Next Generation Sequencing Technologies: Data Analysis and Applications

Data Download and Exploration

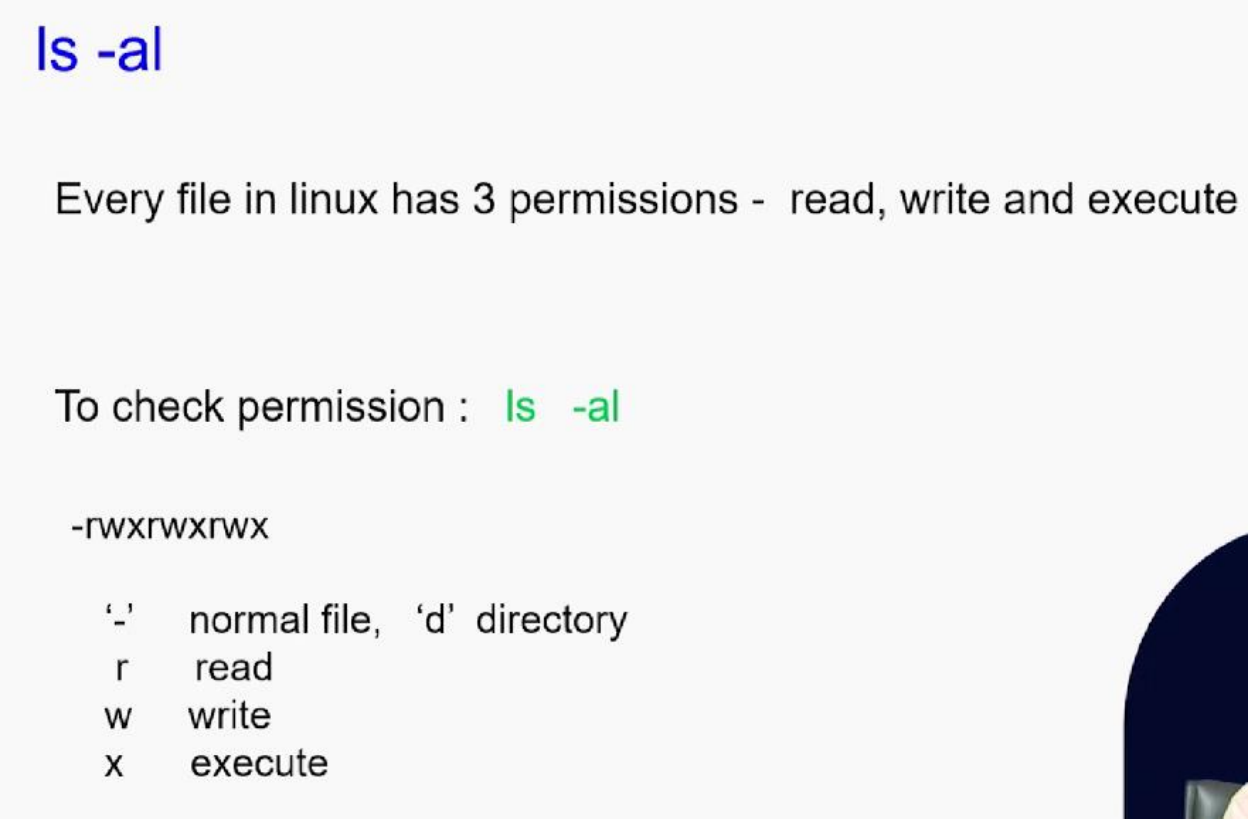
Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Next-Generation Sequencing Technologies: Data Analysis and Applications. We have been discussing basic commands that will be useful for NGA data analysis, and now we will move into data download and exploration. We will also use certain commands that will be useful for data download and exploration. We will talk about file permission, which is very useful. We will talk about system monitoring.

This is something that will come a bit later when we want to run certain commands. We want to monitor the system so that we do not overload it or crash it. And then, finally, we will look into some of the basics of data download and exploration. So these are the keywords for chmod top and vim editor. So these are some of the things that I will come across today.

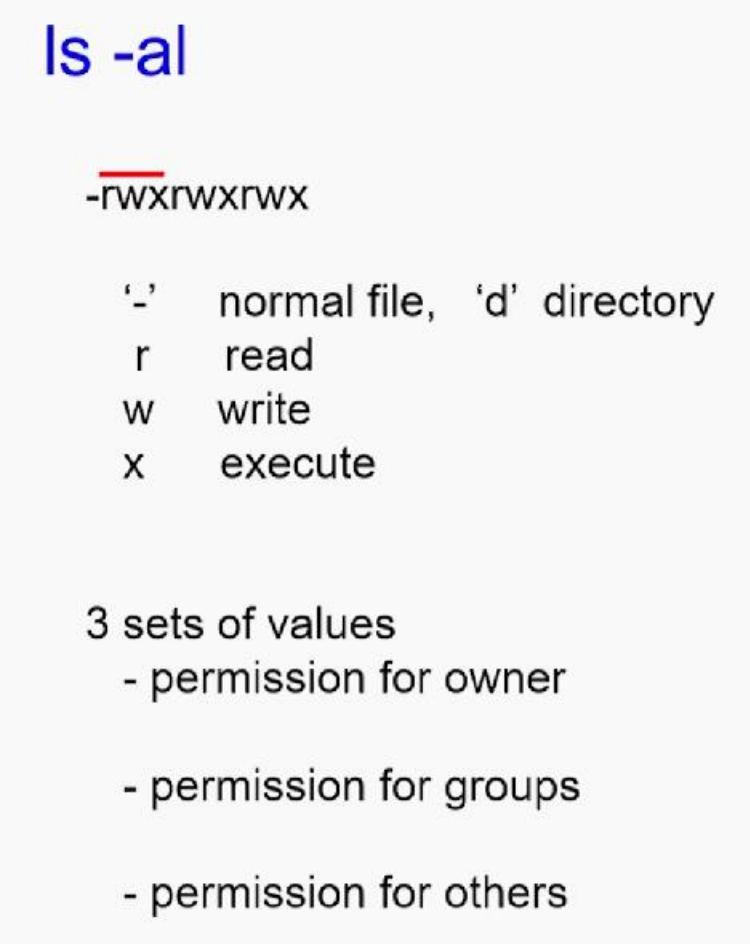
So, starting with file permission, So every Linux file has three permissions. So one is read, the second is write, and the third is execute. How can you check these permissions? These permissions we can check by using something called LS minus al. So this LS command you have seen already.



So this is the list of files; it will usually list these files that are in a directory. And on top of that, we can use these options minus al. When you run this minus al, you will see something like this. The first letter here is this minus sign. So this means this is a normal file.

But sometimes you will see this d in the first letter, and d means directory. So we will see this when we run these commands a bit later. Then you see this rwx repeated. So rwx rwx rwx. So what is this rwx? So r means read, w is write, and x is execute.

Now the question is: why is it repeated three times? So there are three sets of values. So the first set of values, the first set of RWX, is for permission for the owner. So it means the owner of this file has read, write, and execute permission. So the owner can do all these tasks. The second reason is for groups.

