Week 1 FAQs

[Cloudera VM](#)

[Hadoop Architecture](#)

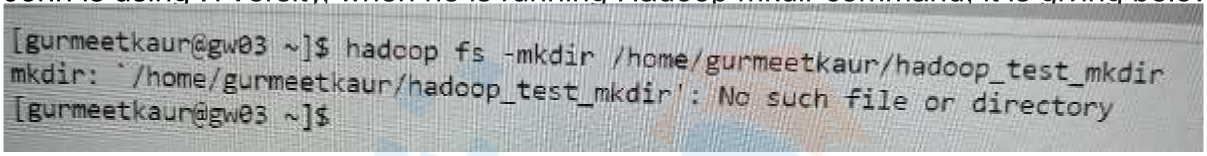
[Linux Commands](#)

[Hadoop Commands](#)

Cloudera VM

1. How to add ovf file to VirtualBox?
   Ans: Open the VirtualBox manager →Click on File & select Import Appliance option. →Browse for the file and select the .OVF file → click next

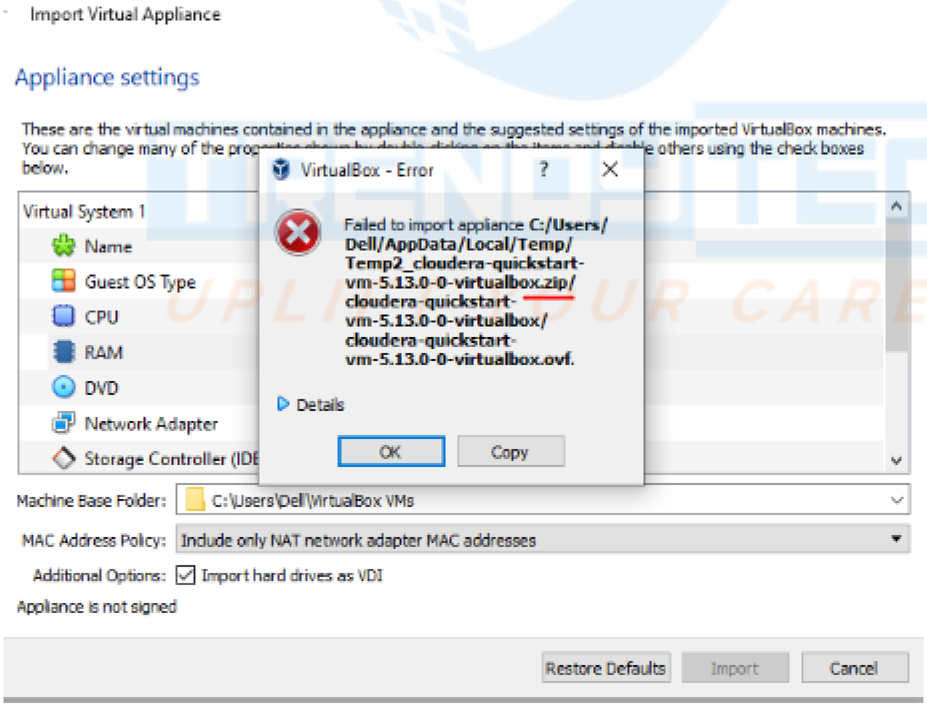
2. John is using ITVersity, when he is running Hadoop mkdir command, it is giving below error.



```
[gurmeetkaur@gw03 ~]$ hadoop fs -mkdir /home/gurmeetkaur/hadoop_test_mkdir
mkdir: `/home/gurmeetkaur/hadoop_test_mkdir': No such file or directory
[gurmeetkaur@gw03 ~]$
```

   Ans: try Hadoop fs -mkdir /user/gurmeetkaur/hadoop_test
   You cannot create directory in home/UserName

3. John is getting below error while importing ovf file?



Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the pro e others using the check boxes below.

