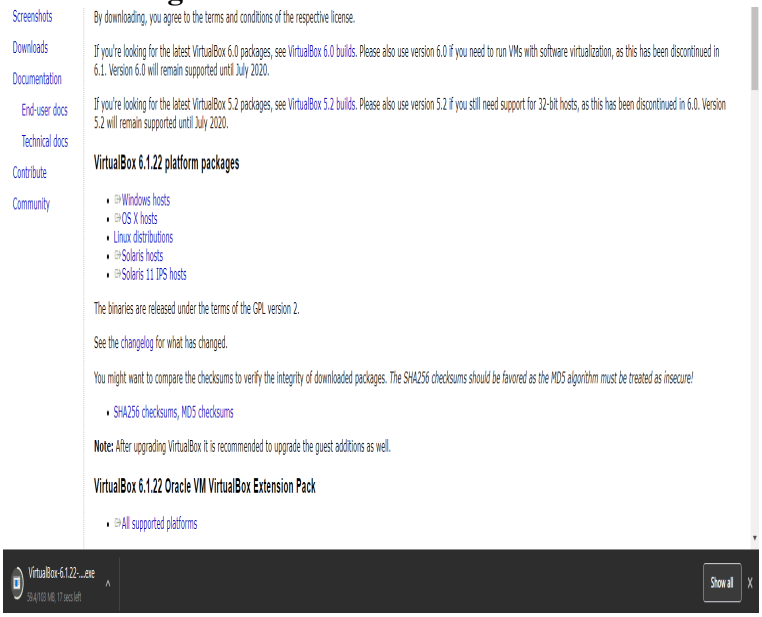
# REU BigDataX 2021 - Linux and Computer Systems HW