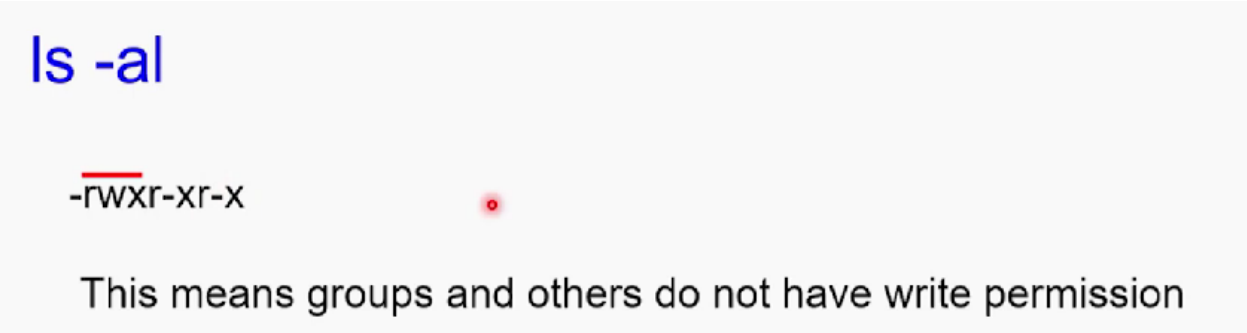


So there could be Linux groups, and they also have read, write, and execute permissions. The third set of rwx is for permission for others. So you have these three sets of values for read, write, and execute. Now the question is: can we change these permissions? This is good; we see these RWX values. Can we change these permissions? For example, I am creating a file as an owner in the Linux system, and I want to keep these permissions to write and execute to myself.

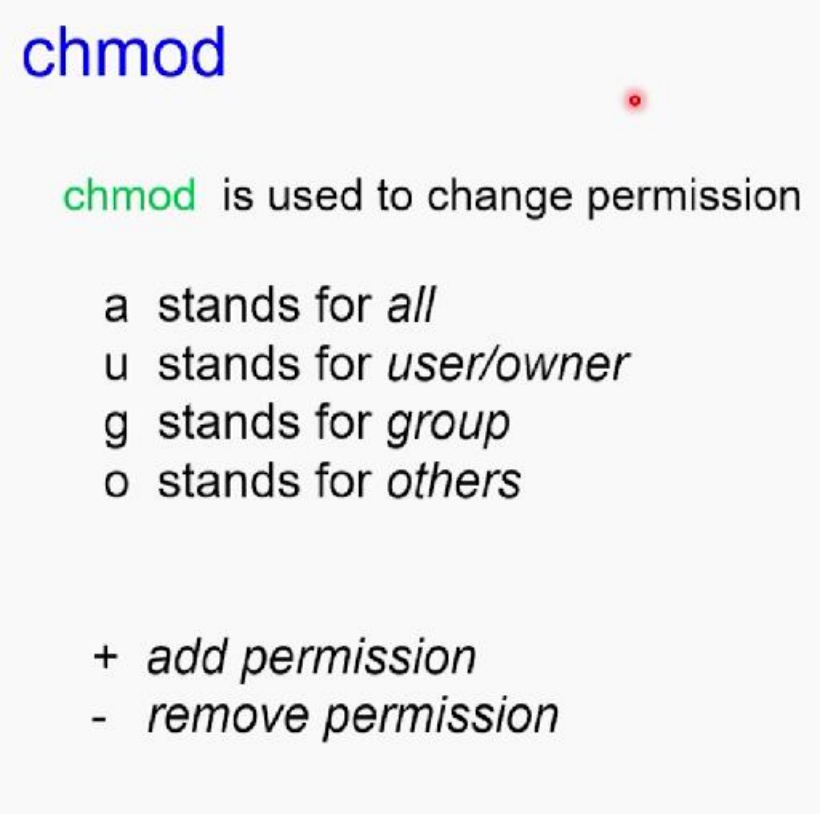
I do not want to give this write-and-execute permission to others. And this could be useful in Linux because sometimes you have multiple users, and you do not want some users, by mistake, to modify the system files, which are important for the running of the operating system. If those files are changed, maybe the system will not work properly and there will be some issues, errors, etc. So, as an administrator, you sometimes want to reserve these permissions for others.

And it turns out there is a way. As the owner of a file, you can actually do something like this. You can say the users might be able to read and execute, but they will not be able to write. They

will not be able to change the content of that file. So how do you do this? So this means if you see this kind of option, it means that this minus sign or dash means that the groups and the others do not have write permission.



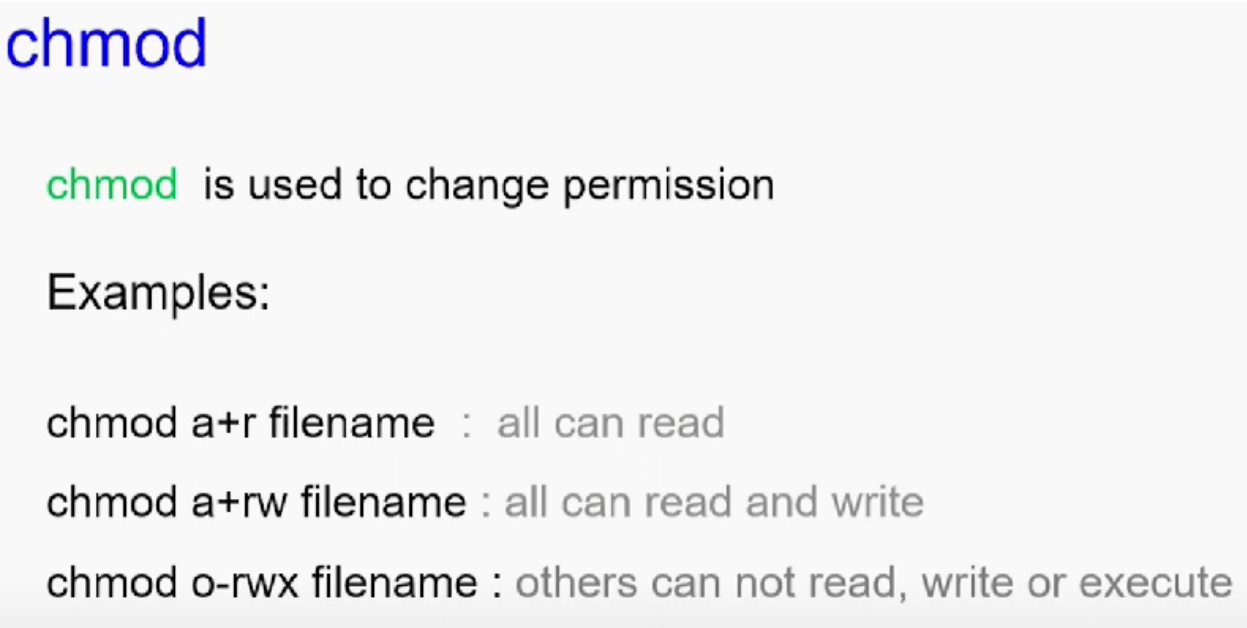
So if you see this is not there, it means these groups or the others will not have this write permission. And the command that we use to change this permission is called chmod. So change the permission, and again, you will have these multiple options with this chmod. You will see something called a right, so this stands for all, so the permission will be changed for everyone. If you see a U, if you give this option, it will stand for user or owner.



If you say G, it will stand for group, so you are changing permission only for groups, and O will stand for others, right? So if you mention this option, it will again change the permission only for

others, not for the owner, not for the groups. You can also use these options: plus, right, see signs plus or minus, so plus means add permission and minus means remove permission. So if you want to add, let us say write permission you say plus W. If you want to remove write permission you say minus W.

So we will see in a moment how these options work. So we will take some examples, and of course we will then run this in the terminal to better understand how this works. So we have chmod a plus r file name right so means you are changing this permission for a means all right plus r means everyone can read ok. So plus r file name so this you are changing this for a specific file name ok. You can also for example say a plus rw so you are giving read and write permission to everyone ok.

