a great resource you can edit on your local machine but saves changes on HPCC. See https://hpcc.ucr.edu/manuals/hpc_cluster/selected_software/vscode/

This should work

```
cd ~/bigdata
```

but if it doesn't

```
cd /bigdata/gen220/$USER # will go into your bigdata folder for the class
```

But if you already had an account on the cluster then

```
# if the above doesn't work you are likely already in a lab group on HPCC
```

```
cd /bigdata/$GROUP/$USER # should work since $USER is your login and $GROUP is your primary lab group
# you can see what groups you are in by typing
groups
```

For your homework:

1. Accept the homework 1 problem - (see link in Canvas).
2. Make a folder for this class and go into it (`~/bigdata` already points to your class folder `/bigdata/gen220/$USER`; see [UNIX I](#)):

```
mkdir -p ~/bigdata/gen220
cd ~/bigdata/gen220
```

3. Get the class data (see [UNIX II](#)):

```
git clone https://github.com/biodataprogram/GEN220.git # small example files in GEN220/data
git clone https://github.com/biodataprogram/GEN220_data.git # genomes and tables
ls GEN220_data/tabular
```

The `tabular` folder has the comma delimited table you will use below.

4. Checkout the homework 1 github repository created in step 1 (if you [setup SSH keys](#) in github)

```
git clone git@github.com:biodataprogram/2026-hw1-YOURGITHUBID.git
```

OR for the https will need to [create a token as your password](#)

```
git clone https://github.com/biodataprogram/2026-hw1-YOURGITHUBID.git
```

5. Go into your folder (`cd 2026-hw1-YOURGITHUBID`). The [Git and GitHub guide](#) (Recipe A) walks through these clone/commit/push steps in more detail.
6. Edit a script in there called `filesize.sh`; you can do this in jupyter on web, you can edit on the command line with `nano`, `vi`, or `emacs`, or you can [use visual studio tunnel](#)
7. Add some code to this script which achieves the directions at the bottom of this page
8. Test it out (run the `./filesize.sh`).
9. To submit your homework (and you can do this more than once), this requires doing `git commit` and then `git push`

```
# stage the file (needed the first time you commit a new file, and after every change)
git add filesize.sh
# this step saves a version of the code
git commit -m "This is a homework 1 solution"
```

```
# this step will push the data from HPCC or your computer UP to the github site
# this step will request your username (YOURGITHUBID) and your password (that TOKEN I mentioned before)
# if you have setup github account with SSH keys then it will ask you for your SSH key password
git push
```

10. You can repeat doing edits to the file, commit, and push to github.

Tasks for Homework 1

1. copy the `threatened-species.csv.gz` file - see info here [HW1](#) or you can just run the included `./setup.sh` script to download. but also encourage you to practice with `cp` command.
2. Write your answers in the script `filesize.sh`:
 - print out the size of the `threatened-species.csv.gz` using `du` or `ls -l`
 - Count the number of lines in file
 - Print out the number of unique phyla, order_name in the table using `cut`, `sort`, `uniq` (see the [UNIX III lab: Data processing with pipes](#) - work through its `cut` and `sort` and `uniq` sections before starting this)
 - count how many kingdom FUNGI are present in the file?
3. Check in your changes with `git add`, `git commit -m 'a message'` and `git push` to save the changes to github. Check the repository page on github.com to make sure your files are there - your last push before the deadline is what gets graded.