Matthew Chen

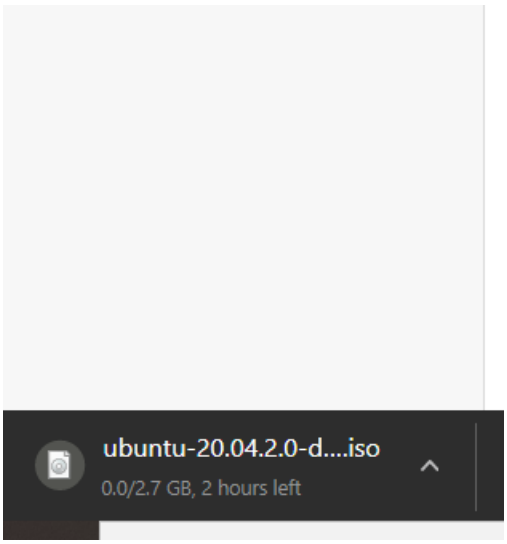
June 1, 2021

# 1 VM Setup and SSH

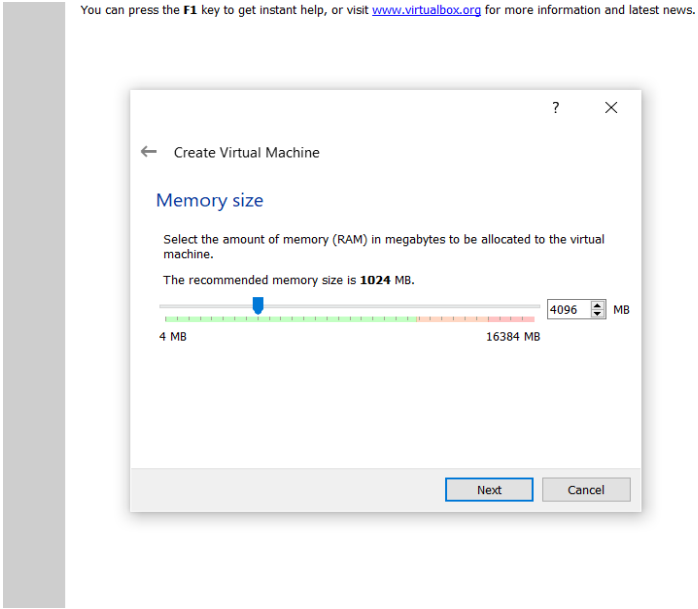
1. **Downloading VirtualBox!!!**



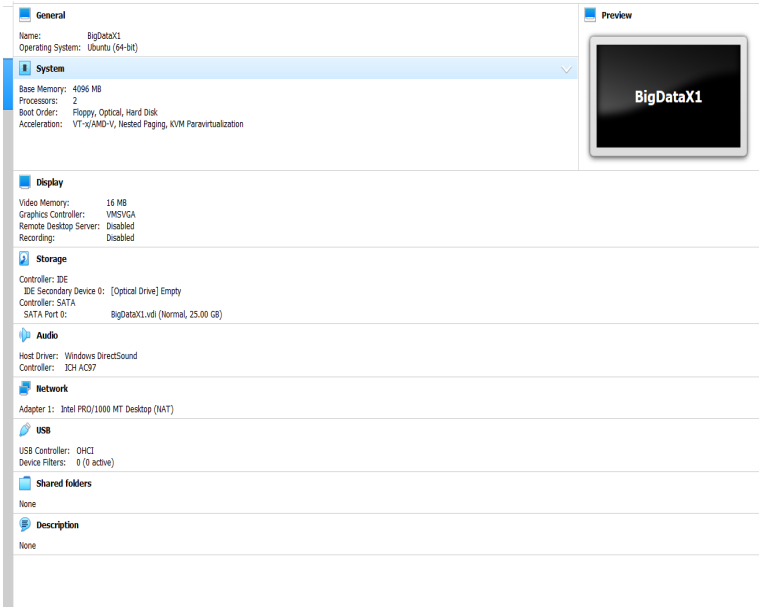

2. **Downloading Ubuntu ISO !!!**



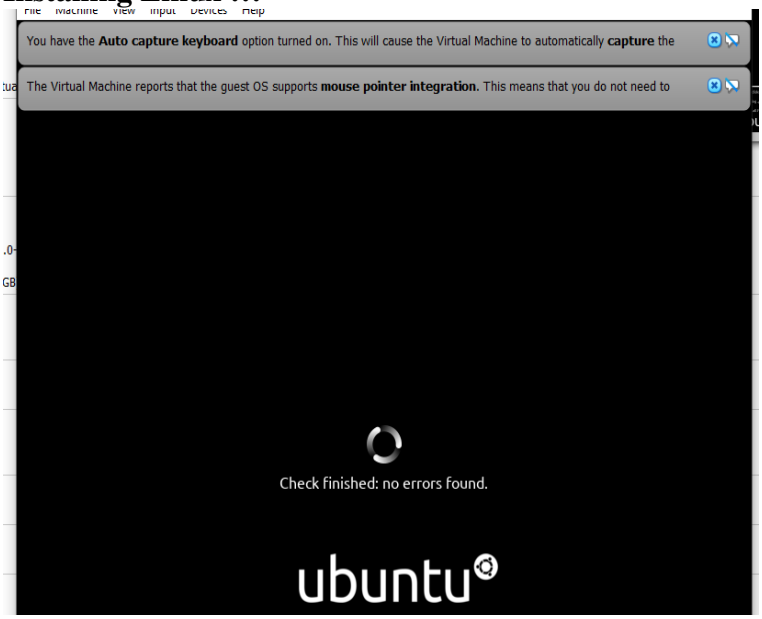

3. **Creating VM !!!**



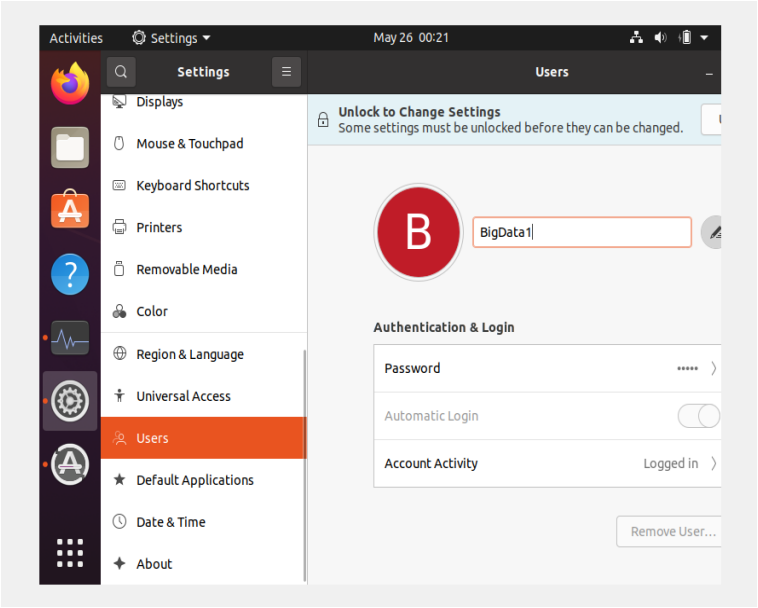

4. **Full VM !!!**



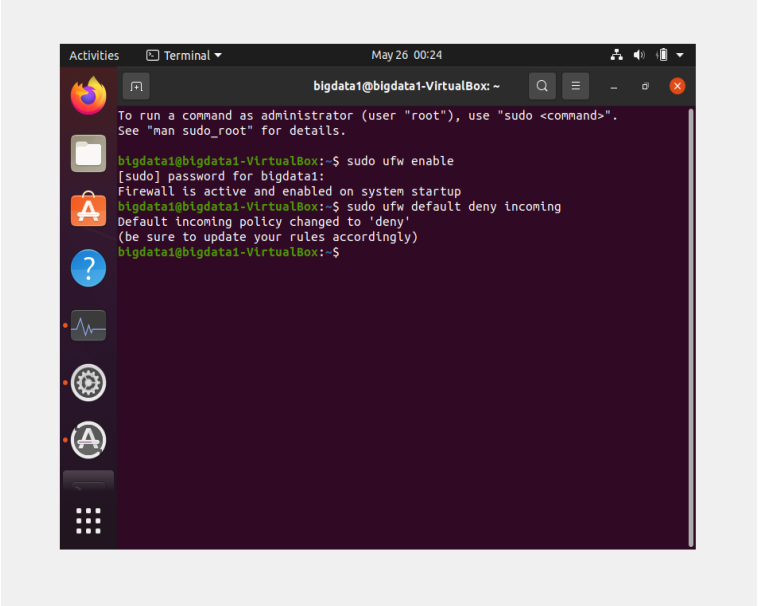

5. **Installing Linux !!!**




6. **Creating user id and password**



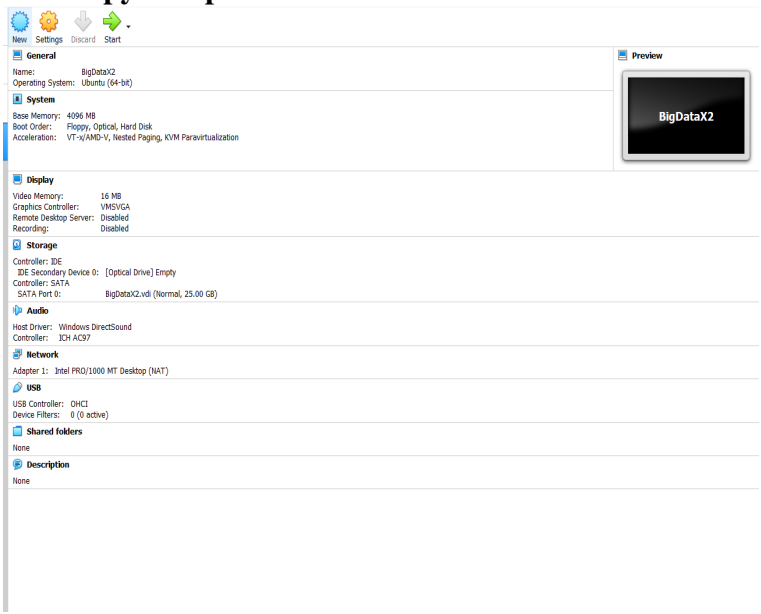

7. **Enabling firewall !!!**



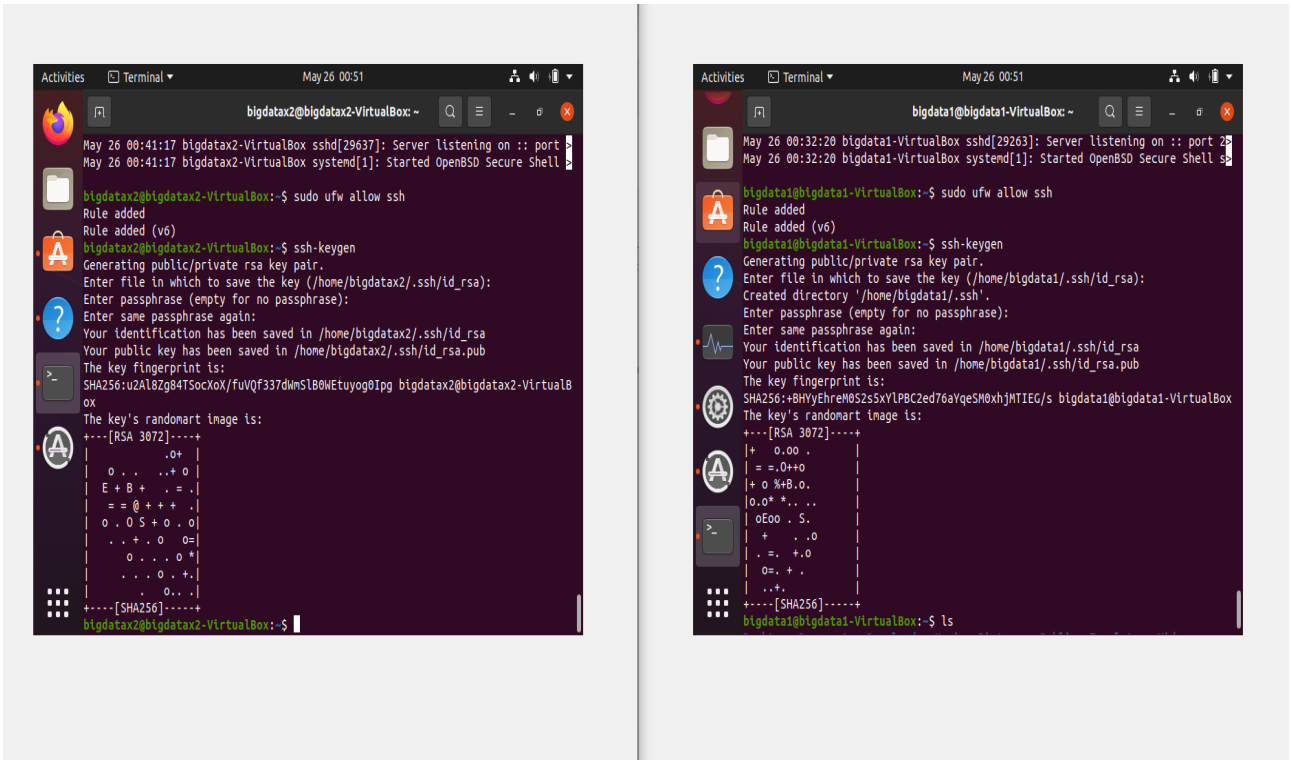

8. **Enabling SSH !!!**

```
Processing triggers for man-db (2.9.1-1) ...
Processing triggers for ufw (0.36-6) ...
bigdata1@bigdata1-VirtualBox:~$ sudo systemctl status ssh
● ssh.service - OpenBSD Secure Shell server
     Loaded: loaded (/lib/systemd/system/ssh.service; enabled; vendor preset: >
     Active: active (running) since Wed 2021-05-26 00:32:20 EDT; 1min 24s ago
       Docs: man:sshd(8)
             man:sshd_config(5)
   Main PID: 29263 (sshd)
      Tasks: 1 (limit: 4653)
     Memory: 1.1M
     CGroup: /system.slice/ssh.service
             └─29263 sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups

May 26 00:32:20 bigdata1-VirtualBox systemd[1]: Starting OpenBSD Secure Shell >
May 26 00:32:20 bigdata1-VirtualBox sshd[29263]: Server listening on 0.0.0.0 p>
May 26 00:32:20 bigdata1-VirtualBox sshd[29263]: Server listening on :: port 2>
May 26 00:32:20 bigdata1-VirtualBox systemd[1]: Started OpenBSD Secure Shell s>

bigdata1@bigdata1-VirtualBox:~$ sudo ufw allow ssh
Rule added
Rule added (v6)
bigdata1@bigdata1-VirtualBox:~$
```