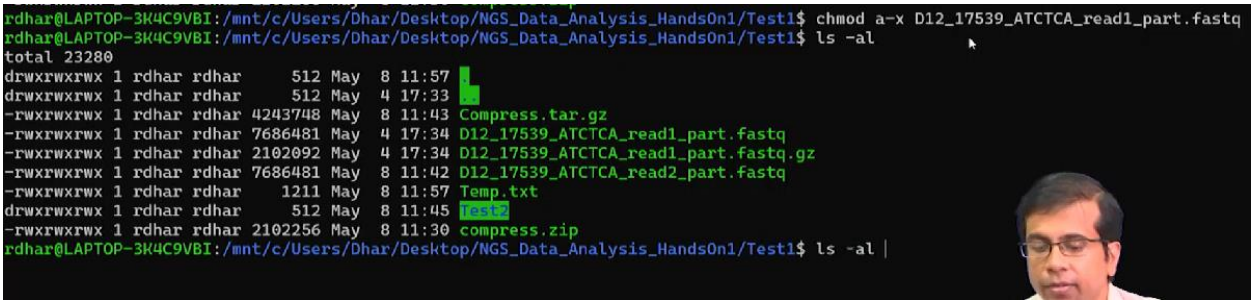


And then the third example you have chmod o minus rwx file name so others cannot read write or execute right you are taking away permission right. So this is this is how the notation look will look like right if you are using this chmod and then you can also use another way of using chmod ok. So I will discuss this bit later first let us see how these commands actually would work right and then we can run ls minus al again to see whether the permission has been changed ok. So let us do that and we will then discuss other options with chmod right there is another way to change permission in chmod ok. So let us go to the terminal now and let us discuss let us run these commands ok.

```
rdhar@LAPTOP-3K4C9VBI:/mnt/c/Users/Dhar/Desktop/NGS_Data_Analysis_HandsOn1/Test1$ ls -al
total 23280
drwxrwxrwx 1 rdhar rdhar     512 May  8 11:57 .
drwxrwxrwx 1 rdhar rdhar     512 May  4 17:33 ..
-rwxrwxrwx 1 rdhar rdhar 4243748 May  8 11:43 Compress.tar.gz
-rwxrwxrwx 1 rdhar rdhar 7686481 May  4 17:34 D12_17539_ATCTCA_read1_part.fastq
-rwxrwxrwx 1 rdhar rdhar 2102092 May  4 17:34 D12_17539_ATCTCA_read1_part.fastq.gz
-rwxrwxrwx 1 rdhar rdhar 7686481 May  8 11:42 D12_17539_ATCTCA_read2_part.fastq
-rwxrwxrwx 1 rdhar rdhar    1211 May  8 11:57 Temp.txt
drwxrwxrwx 1 rdhar rdhar     512 May  8 11:45 Test2
-rwxrwxrwx 1 rdhar rdhar 2102256 May  8 11:30 compress.zip
```

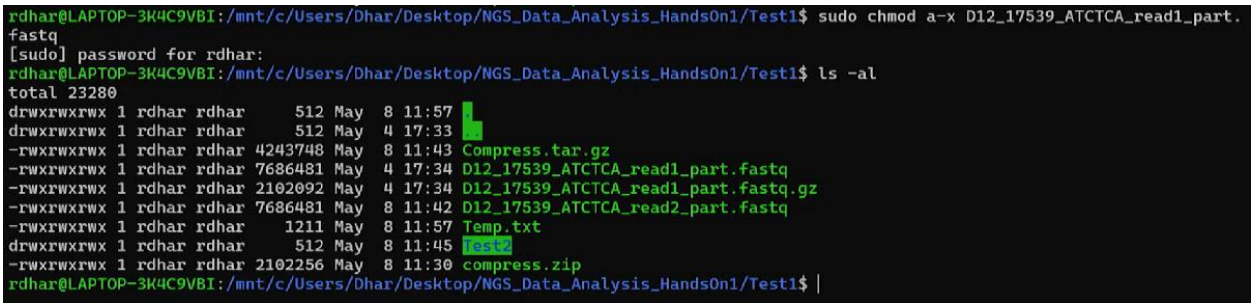
So here the files right we run this using ls minus l so let us just write ls and then we write ls minus al right this is this gives us the settings which permissions are there. So you see like ls minus simply adding this minus al it gives you a whole lot of information ok. So you have on the left side for each file right here the files on the right in green in the left you have this minus rwxrwxrwx right. So this means everyone can read, write, and execute these files ok. We have created this and this is there ok, and we can actually invoke these permissions right if we want ok.

So we will do that now using this command chmod. You also have other permissions right. So here other information, for example, here you can see, like, what is the size of these files? OK. So you have, for example, that these files are quite huge, right? These are fastq files, and the one here is the compression, right? fastq.gz, and you can see how much storage space you save by just using this easy compression, right? So let us try the chmod ok, and we will change permission. Let us say one of these files is one of these fastq files here.

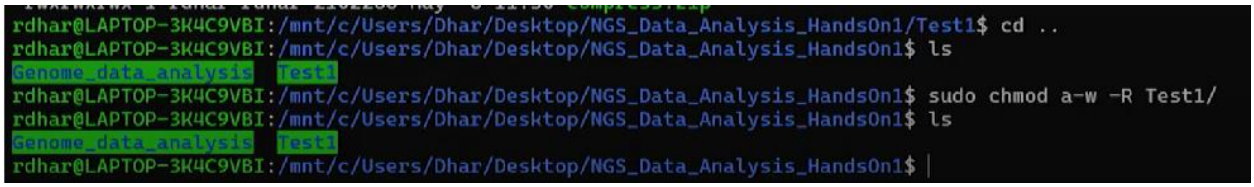


So since all permissions are there, right, so we can say a minus x right, we are removing this permission for everyone. Ok, chmod a minus x ok. So we are removing execution, right? So this is executed now. Now that this is something that will change, let us again check the permission using ls minus al right. So here I do not think actually this has happened this is probably because

the directory actually is created as executable right.

