CS 679 - Homework 0: Getting Set Up

Eric Stevens

1) Logging In

I was able to set up ssh keys so I can log directly into bigbird without a password:

```
bash 20:04:11 ~/Desktop/OHSU/OHSU_Winter_2020/CS679-Dist_Comp  (maste
    ssh bigbird
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-93-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Overheard at KubeCon: "microk8s.status just blew my mind".

     https://microk8s.io/docs/commands#microk8s.status

18 packages can be updated.
17 updates are security updates.

New release '18.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** System restart required ***
Last login: Thu Jan  9 18:59:49 2020 from 10.95.40.35
stevener@bigbird0:~$
```

Question:

How long has bigbird0 been running since its last?

To answer this question I ran the `uptime` command:

```
stevener@bigbird0:~$ uptime
 20:15:08 up 251 days, 10:57,  2 users,  load average: 0.02, 0.01, 0.
```

So it appears the system has been running for **251 days, 10 hours and 57 minutes**.

2) Finding Data

Question:

How many different LDC corpora do we have in `/l2/corpora/ldc` ?

Listing the contents of the `/ld/corpora/LDC` directory we get the output:

```
stevener@bigbird0:/l2/corpora/LDC$ ls -l
total 524
drwxr-sr-x  3     5009   60029  4096 Sep 10  2008 2004L02
drwxr-sr-x  2 bayesteh student  4096 Jul 28  2016 CHiME
drwxr-sr-x  8     5009   60029  4096 Oct 17  2012 English_Gigaword.v3
-rw-r--r--  1    10962 student 34078 Jan 24  2018 grec
-rw-r--r--  1    10962 student 34078 Jan 24  2018 grep
drwxr-sr-x  2     5009   60029  4096 Oct 17  2012 hub4-awol
drwxr-sr-x  4     5009   60029  4096 Oct  7  2010 LDC1998T31
drwxr-sr-x  6     5009   60029  4096 Oct 17  2012 LDC2000S86
drwxr-sr-x  4     5009   60029  4096 Apr 24  2008 LDC2000S92
drwxr-sr-x  3     5009   60029  4096 Jun  9  2009 LDC2000T44
drwxr-sr-x  6     5009   60029  4096 Oct 17  2012 LDC2000T46
drwxr-sr-x  4     5009   60029  4096 Oct 17  2012 LDC2000T48
drwxr-sr-x  3     5009   60029  4096 Oct 17  2012 LDC2000T50
drwxr-sr-x  5     5009   60029  4096 Apr 24  2008 LDC2001S13
drwxr-sr-x  5     5009   60029  4096 Apr 24  2008 LDC2001S15

...

dr-xr-xr-x 26     5009   60029  4096 Oct 30  2013 LDC97S62
drwxr-sr-x  4    10865 student  4096 Jan 28  2016 LDC97T14
drwxr-sr-x 14     5009   60029  4096 Oct 17  2012 LDC97T22
drwxr-sr-x  5     5009   60029  4096 Apr 24  2008 LDC98S71
drwxr-sr-x  4     5009   60029  4096 Apr 24  2008 LDC98T28
drwxr-x--- 34     5009   60029  4096 Jul 31  2009 LDC99S79
drwxr-sr-x  9     5009   60029  4096 Apr 24  2008 LDC99S82
drwxr-sr-x  2     5009   60029  4096 Oct 17  2012 links_kmh
drwxr-xr-x  4     5009   60029  4096 Oct 26  2010 macrophone
drwxr-sr-x  7     5009   60029  4096 Feb 18  2010 MALACH_ENG
drwxr-sr-x  4     5009   60029  4096 Feb 25  2013 old
-rwxr-xr-x  1     5009   60029  1808 Mar 26  2014 readme
drwxr-sr-x  2     5009   60029  4096 Oct 17  2012 swb-awol
drwxr-sr-x  2     5009   60029  4096 Apr 17  2013 tarfiles
```

Assuming that the directories that are prepended with `LDC` are the corpora, then we can use the `grep` with a count option to get the number of corpora:

```
stevener@bigbird0:/l2/corpora/LDC$ ls -l | grep -c LDC
102
```

So it appears that **there are 102 LDC corpora** in the `/l2/corpora/ldc` directory.

Question: Pick any corpus

I stumbled across a folder containing some video files that intrigued me. The directory is located at:

```
bigbird0:/l2/corpora/IDS/IntuitiveTrip2007/videos
```

The contents of this video directory looked like this:

```
-rwxr-xr-x 1 5009 60029 3667562 Apr 25  2008 B_Suturing_1_capture1.av
-rwxr-xr-x 1 5009 60029 3519102 Apr 25  2008 B_Suturing_1_capture2.av
-rwxr-xr-x 1 5009 60029 2074932 Apr 25  2008 B_Suturing_2_capture1.av
-rwxr-xr-x 1 5009 60029 1987516 Apr 25  2008 B_Suturing_2_capture2.av
-rwxr-xr-x 1 5009 60029 2042886 Apr 25  2008 B_Suturing_3_capture1.av
-rwxr-xr-x 1 5009 60029 1954368 Apr 25  2008 B_Suturing_3_capture2.av
-rwxr-xr-x 1 5009 60029 2087098 Apr 25  2008 B_Suturing_4_capture1.av
-rwxr-xr-x 1 5009 60029 1969368 Apr 25  2008 B_Suturing_4_capture2.av
-rwxr-xr-x 1 5009 60029 1976614 Apr 25  2008 B_Suturing_5_capture1.av
...
```

I needed to discover what these video files contained. I was delighted to find that `ffmpeg` was installed on the bigbird, but when using `ffplay` it was very interesting to see how `ffmpeg` attempted to produce the video on the command line using mono-spaced color characters and backgrounds. While I found this impressive, and didn't know `ffmpeg` would do this, it was impossible to make out the video.