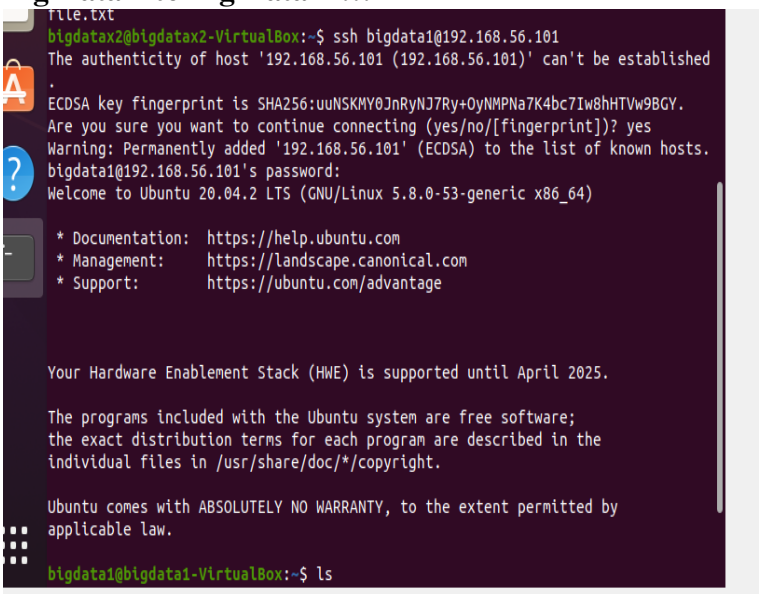
9. **VM Copy Setup !!!**




10. **Generating Key Pairs !!!**




11. **Big Data 2 to Big Data 1 !!!**




# 2 Linux Commands

1. ***ssh***
   **Description** - A program for logging into remote machines and executing commands on

them. Specifically connects and logs into a specified host with an optional username in which the user must provide identificaiton to access the machine.

```
matthewchen@ThePC:~$ ssh mc52@ews.illinois.edu
The authenticity of host 'ews.illinois.edu (192.17.90.133)' can't be established.
ECDSA key fingerprint is SHA256:1T+NCIo/DH7oEdAojbiymaaq5BIwQOuHO9lNkyMV3Ls.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ews.illinois.edu,192.17.90.133' (ECDSA) to the list of kn
mc52@ews.illinois.edu's password:
Permission denied, please try again.
```

2. ***ssh-keygen***
   **Description** - A program that generates, manages, and converts authentication keys for the purposes of ssh. Is adaptable to SSH protocol version 1 and version 2, able to create RSA/DSA keys. It requires and optional passphrase to gain access to the private key. Will store at `.ssh/identity` or `.ssh/id_dsa` or `.ssh/id_rsa`

```
matthewchen@ThePC:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/matthewchen/.ssh/id_rsa):
/home/matthewchen/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/matthewchen/.ssh/id_rsa
Your public key has been saved in /home/matthewchen/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:8MoEf1y5ElCi/Qt65QbZm49POzJJM4DiM0PAA3SZTbA matthewchen@ThePC
The key's randomart image is:
+---[RSA 3072]----+
|=. o*.o..        |
|.o.o.+ o   .     |
| ..Eo.o . o      |
|  o .o.B o .     |
| o .  *.S .      |
|  =  + B+=       |
|   +. +.*+.      |
|     . .++..     |
|        .++.     |
+----[SHA256]-----+
matthewchen@ThePC:~$
```

3. ***scp***
   **Description** - Using ssh for data transfer, the program copies files between hosts on a network. Will need the same authentication as if one is connecting to another machine for ssh.

```
matthewchen@ThePC:~$ scp demoVM.txt mc52@ews.illinois.edu:~
```

4. ***history***

**Description** - Lists all previously used commands, including metadata such as the time and date they were executed.

```
matthewchen@ThePC:~$ history
    1  jupyter notebook --no-browser --port=1234
    2  sudo apt install jupyter-notebook
    3  sudo apt install jupyter-core
    4  sudo apt install jupyter-notebook
    5  sudo apt update
    6  sudo apt upgrade
    7  sudo apt install jupyter-notebook
    8  ip addr show
    9  jupyter notebook --no-browser --port=1234
   10  ssh -N -f -L localhost:1234:localhost:1234 ubuntu@64.131.114.113
   11  ip a
   12  hostname -I
   13  ssh -N -f -L localhost:1234:localhost:1234 ubuntu@192.168.1.11
   14  ls
   15  cd
   16  ls
   17  vim
   18  bash
   19  sudo apt install qemu-kvm libvirt-bin virtinst cloud-utils
   20  sudo apt-get install qemu-kvm libvirt-daemon-system libvirt-clients bridge-uti
   21  sudo reboot
   22  wget https://cloud-images.ubuntu.com/bionic/current/bionic-server-cloudimg-amd
   23  ls
   24  sudo qemu-img info bionic-server-cloudimg-amd64.img
   25  sudo qemu-img resize bionic-server-cloudimg-amd64.img 40G
   26  sudo qemu-img info bionic-server-cloudimg-amd64.img
   27  sudo mkdir -p /var/lib/libvirt/images
   28  sudo qemu-img convert -f qcow2 bionic-server-cloudimg-amd64.img > /var/lib/lib
   29  sudo qemu-img convert -f qcow2 bionic-server-cloudimg-amd64.img \ > /var/lib/l
```

5. ***sudo***
   **Description** - Enables the user to execute a command as the superuser/other user

```
19  sudo apt install qemu-kvm libvirt-bin virtinst cloud-utils
```

6. ***ip***
   **Description** - Given a user-specified object, the program allows a user to view or manipulate routing, network devices, interfaces and tunnels of a machine

```
matthewchen@ThePC:~$ ip a
6: eth0: <> mtu 1500 group default qlen 1
    link/ether 60:a4:4c:c9:81:81
    inet 169.254.103.200/16 brd 169.254.255.255 scope global dynamic
       valid_lft forever preferred_lft forever
    inet6 fe80::7dfa:62d5:1773:67c8/64 scope link dynamic
       valid_lft forever preferred_lft forever
1: lo: <LOOPBACK,UP> mtu 1500 group default qlen 1
    link/loopback 00:00:00:00:00:00
    inet 127.0.0.1/8 brd 127.255.255.255 scope global dynamic
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host dynamic
       valid_lft forever preferred_lft forever
15: wifi0: <BROADCAST,MULTICAST,UP> mtu 1500 group default qlen 1
    link/ieee802.11 80:1f:02:95:66:67
    inet 192.168.1.11/24 brd 192.168.1.255 scope global dynamic
       valid_lft 2226sec preferred_lft 2226sec
    inet6 fd22:4e39:e630:1:923:c7f8:a58f:2ae8/64 scope global dynamic
       valid_lft forever preferred_lft forever
    inet6 fd22:4e39:e630:1:54f6:2020:7ed:1463/128 scope global dynamic
       valid_lft 556628sec preferred_lft 38181sec
    inet6 fe80::923:c7f8:a58f:2ae8/64 scope link dynamic
       valid_lft forever preferred_lft forever
5: wifi1: <> mtu 1500 group default qlen 1
    link/ieee802.11 80:1f:02:95:66:61
    inet 169.254.209.158/16 brd 169.254.255.255 scope global dynamic
       valid_lft forever preferred_lft forever
    inet6 fe80::fcda:5649:4fad:d19e/64 scope link dynamic
       valid_lft forever preferred_lft forever
9: wifi2: <> mtu 1500 group default qlen 1
    link/ieee802.11 80:1f:02:95:66:62
    inet 169.254.79.197/16 brd 169.254.255.255 scope global dynamic
       valid_lft forever preferred_lft forever
    inet6 fe80::3488:a14b:74b0:4fc5/64 scope link dynamic
       valid_lft forever preferred_lft forever
matthewchen@ThePC:~$
```