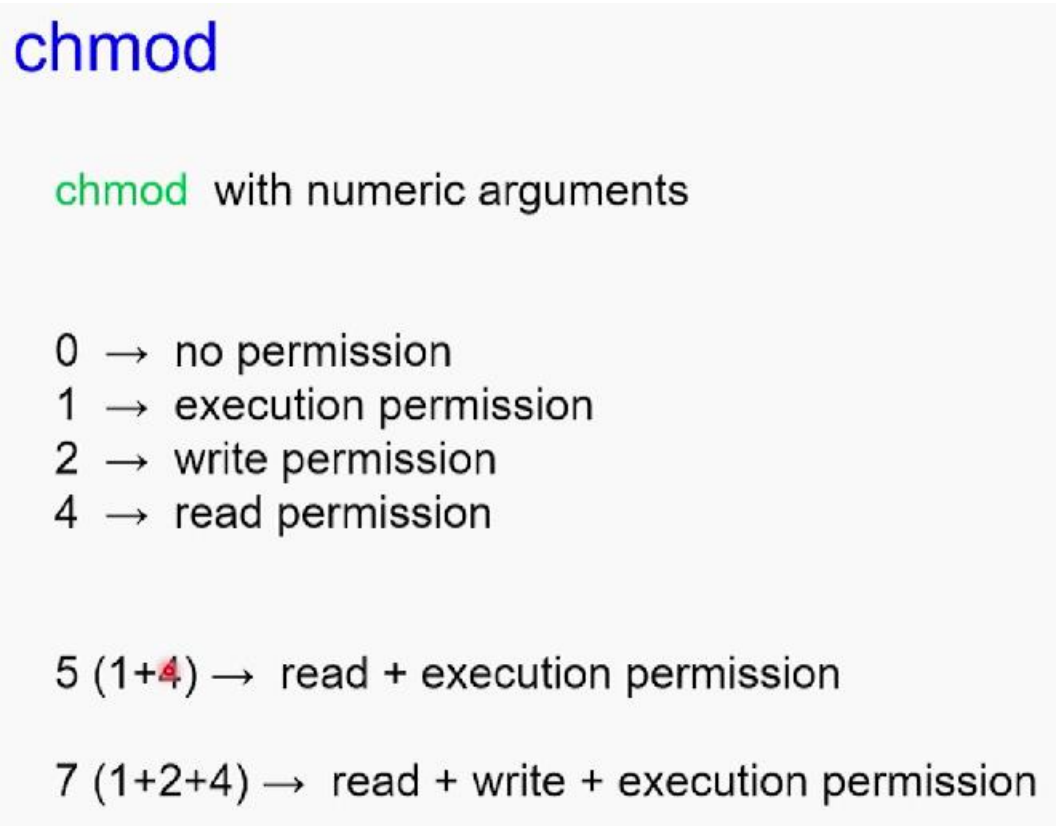


So one of the things we have probably not discussed yet is that some of the things you need to change this permission for are administrative privileges, ok? We will come to that in a moment. So you will have to have that administrative privilege to change these permissions, ok, and this privilege actually comes through sudo, ok. So the moment you use this sudo, it will actually ask for this permission right password, and that password has to be set, ok. So this is something that we use, and there is another one that we use, called chmod. If you want to use remove this permission or do this for a folder we use this chmod minus capital R ok. So how do you do this if you want to do this for a full folder ok. So this is the format ok sudo chmod a minus w minus r at is to unfold ok. So this is how it will run so unfortunately because these files were created in another system so their permissions are not changing you probably see you still see that you have all these permissions here ok. But nevertheless this is how you will actually change these permissions ok.

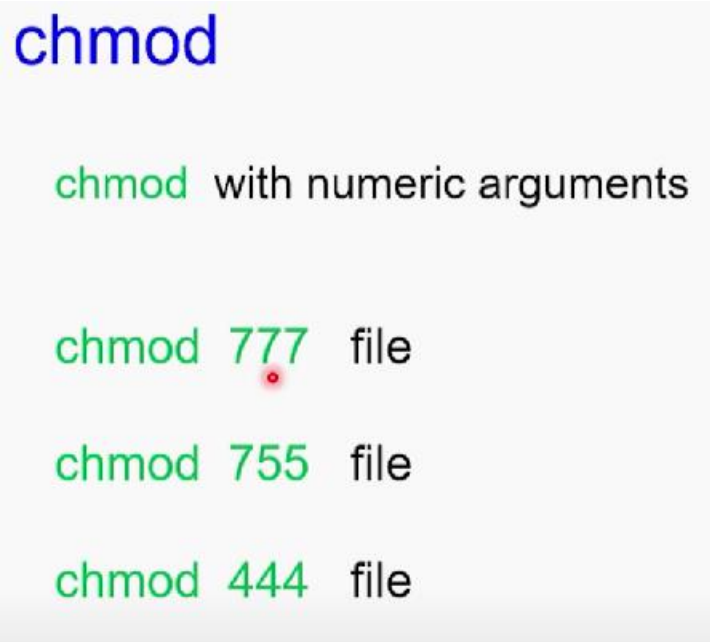


I can probably show this in a server system where it these files were created and that will probably be useful if you want ok. So maybe we will do that later ok. So these are the commands you can try yourself once you have this Linux system installed ok all right. So you can change permission in another way with chmod using something called numeric arguments ok? So here is 5, which would be then 1 plus 4; this means you have read plus execution permission, ok? So we can sum up these permissions, and we can say, well, give this read plus execution permission together, right? So if you mention 5 in the command line along with chmod, this means you are giving read and execution permission, ok. If you are using 7, this means you are giving read plus write plus execution permission, ok. So you can use these combinations right. If you say you want to give

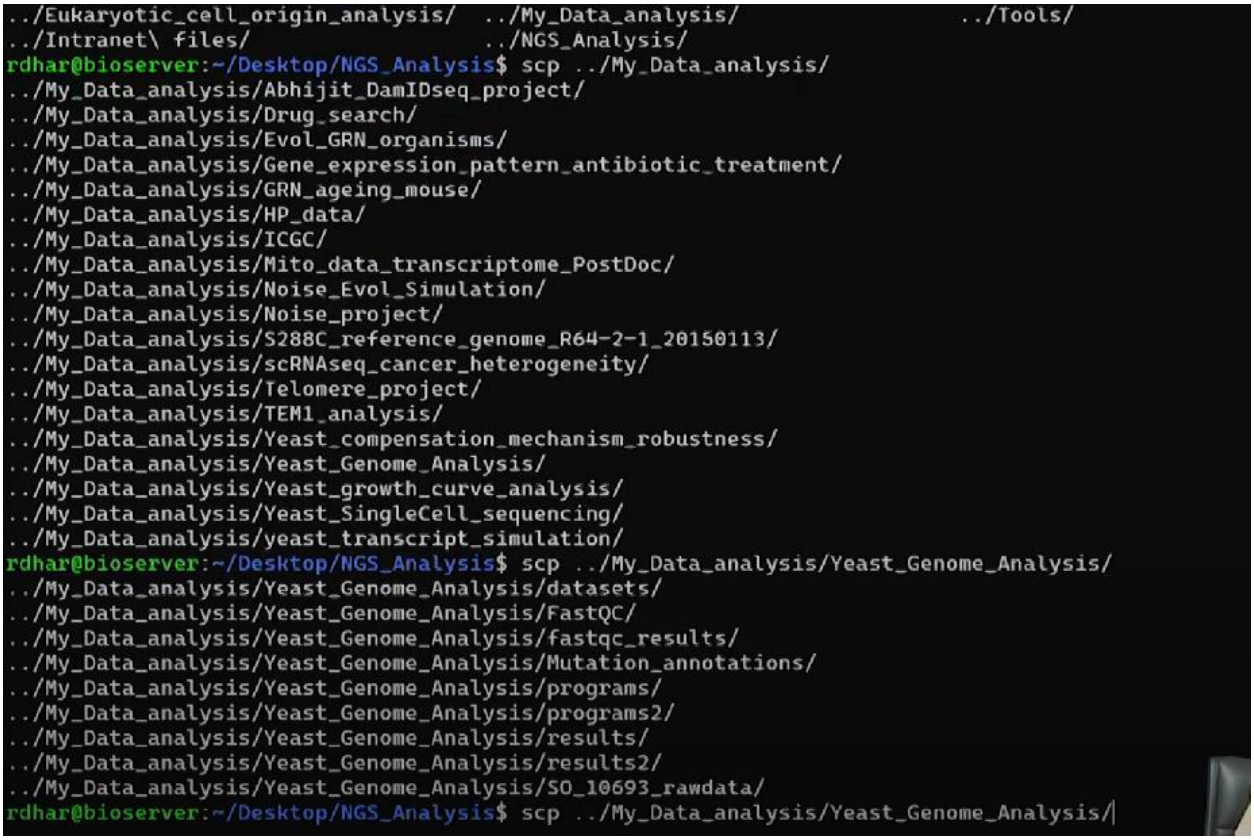
this 1, 2, write, execution, write, you can just simply set 3 right.