I had to come up with a way of running the video files. While port forwarding had been suggested, I didn't know whether the older browser on the bigbird system would support the codec for the specific video files. I did, however, know that FFMPEG would support the codec, since it was attempting to generate video out on the command line in bigbird.

The solution, mount the remote file system on my local machine and use my local FFMPEG to run the video files. This is accomplished using the `sshfs` utility. On my Mac I used home brew to install the `sshfs` utility by running the commands:

```
brew cask install osxfuse
```

followed by `brew install sshfs`

I then created a mounting directory called `droplet` on my local machine: `mkdir droplet`

From this point I can mount the remote file system from my local machine. This will give my local applications access to the remote files.

```
bash 21:44:51 ~  $:
    sshfs bigbird:/l2/corpora/IDS/IntuitiveTrip2007/videos droplet/

bash 21:45:31 ~  $:
    ls droplet | head
total 290816
-rwxr-xr-x  1 5009  60029  3667562 Apr 25  2008 B_Suturing_1_capture1
-rwxr-xr-x  1 5009  60029  3519102 Apr 25  2008 B_Suturing_1_capture2
-rwxr-xr-x  1 5009  60029  2074932 Apr 25  2008 B_Suturing_2_capture1
-rwxr-xr-x  1 5009  60029  1987516 Apr 25  2008 B_Suturing_2_capture2
-rwxr-xr-x  1 5009  60029  2042886 Apr 25  2008 B_Suturing_3_capture1
-rwxr-xr-x  1 5009  60029  1954368 Apr 25  2008 B_Suturing_3_capture2
-rwxr-xr-x  1 5009  60029  2087098 Apr 25  2008 B_Suturing_4_capture1
-rwxr-xr-x  1 5009  60029  1969368 Apr 25  2008 B_Suturing_4_capture2
-rwxr-xr-x  1 5009  60029  1976614 Apr 25  2008 B_Suturing_5_capture1
```

I can now play the videos with my local `ffplay` and see the full resolution.

1) What kind of data does it contain?

As it turns out, this data consists of many videos of people of various skill levels attempting a suturing task on an artificial wound using the robotic system da Vinci.

2) Where does the data come from?

Elsewhere in the parent directory I was able to track down a document that explains what this data is:

da Vinci Data Collection Session 8/18/04

“Goal: collect from three surgeons of differing skill levels during performance of a simple suturing task. Hopefully we’ll be able to identify the same skill differences using our automatic and objective technique.”

3) How big is the data?

First looking at the size on disk, we will ignore the fact that there are a few html files in there. Their size is negligible with respect to the video files. **We see that there are about 143 Megabytes** worth of video data.

```
bash 22:20:03 ~/droplet/IntuitiveTrip2007  $:
    du -sh videos/
143M    videos/
```

To get the number of video files we can use the same method we used to count the number of copra in the previous section.

```
bash 22:24:23 ~/droplet/IntuitiveTrip2007/videos  $:
    ls | grep -c .avi
50
```

We see that **there are 50 videos** that make up these 143 megs of data.

3. Running a Slurm job

```
stevener@bigbird0:~$ srun -N 20 hostname
bigbird62
bigbird1
bigbird16
bigbird9
bigbird20
bigbird3
bigbird17
bigbird12
bigbird13
bigbird5
bigbird11
bigbird2
bigbird8
bigbird19
bigbird15
bigbird4
bigbird14
bigbird18
bigbird7
bigbird6
```

40 appears to be to many nodes:

```
stevener@bigbird0:~$ srun -N 40 hostname
srun: error: Unable to allocate resources: Node count specification i
```

4. Viewing files in HDFS

Question: How many file and subdirectories are there in /data?

```
stevener@bigbird0:~$ hadoop fs -ls /data
Found 7 items
drwxr-xr-x   - hdfs hdfs           0 2018-01-23 14:33 /data/medline
-rw-r--r--   3 hdfs hdfs  2173502659 2018-01-09 14:09 /data/nyt_eng.t
drwxr-xr-x   - hdfs hdfs           0 2018-01-17 15:41 /data/nyt_split
drwxr-xr-x   - hdfs hdfs           0 2018-01-08 19:59 /data/pmc
-rw-r--r--   3 hdfs hdfs 20718119505 2018-01-26 11:06 /data/pmc_full.
drwxr-xr-x   - hdfs hdfs           0 2018-01-24 16:35 /data/reddit
drwxr-xr-x   - hdfs hdfs           0 2018-01-24 14:11 /data/scihub
```

There are **7 files and subdirectories** in the /data directory.

Question: How much space is available on our HDFS system?

```
stevener@bigbird0:~$ hadoop fs -df -h
Filesystem                              Size   Used  Available  Use%
hdfs://hadoopns1.cslu.ohsu.edu:8020  18.0 T  6.2 T     11.3 T   34%
```

There is **11.3 Terabytes** of space available on the HDFS.