7. ***dd***
   **Description** - A program that allows a user to copy and convert a file to a different file type

```
or available locally via: info '(coreutils) dd invocation'
matthewchen@ThePC:~$ sudo dd if=demoVM.txt bs=1M of=pagman.txt
[sudo] password for matthewchen:
0+1 records in
0+1 records out
90 bytes copied, 0.0030431 s, 29.6 kB/s
```

8. ***fdisk***
   **Description** - Enables user to manipulate disk partition table

```
matthewchen@ThePC:~$ fdisk /dev/sda

Welcome to fdisk (util-linux 2.34).
Changes will remain in memory only, until you decide to write them.
Be careful before using the write command.

fdisk: cannot open /dev/sda: No such file or directory
```

9. ***apt***
   **Description** - A high level command line interface for the operating system's package management system.

```
matthewchen@ThePC:~$ sudo apt update
Get:1 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Hit:2 http://archive.ubuntu.com/ubuntu focal InRelease
Get:3 http://archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:4 http://archive.ubuntu.com/ubuntu focal-backports InRelease [101 kB]
Get:5 http://security.ubuntu.com/ubuntu focal-security/main amd64 Packages [667 kB]
Get:6 http://archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [990 kB]
Get:7 http://security.ubuntu.com/ubuntu focal-security/main Translation-en [135 kB]
```

10. ***vi***
    **Description** - A screen-oriented text editor where the terminal acts as a window into the editing buffer.

```
$ vi demoVM.txt
```

11. ***time***
    **Description** - Given a command with specified arguments, the program will print to standard error the timing statistics of the program run. These include elapsed real time, user CPU time, and system CPU time.

```
matthewchen@ThePC:~/BigDataLinuxHW$ time cat wow.txt
WOW



real    0m0.009s
user    0m0.000s
sys     0m0.016s
```

12. ***tar***
    **Description** - An archiving program designed to store multiple files into a single file, essentially compressing the files together.

```
matthewchen@ThePC:~$ tar -cvf home-5-27-21.tar ~/BigDataLinuxHW
tar: Removing leading `/' from member names
/home/matthewchen/BigDataLinuxHW/
/home/matthewchen/BigDataLinuxHW/.git/
/home/matthewchen/BigDataLinuxHW/.git/branches/
/home/matthewchen/BigDataLinuxHW/.git/COMMIT_EDITMSG
/home/matthewchen/BigDataLinuxHW/.git/config
/home/matthewchen/BigDataLinuxHW/.git/description
/home/matthewchen/BigDataLinuxHW/.git/FETCH_HEAD
/home/matthewchen/BigDataLinuxHW/.git/HEAD
/home/matthewchen/BigDataLinuxHW/.git/hooks/
/home/matthewchen/BigDataLinuxHW/.git/hooks/applypatch-msg.sample
```

13. ***cat***
    **Description** - Concatenates specified files and prints it to standard output (typically used to display contents of a file and sometimes create a new file)

```
matthewchen@ThePC:~$ cat demoVM.txt
#cloud-config
password:123456
chpasswd: {expire: False}
ssh_pwauth: True
hostname: demoVM
```

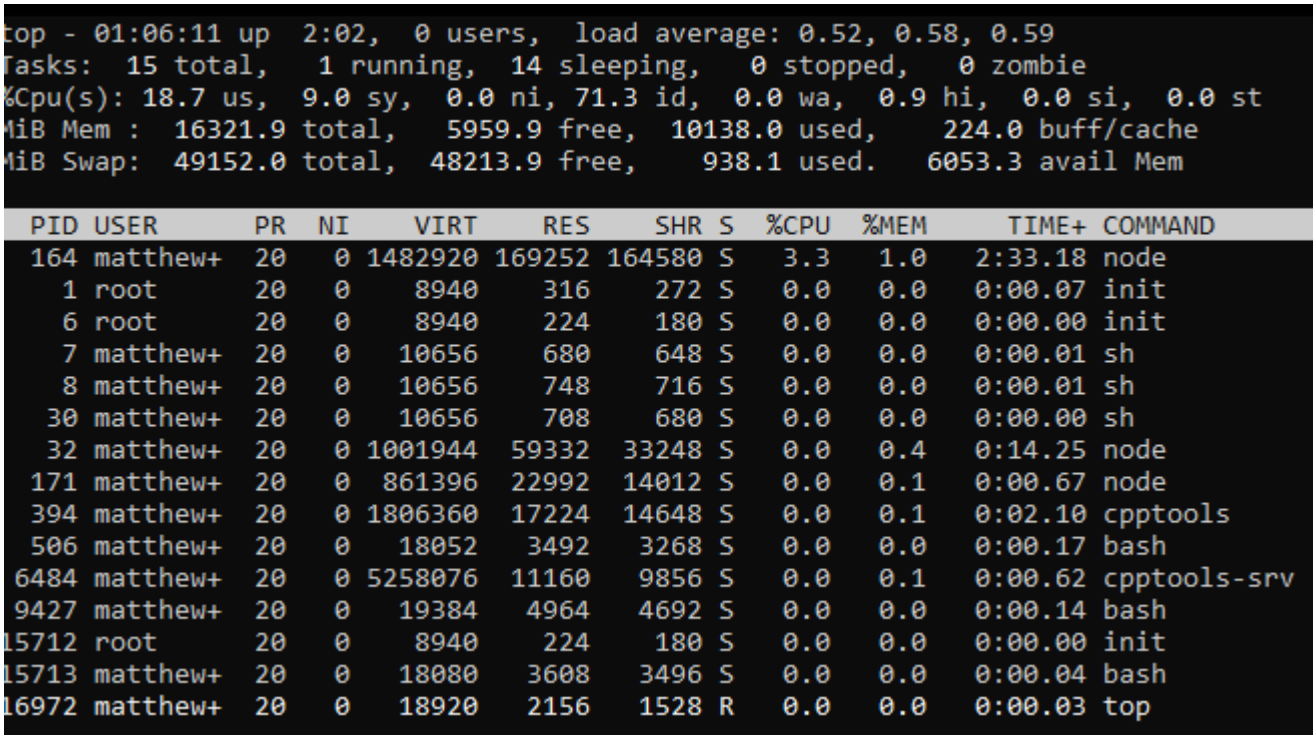
14. ***watch***
    **Description** - Runs a command repeatedly, displaying its output every time. By default it will execute a command every 2 seconds and repeat until interrupted.

```
Every 2.0s: cat demoVM.txt

#cloud-config
password:123456
chpasswd: {expire: False}
ssh_pwauth: True
hostname: demoVM
```