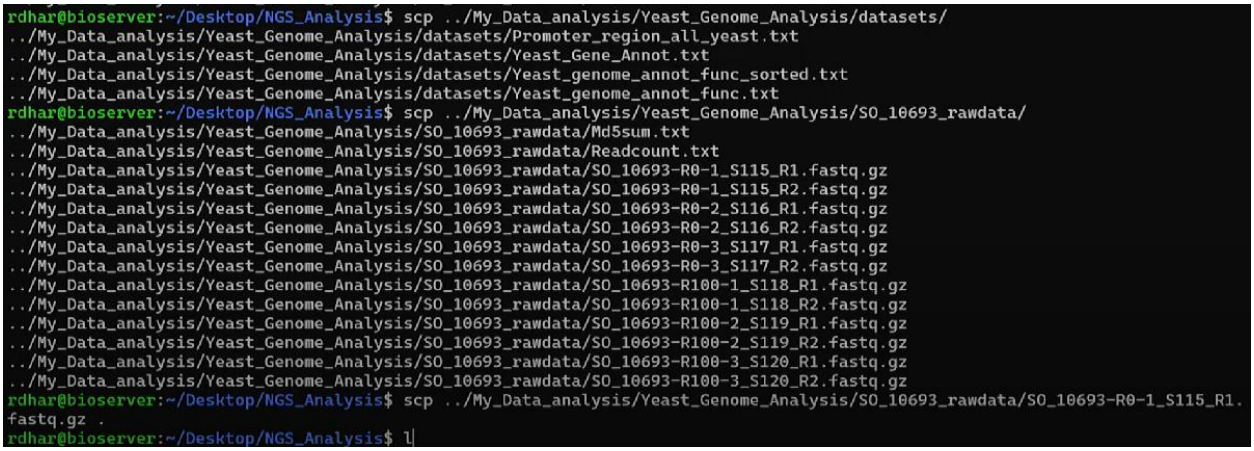


So how do you specify these again for these 3 sets of values? So it turns out you simply mention them one after another, right? So here are some examples, right? So you have chmod 777 ok. So 7 means you are giving read plus write plus execution option ok.



So the first 7 is for the owner right, the second 7 is for the group, and the third 7 is for others right. So here are these 3 sets of values, right? Like before, these are mentioned, and this way you can change permission for everyone. You are setting permission for everyone. So when you say chmod 7775, this will actually change permission for everyone, and they will get read, write, and execute permission, ok? You can say chmod 755 right. So 7 means again that for the owner, you have all permissions, but for the others and for groups, you are giving just read and write permission, ok? So you are not giving them execution permission if you remember these notations: 0, 1, 2, 4, ok. And similarly, for chmod 444, you have read permission for everyone, ok? So this is how the whole thing will work, right? So let us see if it works, and we can actually also go to a server system where we can demonstrate this better, where these files were created. So, how do you connect to a server? It is the SSH, right? So this is where this is something we discussed last time, right, and we can actually go to this folder. We can again navigate right using this cd command, right, and we can see this for example, cd desktop. So it is giving us all the NGS analysis, right? This is where we can go, and we can actually copy certain files from other places, and we can demonstrate this data, right? So for example, we can go into there is some genome data right hist genome analysis data sets we have some genome data right and from there we can actually yeah.





So we have actually the yeast genome data analysis; we have the raw data here, ok. So you can see this raw data here, and they are all FASTQ data sets, ok, and we can copy them here using this cp or scp command, and this data is here, right, and we can check if we have this permission; what is the permission status? So here you can see some of the permissions are not there, for example, the write permission. For others, we do not have write permission here. So let us try and see if this we can change. Let us say we say 777. So we are giving read, write, and execute permission to

everyone, ok, and we have run this to see if we have changed, ok. Now you see the difference here, ok. So earlier we did not see the difference because the file was created somewhere else and all permissions were given right.

```
rdhar@bioserver:~/Desktop/NGS_Analysis$ ls -al
total 458464
drwxrwxr-x  2 rdhar rdhar      4096 May  8 12:23 .
drwxr-xr-x 18 rdhar rdhar     20480 May  3 18:35 ..
-rwxrwxr-x  1 rdhar rdhar 469434468 May  8 12:23 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ chmod 777 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ ls -al
total 458464
drwxrwxr-x  2 rdhar rdhar      4096 May  8 12:23 .
drwxr-xr-x 18 rdhar rdhar     20480 May  3 18:35 ..
-rwxrwxrwx  1 rdhar rdhar 469434468 May  8 12:23 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$
```

So here you can see this file permission has been changed ok. Now let us see if we can set it back to something else ok if we can say 755 ok and see and you can see immediately right this write permissions are gone ok this write permissions are not there. Can we make it 700?

```
rdhar@bioserver:~/Desktop/NGS_Analysis$ chmod 755 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ ls -al
total 458464
drwxrwxr-x  2 rdhar rdhar      4096 May  8 12:23 .
drwxr-xr-x 18 rdhar rdhar     20480 May  3 18:35 ..
-rwxr-xr-x  1 rdhar rdhar 469434468 May  8 12:23 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ chmod 700 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ ls -al
total 458464
drwxrwxr-x  2 rdhar rdhar      4096 May  8 12:23 .
drwxr-xr-x 18 rdhar rdhar     20480 May  3 18:35 ..
-rwx------  1 rdhar rdhar 469434468 May  8 12:23 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$
```

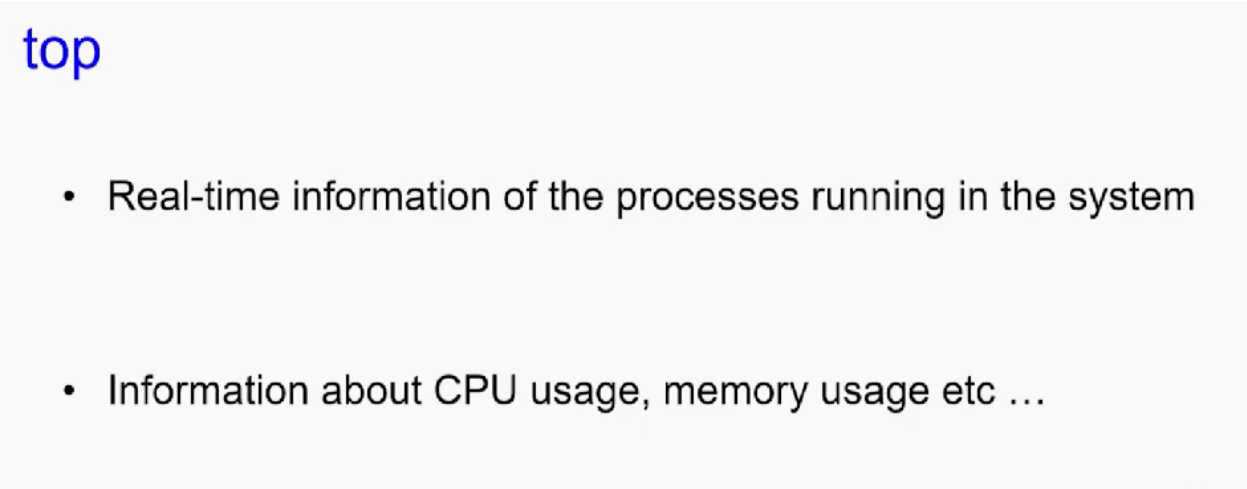
Let us say only the owner of this file will be able to read, write, or execute it, and the rest of the users may not be able to do that, ok, and you see, this permission is gone. Ok, only the owner has this read, write, and execute permission. Sometimes you want to also involve writing permission for the owner, and this is something that is sometimes advisable because you do not want to write anything by mistake. Anyway, so we have demonstrated right this chmod 777 or 755 etcetera how

they work ok.

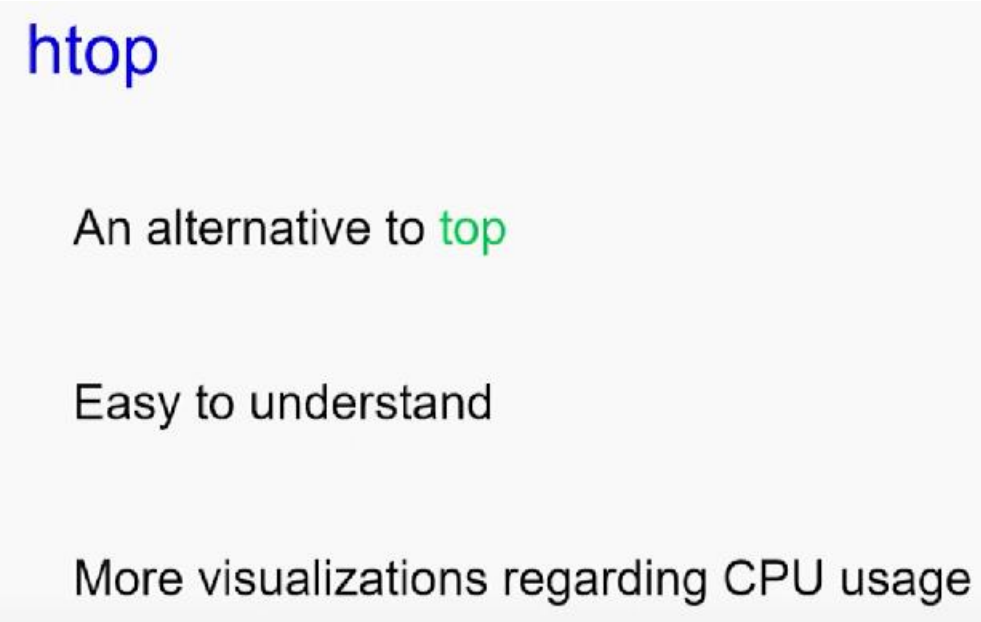
```
rdhar@bioserver:~/Desktop/NGS_Analysis$ chmod a+x SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$ ls -al
total 458464
drwxrwxr-x  2 rdhar rdhar      4096 May  8 12:23 .
drwxr-xr-x 18 rdhar rdhar     20480 May  3 18:35 ..
-rwx--x--x  1 rdhar rdhar 469434468 May  8 12:23 SO_10693-R0-1_S115_R1.fastq.gz
rdhar@bioserver:~/Desktop/NGS_Analysis$
```

We can also try this now again with this sort of notation: if you say a plus x, so we are giving execution permission to everyone. Right, a is all right, and plus means we are giving permission, and then we again check what is actually happening ok. So you see the x is there for everyone now, ok, so they can execute. So some of this is the way you can actually control the file permission, and sometimes this is useful. You will see certain software's executable files when you download and install them; they may not have permission to execute because the Linux system by default will not give them permission to execute. So if you want to set those permissions to execute right, you have to use this command: chmod. Well, you will probably use them in certain instances, ok? So that is for file permission and how we actually change these permissions.

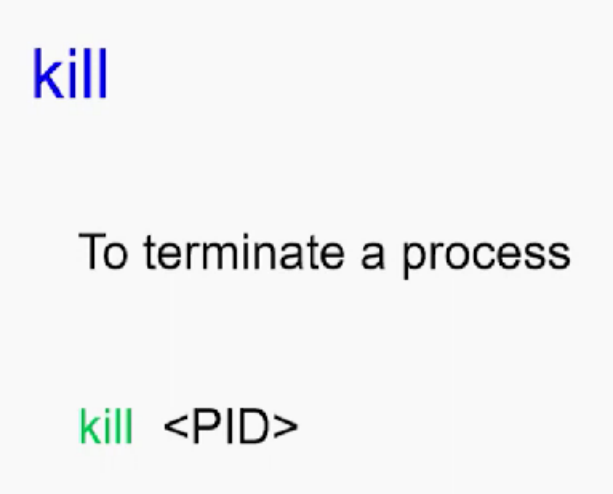
Coming to the other commands that are very useful, for example, the top command, ok. So this is a command that we use to monitor the system, so it gives us real-time information about the processes that are running in the system, and it also gives us information about CPU usage, memory usage, etcetera, ok.



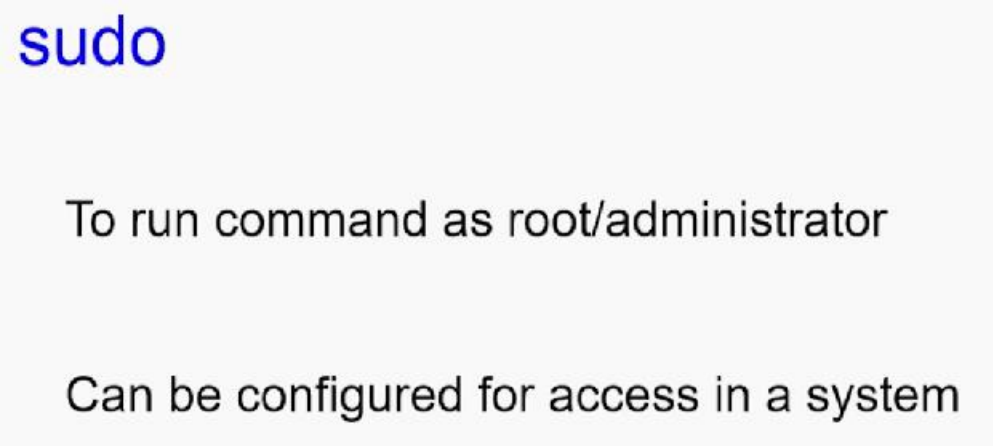
So this is something that is also useful if you are running a code or a software tool you do not want to use the maximum amount of memory, and maybe you do not want to crash the system if you are going over the limit. So this is something that is very useful to monitor if you are running some sense software or certain code. There is an alternative to top, which is called the htop.



Why do we use this? Because it is usually easier to understand and gives you a lot more visualization regarding CPU usage, such as how much CPU usage is there, etcetera. So we will see, like, how this actually looks here. Now when you use this top right or edge top, if you see some process is going to crash the system right, its memory usage is increasing, like very, very fast. You use this command kill and the PID right process ID. Now I will show you how to do this, but you have to be very careful right; you do not want to kill certain processes that are very critical. So this is something that you have to be careful about when you are using this kill command, similar to the rm command we have talked about.