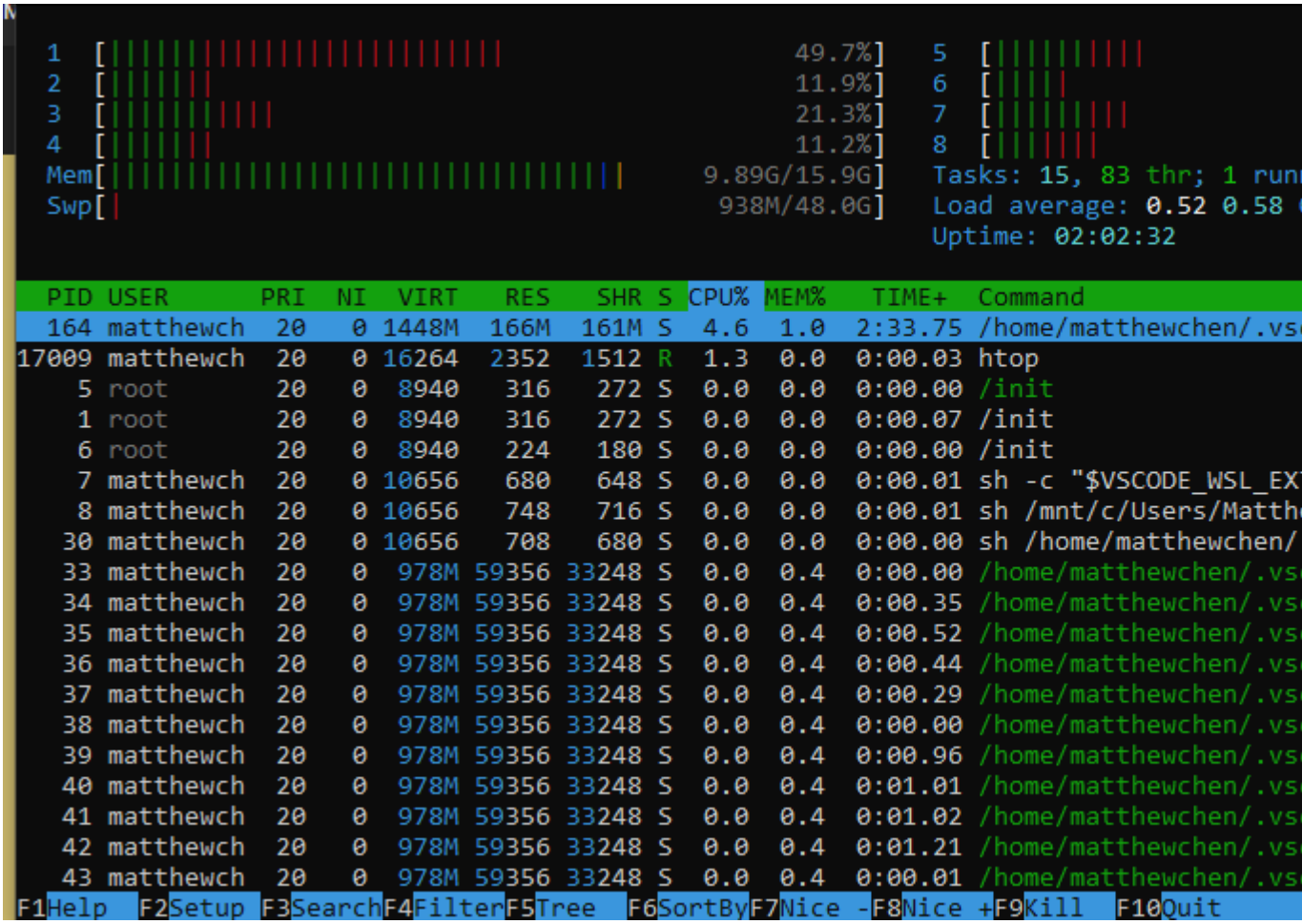
15. ***ps***
    **Description** - Reports a snapshot on a selection of active processes. by default selecting all processes associated with the current user and same terminal as the invoker.

```
matthewchen@ThePC:~$ ps
  PID TTY          TIME CMD
15713 tty2     00:00:00 bash
16935 tty2     00:00:00 ps
matthewchen@ThePC:~$
```

16. ***top***
    **Description** - A program that provides a dynamic real-time view of the running system by including information such as system summary information, lists of processes and threads and even an interface for process for process manipulation.




17. ***htop***
    **Description** - Similar to ***top***, it is an interactive process viewer for Linux that displays all processes in a concise, human-readable form that allows for easy process selection.




18. ***gcc***
**Description** - The GNU project C and C++ compiler which does preprocessing, compilation, assembly, and linking of files which outputs an executable of your compiled code.

```
matthewchen@ThePC:~/BigDataLinuxHW$ gcc sorted_data.c
```

19. ***tail***
**Description** - Given a file, it will print the last 10 lines of the of the file to standard output, otherwise it will read from standard input.

```
matthewchen@ThePC:~$ tail demoVM.txt
#cloud-config
password:123456
chpasswd: {expire: False}
ssh_pwauth: True
hostname: demoVM
```

20. ***grep***
**Description** - A program that searches for a specified pattern in each file and then prints

each line matching the pattern in a file. If no file is given then it will either read from the working directory or from standard input.

```
matthewchen@ThePC:~$ grep "password"
```

21. ***kill***
**Description** - Sends a system call to a group of process or process. Requires a parameter of pid and signal. If pid ¿ 0 then signal is sent to process with ID with pid. Else if pid = 0 then signal is sent to every process group of the calling process. If pid = -1 the signal is sent to all processes it has permission to send to execept for process 1. If pid ¡ -1 the signal is send to every process in process group with ID -pid.

```
matthewchen@ThePC:~$ kill 1212 1313 1414
```

22. ***killall***
**Description** - Performs the kill program on all processes performing user-specified commands. If not signal is specified SIGTERM is sent.

```
matthewchen@ThePC:~$ killall -o 5h
```

23. ***du***
**Description** - Summarizes the disk usage of a set of files, recursively from working directory.

```
matthewchen@ThePC:~$ du
0       ./.cache/vscode-cpptools/ipch/f0708cf7dd4b84a2
0       ./.cache/vscode-cpptools/ipch
0       ./.cache/vscode-cpptools
0       ./.cache
4       ./.config/htop
0       ./.config/procps
4       ./.config
0       ./.landscape
0       ./.local/share/jupyter/runtime
0       ./.local/share/jupyter
0       ./.local/share
0       ./.local
8       ./.ssh
0       ./.vscode-server/bin/ea3859d4ba2f3e577a159bc91e3074c5d85c0
0       ./.vscode-server/bin/ea3859d4ba2f3e577a159bc91e3074c5d85c0
36      ./.vscode-server/bin/ea3859d4ba2f3e577a159bc91e3074c5d85c0
28      ./.vscode-server/bin/ea3859d4ba2f3e577a159bc91e3074c5d85c0
```

24. ***df***
**Description** - A program that displays the amount of disk space available on the file system including each file name and metadata on the files.

```
matthewchen@ThePC:~$ df
Filesystem     1K-blocks      Used Available Use% Mounted on
rootfs         976136092 584939432 391196660  60% /
none           976136092 584939432 391196660  60% /dev
none           976136092 584939432 391196660  60% /run
none           976136092 584939432 391196660  60% /run/lock
none           976136092 584939432 391196660  60% /run/shm
none           976136092 584939432 391196660  60% /run/user
tmpfs          976136092 584939432 391196660  60% /sys/fs/cgroup
C:\            976136092 584939432 391196660  60% /mnt/c
```

25. ***screen***
    **Description** - A terminal multiplexer where one can start a screen session and then open virtual terminals inside the session which will run even if not visible or one gets disconnected.

```
GNU Screen version 4.08.00 (GNU) 05-Feb-20

Copyright (c) 2018-2020 Alexander Naumov, Amadeusz Slawinski
Copyright (c) 2015-2017 Juergen Weigert, Alexander Naumov, Amadeusz Slawinski
Copyright (c) 2010-2014 Juergen Weigert, Sadrul Habib Chowdhury
Copyright (c) 2008-2009 Juergen Weigert, Michael Schroeder, Micah Cowan, Sadrul Habi
Copyright (c) 1993-2007 Juergen Weigert, Michael Schroeder
Copyright (c) 1987 Oliver Laumann

This program is free software; you can redistribute it and/or modify it under the te
License as published by the Free Software Foundation; either version 3, or (at your

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRA
warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General

You should have received a copy of the GNU General Public License along with this pr
not, see https://www.gnu.org/licenses/, or contact Free Software Foundation, Inc., 5
Boston, MA  02111-1301  USA.

Send bugreports, fixes, enhancements, t-shirts, money, beer & pizza to screen-devel@

Capabilities:
+copy +remote-detach +power-detach +multi-attach +multi-user +font +color-256 +utf8
```

26. ***vim***
    **Description** - Similar to vi, vim is a screen oriented text editor but is the improved version (vi Improved)

```
matthewchen@ThePC:~$ vim demoVM.txt
```

27. ***chmod***
    **Description** - Allows a user to change the permissions of a file aka file mode bits

```
matthewchen@ThePC:~$ sudo chmod 777 demoVM.txt
[sudo] password for matthewchen:
```

28. ***chown***
    **Description** - Allows a user to change the owner and group of each user-specified file.

```
matthewchen@ThePC:~$ sudo chown linuxize demoVM.txt
```

29. ***useradd***
    **Description** - Creates a new user account using specified parameters or could change default new user information

```
matthewchen@ThePC:~$ sudo useradd forsen
```

30. ***man***
    **Description** - An interface to access Linux's system manual and learn of the different commands via a manual page

```
MAN(1)                                          Manual pager utils

NAME
       man - an interface to the system reference manuals

SYNOPSIS
       man [man options] [[section] page ...] ...
       man -k [apropos options] regexp ...
       man -K [man options] [section] term ...
       man -f [whatis options] page ...
       man -l [man options] file ...
       man -w|-W [man options] page ...
```

31. ***locate***
    **Description** - Lists files in databases that match a given pattern (usually via a regex expression)

```
matthewchen@ThePC:~$ locate "password"
/etc/pam.d/common-password
/home/matthewchen/.vscode-server/extens
ango-stubs/django-stubs/contrib/auth/pa
/home/matthewchen/.vscode-server/extens
ango-stubs/django-stubs/contrib/auth/ma
```

32. ***find***
    **Description** - Searches for files in a directory hierarchy

```
matthewchen@ThePC:~$ find BigDataLinuxHW/
BigDataLinuxHW/
BigDataLinuxHW/.git
BigDataLinuxHW/.git/branches
```

33. ***sed***
**Description** - A stream editor that can perform functions on streams such as searching, finding and replace, insertion and deletion.

```
matthewchen@ThePC:~$ sed 's/password/pissword/' demoVM.txt
#cloud-config
pissword:123456
chpasswd: {expire: False}
ssh_pwauth: True
hostname: demoVM
```

34. ***awk***
**Description** - A program that runs the awk scripting language which allows a user to write a short effective program that can scan a file line by line, splits each input line into fields, compares input line/fields to a pattern, and performs actions on found lines.

```
matthewchen@ThePC:~$ awk '/a/ {print $0}' demoVM.txt
password:123456
chpasswd: {expire: False}
ssh_pwauth: True
hostname: demoVM
```

35. ***diff***
**Description** - Compares files line by line and shows differences between the two if they are different and nothing if they are not

```
matthewchen@ThePC:~$ diff demoVM.txt pagman.txt
```

36. ***sort***
**Description** - Sorts a file assuming it contains ASCII characters in order but can have different sorting orders if the user specifies.

```
sort -b -no sorted.txt "$filename"
```

37. ***export***
**Description** - Marks the environment variables to be exported to child-processes

```
matthewchen@ThePC:~$ export
declare -x HOME="/home/matthewchen"
declare -x HOSTTYPE="x86_64"
declare -x LANG="C.UTF-8"
declare -x LESSCLOSE="/usr/bin/lesspi
declare -x LESSOPEN="| /usr/bin/lessp
```

38. ***pwd***
**Description** - prints path of current working directory

```
matthewchen@ThePC:~$ pwd
/home/matthewchen
```

39. ***crontab***
**Description** - Opens the cron table, which is a table that lists tasks scheduled to run at regular time intervals

```
matthewchen@ThePC:~$ */3 * * * * BigDataLinuxHW/generate-dataset.sh s
```

40. ***mount***
**Description** - mounts a file system from a device to the Linux big tree structure file system

```
matthewchen@ThePC:~$ sudo mount -l -t fuseblk
```

41. ***passwd***
**Description** - Allows a user to change their password

```
matthewchen@ThePC:~$ sudo passwd forsen
[sudo] password for matthewchen:
New password:
Retype new password:
passwd: password updated successfully
```

42. ***uname***
**Description** - Prints certain system information

```
matthewchen@ThePC:~$ uname
Linux
```

43. ***whereis***
**Description** - locates the binary, sources, and manual page files for a command

```
cat: /usr/bin/cat /usr/share/man/man1/cat.1.gz
```

44. ***whatis***
**Description** - displays the short manual page description associated with a name/command

```
matthewchen@ThePC:~$ whatis cat
cat (1)              - concatenate files and print on the standard output
matthewchen@ThePC:~$
```

45. ***su***
**Description** - run a command with a substitute user and group ID. If no user is specified it will run as root.

```
matthewchen@ThePC:~$ su matthewchen cat demoVM.txt
Password:
/usr/bin/cat: /usr/bin/cat: cannot execute binary file
```

46. ***ping***
**Description** - Elicits a response from a destination machine by sending it an `ICMP ECHO_REQUEST`

```
matthewchen@ThePC:~$ ping 104.160.131.3
PING 104.160.131.3 (104.160.131.3) 56(84) bytes of data.
```

47. ***traceroute***
**Description** - Prints the route a packet takes to reach a host with details at each of its stops

```
matthewchen@ThePC:~$ traceroute example.com
traceroute to example.com (93.184.216.34), 30 hops max, 60 byte packets
```

48. ***date***
**Description** - Print or set the system time

```
matthewchen@ThePC:~$ date
Thu May 27 02:14:41 EDT 2021
```

49. ***time***
**Description** - Given a command with specified arguments, the program will print to standard error the timing statistics of the program run. These include elapsed real time, user CPU time, and system CPU time.

```
matthewchen@ThePC:~/BigDataLinuxHW$ time cat wow.txt
WOW



real    0m0.009s
user    0m0.000s
sys     0m0.016s
```

50. ***wget***
**Description** - A non-interactive network downloader that allows a user to download files from the internet.

```
matthewchen@ThePC:~$ wget -P /mnt/iso http://mirrors.mit.edu/centos/7/isos/x86_64/Cen
--2021-05-27 02:15:19--  http://mirrors.mit.edu/centos/7/isos/x86_64/CentOS-7-x86_64-
```

51. ***wc***
**Description** - Prints newline, word, and byte counts for each file

```
matthewchen@ThePC:~$ wc demoVM.txt
 5  9 90 demoVM.txt
```

52. ***pwgen***
**Description** - A program that generates a password easily memorized by humans while being as secure as possible.

```
matthewchen@ThePC:~$ pwgen
Aad4aiV7 eesh1Ohy Ahriew5c zi8aeCha seileeS6 Aiyi1boo Iishu1Ee shoh8Ium
eb0aiNga eiR3oaM9 ahmah6Ai Eef3aT5z we5Si3Go Ae7wai3Z uo3OGhee iHah7eiv
Iiphoo4u tiu6heeH FaiXohv7 thooP6fo suW5kazi oophie2Y oidoNoh4 Eu7eifoh
Eegu2cho aeka6Aip uepe5uR5 Iag1yadu aic2PohK AhGon7xo niy1ieR0 hizeiS5L
Thav9eix bae6ooBi sen7eLei laD3bahj eGhi5shi Ahn0ucei cieTah3z Nah5eeNi
RiCh6cae eHav4zah oChuad4u zohJo5oo Eefoong3 Tai4oode ohk7eoN4 Coh3quie
Eephong0 Ajeep7jo iengoCo9 se0Nepov sexa5Iex ji2tePee Shu7caek eiX0cohm
Teexohy0 cais6AiL Ga2ach3f AeNg3noo GeF8yeiJ ahche6Lo Et8cee5p Aom6diem
jaN3koeg uasahG5n oow7ia9A hooXai0g Quoh0eik eekooR8L ciZ5lohp HeiW0oeb
Dei3LooD uJaipei2 ahPh9quo yae4AiPe haiyaeJ3 aShua8Ai kooLe8Oh Ooph4aix
kef4Ohb6 jei9beiF Pooghie0 SaQuaek7 mei3Dua4 aiN8ar0m oW6ahRah aek8Aib8
jee4Eila ayaeY5ai exaeJu0e Feithaa5 fohPhi6N Shoh0ia3 Chi8eeli Vo9aer1u
ahb0Loom Uam7choh vai5iMoo quuHei1o ooCh1Yoi aela7ieJ Ohxo2iup naimi4Hi
Chu4ein2 ooPid2ki pi3Poh0i quolee4T xeHah1ph pai9siCu oPhie4ez Hie9Pei3
weeDop7e Aex7aoZi ce8ku7Ph feiChae3 Utohsh4n CeiYi3wa eech3Dah Ei0Chae3
Ta9fohme eiZaej4e it8Opuur Veo4reo7 eg7DeiPu cho0Yaay eeC4wahT vaRohl0i
Ohquie7u Phai7ahp Mu1ooqui iePh5Ahp RaeJ8ieM uNoh6toh eiPh0doo Their6ko
ohL3ahch boo6Ques ohh5Ziep PhieNgi7 OM4Tooto eeHeeD8o naing8Lu Chai6nai
Ubohm9wi Ajoof5he shae8Dea Eid8cha4 Goo8eipe Airu4Quo Ahph8voo Jooquie9
ooDa1eis lav7OhD3 leH8Chuo Heiphic7 Doo6Eizo zee1ooM5 geeT3quo Ifu0veeg
```