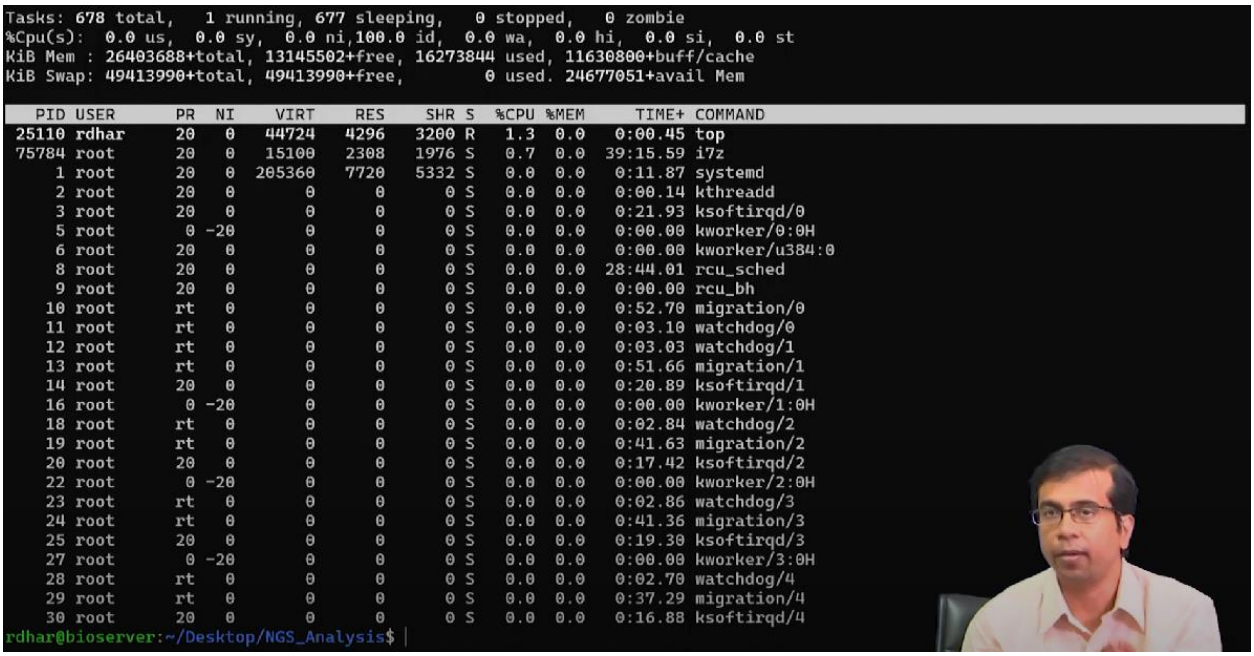


When you are removing a certain file, be very sure that you are removing the right one and you want to remove it right because there will not be any backup anymore, ok? And as I mentioned, when running some of these commands, we also need to be the root or the administrator, and we need to use this called sudo.

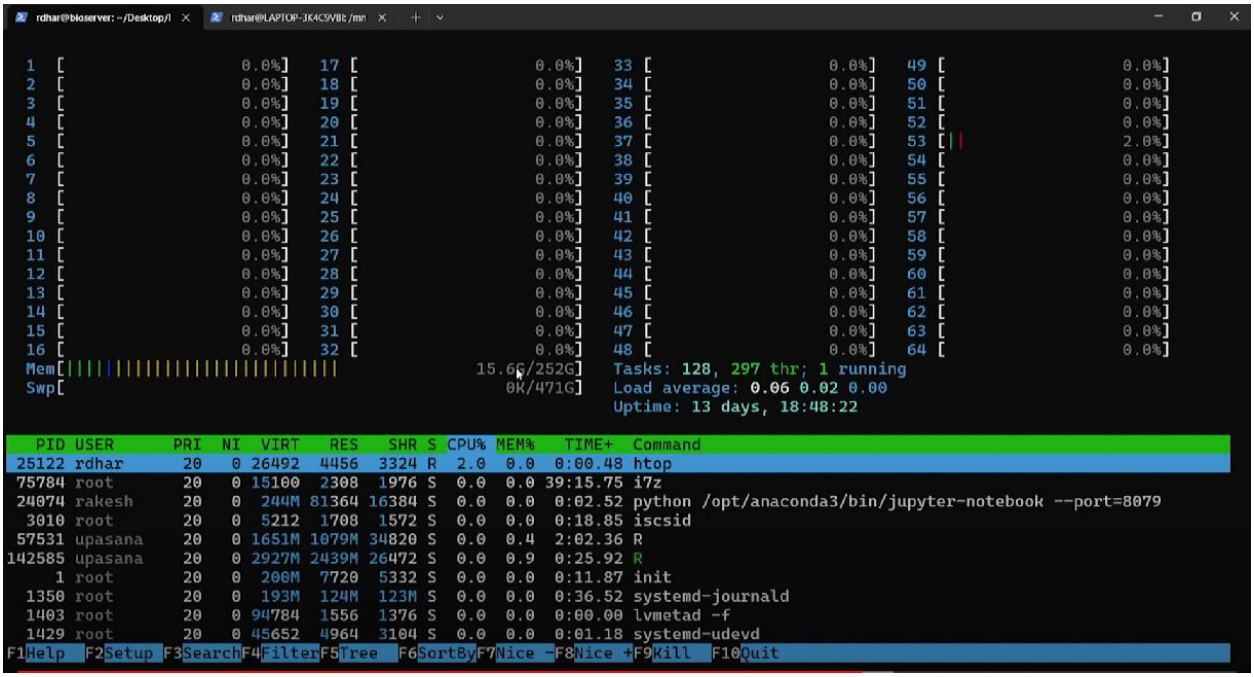


This sudo command gives you administrator privileges, ok? For example, you can change certain system files that any other users will not be able to do. Now again, this can be configured right, whether you will have sudo access or not; it is decided by the root or the administrator of the system. So the administrator can give sudo access to others or may not give sudo access to others. So this is something that we will see ok. So let us first look at these commands in the command line top and each top, etcetera, and then we will talk about the next steps ok. So since we are here on the server, we can simply write this right, we can simply write top and we can see the processes

right. So immediately, it will open a window for you, and it will tell you what processes are running. It will give you some statistics, such as what the memory usage is, etcetera. Again, these are really big numbers and sometimes very difficult to interpret, depending on how much CPU is used or how much memory each process is taking.

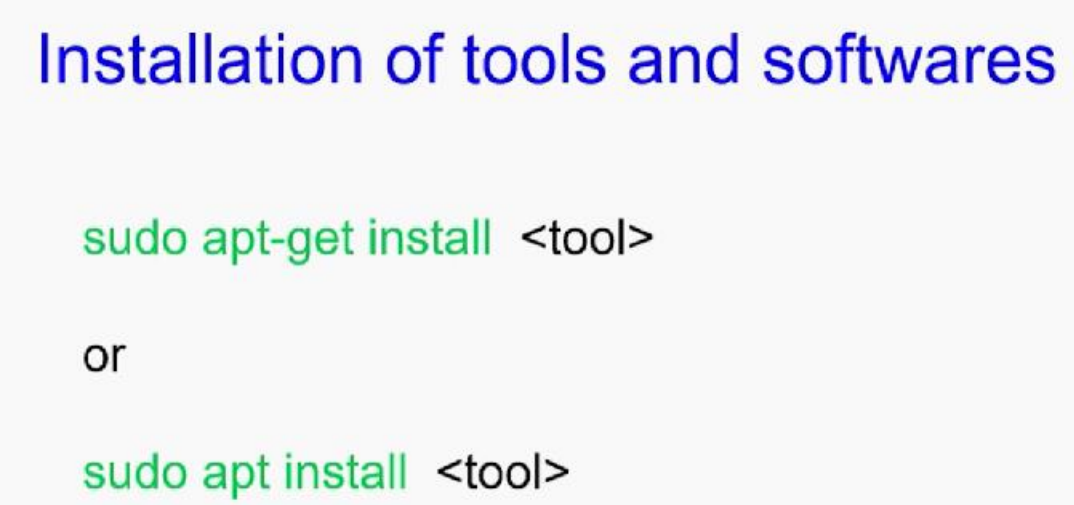


So you have this process list right, and the process ID is OK. If you want to come out, you simply type q, and you come out of that. So that is real-time monitoring of the system; if you type q, you just come out of it. You said that we also talked about each other. So we can run each other up. So one of the things I will tell you right now is that each top does not come as a default; you have to install it, ok? In a moment, I will show you how to install any tool, etcetera, on a Linux system. If you run each top, the lower part is very similar to what we have seen, but with much more detail. So here you can see the CPU usage, how many cores, etcetera are used, and what is the memory usage now in GB. So this server system has about 250 GB and current utilization is about 15.