# 3 Data Generation and Sorting analysis

1. Attached is the performance time for generating 50 records using my bash script and then sorting them on my computer (Intel i7 4770):

```
Matthew Chen@ThePC MINGW64 ~/Documents/BigDataLinuxHW (main)
$ time sh generate-dataset.sh poop.txt 50

real    0m10.212s
user    0m2.614s
sys     0m7.581s

Matthew Chen@ThePC MINGW64 ~/Documents/BigDataLinuxHW (main)
$ time sh sort-data.sh poop.txt

real    0m0.140s
user    0m0.000s
sys     0m0.076s

Matthew Chen@ThePC MINGW64 ~/Documents/BigDataLinuxHW (main)
$
```

2. Here are statistics for generating 1000, 100000, and 100000000 records as well (generated on Chameleon with Intel Haswell instance):

```
cc@mchenhw:~/BigDataLinuxHW$ time bash generate-dataset.sh poop1000.txt 1000

real    0m8.614s
user    0m8.559s
sys     0m6.334s
```

```
cc@mchenhw:~/BigDataLinuxHW$ time bash generate-dataset.sh poop100000.txt 100000

real    14m38.529s
user    14m30.269s
sys     10m48.412s
```

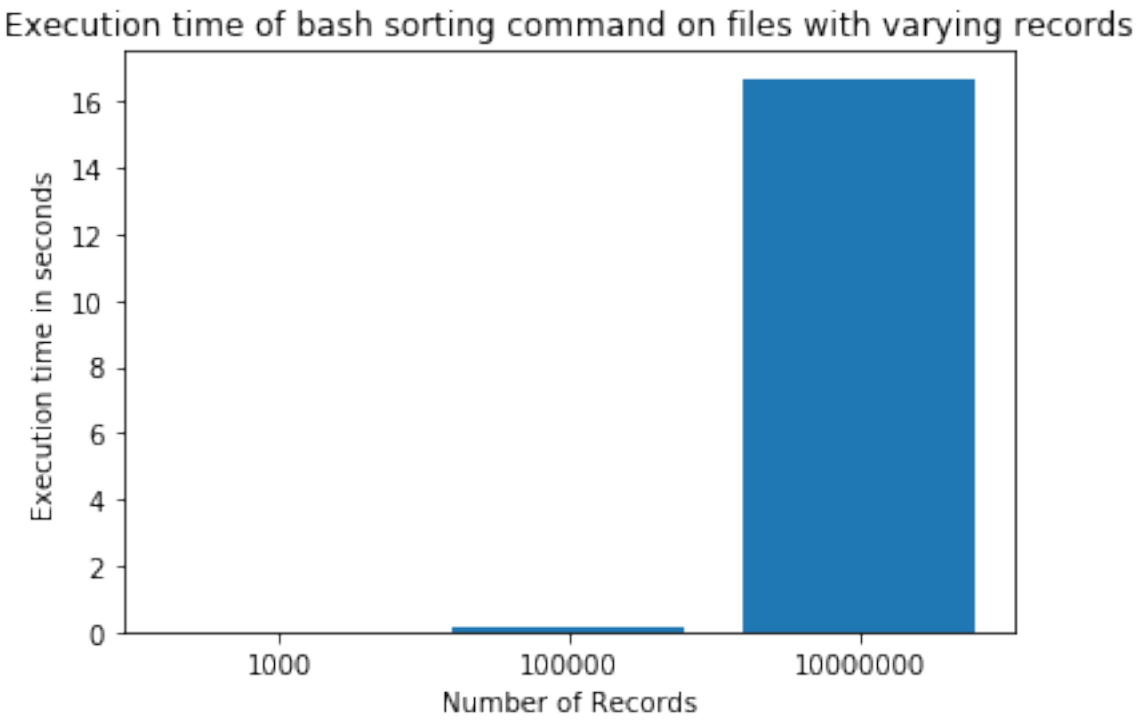
```
cc@mchenhw:~/BigDataLinuxHW$ time bash generate-dataset.sh poop10000000.txt 10000000

real    1470m27.442s
user    1462m36.468s
sys     1089m20.615s
cc@mchenhw:~/BigDataLinuxHW$
```

3. Here are the execution times for sorting the data via the bash command `sort`:

```
cc@mchenhw:~/BigDataLinuxHW$ time bash sort-data.sh poop1000.txt

real    0m0.006s
user    0m0.003s
sys     0m0.003s
```

```
cc@mchenhw:~/BigDataLinuxHW$ time bash sort-data.sh poop100000.txt

real    0m0.121s
user    0m0.097s
sys     0m0.024s
```

```
cc@mchenhw:~/BigDataLinuxHW$ time bash sort-data.sh poop10000000.txt

real    0m16.698s
user    0m22.488s
sys     0m3.091s
```

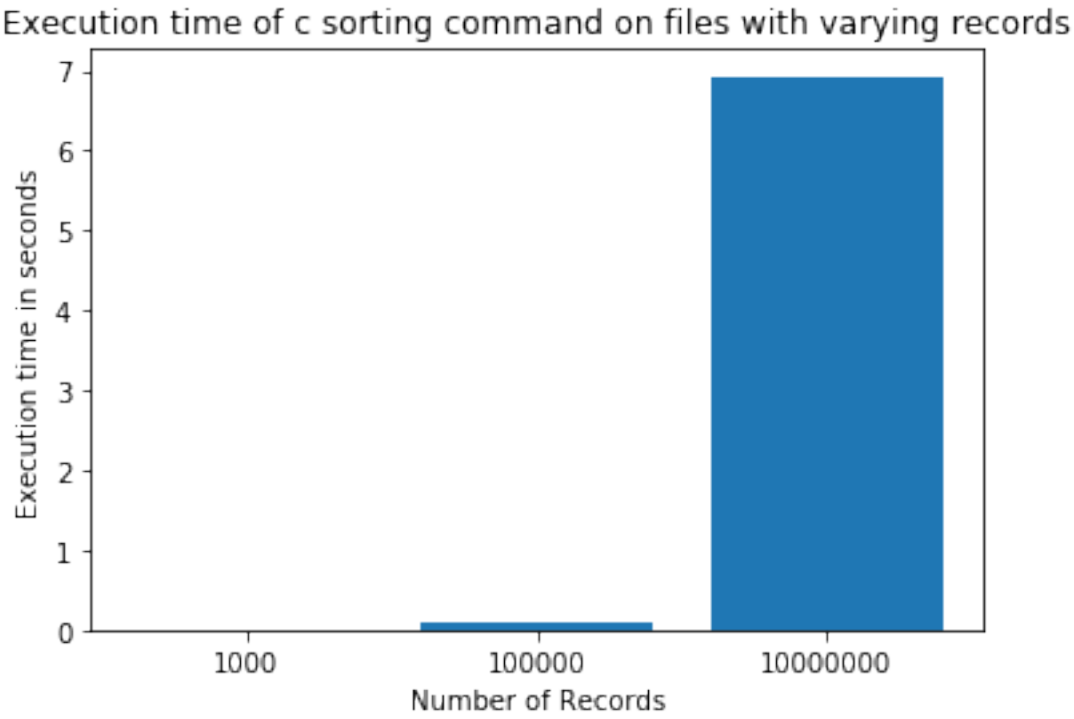



4. Here are the execution times for sorting the data via the c command `qsort`:

```
cc@mchenhw:~/BigDataLinuxHW$ time ./a.out
The program took 0.000284 seconds to execute

real    0m0.003s
user    0m0.003s
sys     0m0.000s
cc@mchenhw:~/BigDataLinuxHW$
```

```
cc@mchenhw:~/BigDataLinuxHW$ time ./a.out
The program took 0.023657 seconds to execute

real    0m0.091s
user    0m0.091s
sys     0m0.000s
cc@mchenhw:~/BigDataLinuxHW$
```

```
The program took 2.486576 seconds to execute

real    0m6.927s
user    0m6.635s
sys     0m0.292s
```

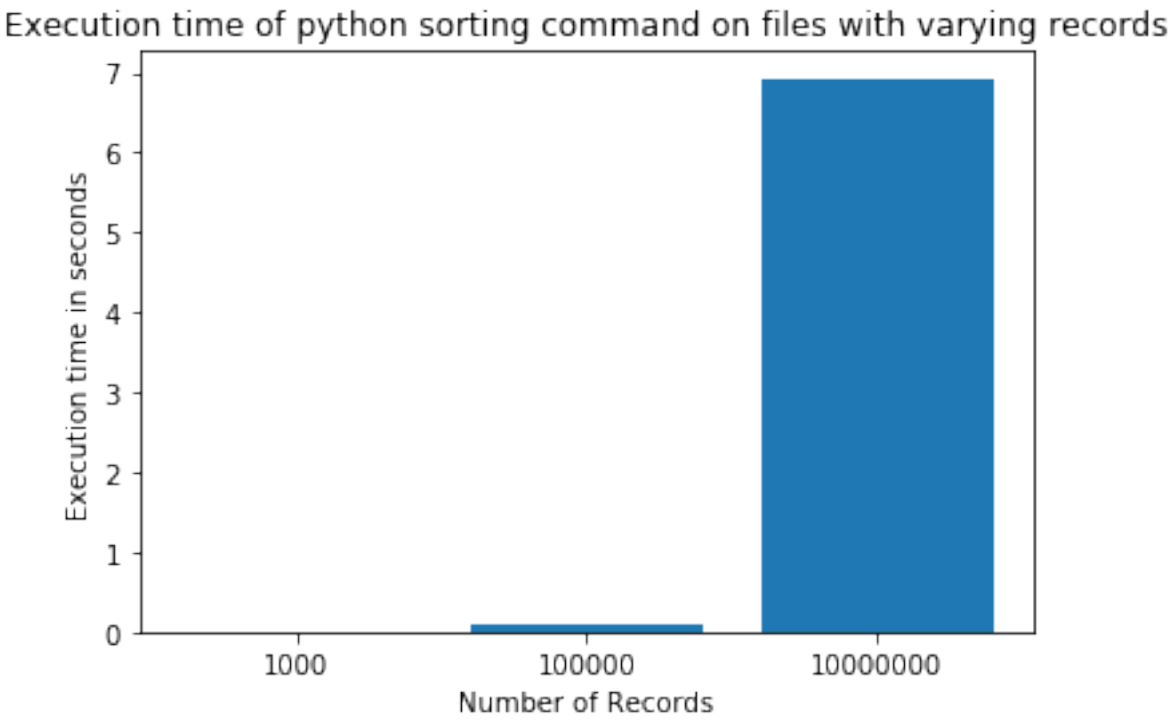



5. Here are the execution times for sorting the data via the python built-in function `sort`:

```
cc@mchenhw:~/BigDataLinuxHW$ time python sort-data.py
time-elapsed 0.0002200603485107422

real    0m0.036s
user    0m0.024s
sys     0m0.012s
cc@mchenhw:~/BigDataLinuxHW$ vim poop1000.txt
```

```
cc@mchenhw:~/BigDataLinuxHW$ time python sort-data.py
time-elapsed 0.03398537635803223

real    0m0.175s
user    0m0.159s
sys     0m0.016s
cc@mchenhw:~/BigDataLinuxHW$
```

```
cc@mchenhw:~/BigDataLinuxHW$ time python sort-data.py
time-elapsed 6.471749305725098

real    0m15.627s
user    0m15.158s
sys     0m0.468s
cc@mchenhw:~/BigDataLinuxHW$
```




# 4 Questions on VMs

1. Increasing the number of processors on the VM reduces the time it usually takes to execute a program because adding cores allows a computer to run processes in parallel. One scenario in which one might only need the minimum amount of cores is if one knows that their code would not benefit from parallel processing and wants to minimize the amount of resources consumed by the VM. On the other hand, one might need to use the maximum amount of processors if they do know that they can perform some parallel processing on their code which could drastically improve their execution speeds. Its potentially a bad idea to use the maximum amount of processors because sometimes the maximum is more than the cores present in your physical processor, so VirtualBox will have to either simulate extra cores or refuse to compute on these cores, meaning there would be no inherent benefit to adding more ”virtual” cores.

2. Specifying **None** turns off any paravirtualization interface. Specifying **Legacy** is an option for VMs created with older VirtualBox versions and will automatically pick a paravirtualization interface while starting the VM with VirtualBox 5.0. Specifying **Minimal** announces the presence of a virtualized enviroment, reporting the TSC and APIC frequency to the guest operating system. Furthermore **Minimal** is **mandatory for running any Mac OS X guests**. Specifying **Hyper-V** presents a Microsoft Hyper-V hypervisor interface that can be recognized by Windows 7 operating systems and newer. Furthermore **Hyper-V** is **reccomended for Windows guests**. Finally, **KVM** presents a Linux KVM hypervisor interface

which is recognized by Linus kernels starting from version 2.6.25. Furthermore **KVM** is **reccomended for Linux guests**. Because **KVM** is a paravirtualization provider designed for Linux systems, it would be the best provider for Ubuntu Linux.

3. Virtual Box must present real storage to a user as a virtual hard disk, and so needs a hard disk controller. Some examples of hard disk controllers include: **IDE (ATA), Serial ATA (SATA), and Non volatile memory express (NVMe)**. An **IDE** controller is a backwards compatible controller that works with hard disks, CD-ROM drives and other removable media. One could possibly use this if they want to best simulate a computer containing a CD-ROM drive. A **SATA** controller is a more recent standard of the **IDE** that supports higher speeds and more devices per controller. One could possibly use **SATA** to simulate a computer with a normal hard disk drive. On VirtualBox, the virtual **SATA** controller performs much faster and consumes less resources than a virtual **IDE** controller. A **NVMe** controller is a controller standard used for connecting non volatile memory over PCI express to exceed the bandwidth limits of the **SATA** protocol. A use case for the virtual **NVMe** controller would be to simulate high performance data retrieval and storage with non-volatile memory on a computer.

4. Virtual Box virtualizes several different networking modes including **NAT, Bridged Networking, Internal Networking, and Host-Only Networking**. **Network Address Translation (NAT)** is a simple networking mode that enables one to have internet access and prevents a guest from accessing other users/the host. Typically, one would use this mode if all they needed to do was browse the Web, download files, or view emails. **Bridged Networking** is a more advanced networking mode that allows for network simulations and running servers as a guest. Typically, one would use this networking mode when they have network heavy tasks in their VM, such as creating a server inside their VM that hosts an application. **Internal Networking** is a networking mode that helps create a software-based network which is only visible to select virtual machines, which means that the host and outside world are excluded. One could use this as a lightweight solution that enables users to collaborate and connect in privacy such as in a lab or private database. **Host-only Networking** is a networking mode that enables one to create a network with a host and a set of virtual machines without the need of the host's physical network interface (it creates a virtual network interface instead).

5. Virtual Box enables a virtual machine to directly interact with USB devices connected to the host machine via three configurations: **USB 1.1, 2.0, and 3.0**. By setting the USB controller to either OHCI, EHCI, or xHCI, the virtual USB controller will support the levels of USB 1.1, USB 2.0, or USB 3.0 respectively. Additionally, xHCI enables support for all USB speeds while EHCI enables support for OHCI.