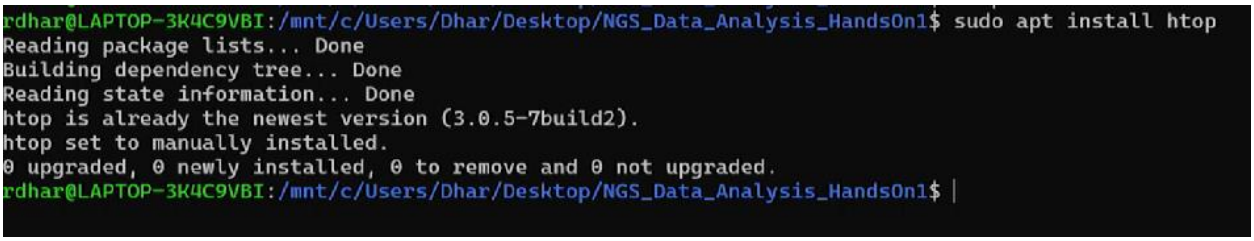


6 GB, ok, and also other information that will be here, ok. So again, to come out, we just type q and you come out of the system, right? So here you have these two tools for real-time monitoring of the system, ok. We just come out of the server and we will again work on the system, where I will show you how to install some of these tools that are not already present, ok. So kill, I did not demonstrate because we do not want to kill any processes that are running right now unless we start something ourselves, and then we can use this kill command, ok.

Coming to the installation of tools that we want to use, So these are the commands that we usually use for Linux, ok? So this is again slightly dependent on the Linux system that you are working with; it might be apt install, for example; in Ubuntu, it's apt install or apt get install; in other Linux, this could be different. So for Debian or Ubuntu, we use this apt install.



If you want to install system-wise, we need this sudo command, right? So we need to have this administrator privilege, ok? So this is something we will do now. We will install this htop tool in this system here, where this htop tool is not there, or htop is there, but we can try again. So sudo apt install htop, ok? So it is saying that because it is installed, it is not installing this htop again, right? It is just saying htop is already the newest version, OK, and it will not be installed. But if you want to install a program that is not already installed, it will show you the process, the steps, etcetera. It will give you the full output.



So it is a very simple step, right? For many tools in Linux, you just say sudo apt install htop as long as you have this sudo punch, ok? So these are some of the commands that are very useful in brief, ok? There are of course many more commands and many more utilities that you can use through the command line, but those you can use or you can learn later will be enough for NGS data analysis, ok? Now we will talk about certain text editors and data exploration. So these text editors we use for exploring NGS data sometimes ok. So for example, we have gedit ok. So gedit will not use or call nano this is another text editor and the one editor that will use is vim or vi ok. This is a very popular text editor in Linux and people use it all the time for looking at files or writing codes etcetera ok. So this is something that will come up when we talk about these tools

ok.

So let us look at nano first. So nano and vim or vi that we will be exploring we will not talk about gedit, gedit we have used earlier but now most of the people will prefer this nano or vim or vi ok. So let us explore vim or vi little bit right and or nano first and then vim or vi ok. So let us go to the terminal right and see how they actually look like ok. So let us let us go inside this folder right where we have certain files between so that we can actually open this and see ok. So to start the editor we simply say nano right followed by the file name ok.

Here is the FASTQ file right and we open that FASTQ file ok and here it is right this is the file that is opened ok. Now what it says that file read one part FASTQ dot FASTQ is unwriteable ok. So that is ok we just want to look at this and below this what you will see is the different options that you have ok and these are certain options and how you run these options ok. So here for example, we have this x right.

So you see this sign, but this actually means control. So you have to use control plus x to do this. If you want to execute, you have to do control plus x. If you want to write out right if you want to read this control r you have if you want to cut or paste you have this control k control u etc many

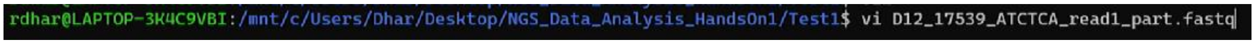
different options ok. If you want to undo right you have to m u right.

If you do, there are a lot of functions here, ok? So it gets a bit cumbersome because you have to always use this control to do things here right? So maybe we will talk about it, right? So here is the file name to write, yeah. So in which append whether you want to cancel or if you want to write, you just pressed control o right we want to write, and then it gives you certain options right.

What are the options that you want to take, which format, etcetera? You do not want to write this file, right? You just say cancel, right-click, control-C, and it will come back to the original option, right? So you will have these options at the bottom of this file, okay? This editor is giving you this option, right? So you can then actually execute it. So we do not want to write anything; we just want to go down and scroll through this file, and we can then exit using the control x command. So this is an editor that runs within the terminal, okay? We are not going out of the terminal. We are writing within the terminal; we can save, we can exit, and we can come back to the terminal. So this is very, very useful, right? You are writing within the terminal and doing all sorts of things that you can do, and you can come back and run these commands that you want to run here.

The other one is the vim, or vi right. It simply writes Vim, and it just gives you OK with what you have, and then here it will give you certain commands: right colon, q colon, help, etcetera, etcetera. So if you want to exit, you have to write a colon q here, right? So here, if we just want to exit, this

is a blank screen; there is nothing here. So we just say colon q right and for if you want to open this file we just say vim or sometimes simply we say vi ok. We do not need to then write vim d12 the file name dot first ok and again like earlier like nano it will open a file like this ok.



And we can scroll this file like before we just up arrow or down arrow and we can then again come out of this file ok. So here you will see certain statistics etcetera and then there are certain commands that will be actually very useful when you talk about the vi editor in details ok. So to come out the option is colon q right. So it mentions that if you want to come out of this file, you just say colon q, ok, and we are back to the terminal again, right? So these two editors are really useful, and we will talk about one of them in the next class, which is Vim. But it turns out Vim is really powerful; it has a lot of options that you can use within the editor, ok? So these are the agenda items, right? So in the next class, we will talk about this data download, and we will explore this data using the vi editor. This is again the reference that we have used, and to conclude, we have talked about the shell commands for file permission control. And these are very useful when we are downloading software tools that might not have the necessary permission to run on your system.

So you want to give those permissions before you can execute them, ok? So we use this command

ls minus l to check the permission, and then we use chmod to change permission. There are two ways of changing permission we talked about, right? We can go about a plus x or a plus root, or we can use simply the number system, which is the 777 or 755 system, which can change permission for all users at once. So this is what we have talked about, and then we also talked about certain commands for system monitoring, for example, top or each top, which will allow us to monitor the system right for system usage. For example, CPU usage if you are running some program, how much memory is utilized, etc., and finally, you can also control right; you can kill processes if you see they are going to go over memory and they might crash the system. If your memory usage is around 90 percent to 95 percent from a single program, you might want to stop that program, ok? And we introduced the text editors in the next very briefly, which we will discuss a lot more in the next class. Thank you.