Virtual System 1
- Name
- Guest OS Type
- CPU
- RAM
- DVD
- Network Adapter
- Storage Controller (IDE

VirtualBox - Error

Failed to import appliance **C:/Users/Dell/AppData/Local/Temp/Temp2_cloudera-quickstart-vm-5.13.0-0-virtualbox.zip/cloudera-quickstart-vm-5.13.0-0-virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.ovf.**

Machine Base Folder: C:\Users\Dell\VirtualBox VMs

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

   Ans: First Unzip the file and then import it

4. After installing Cloudera VM. It’s showing in a small window and not full screen. Is there any setting I need to do to open the full screen?

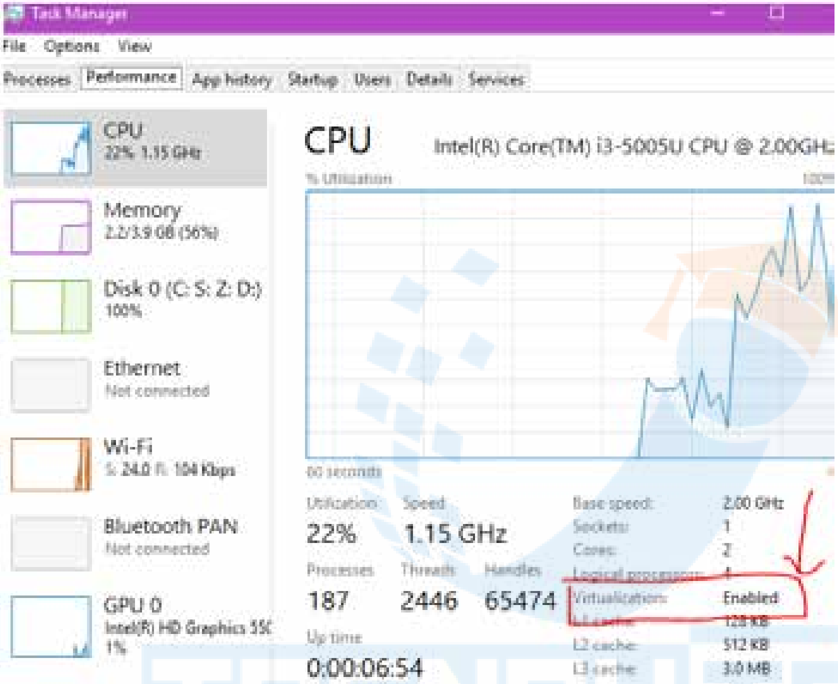
   Ans: View -> Switch to Fullscreen Mode (or HOST+F)

   Right CTRL key is normally the default HOST key.

5. When I Powered off the Cloudera Quickstart VM and started it again the shared folder has been removed. what can be the solution for that?? I need to run - mount -t vboxsf Shared folder command again to create it.

   Ans: Make sure you have checked “Make Permanent “ Checkbox while creating shared folder

6. How to Check Virtualization is enabled or not?
   Ans: Goto Task Manager → Performance tab

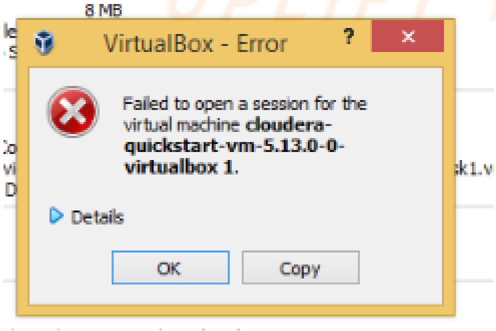

Task Manager
File Options View
Processes Performance App history Startup Users Details Services
CPU 22% 1.15 GHz
Memory 2.2/3.9 GB (56%)
Disk 0 (C: S: Z: D:) 100%
Ethernet Not connected
Wi-Fi S: 24.0 R: 104 Kbps
Bluetooth PAN Not connected
GPU 0 Intel(R) HD Graphics 55(1%
% Utilization
100%
60 seconds

CPU Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz

Utilization 22%
Speed 1.15 GHz
Processes 187
Threads 2446
Handles 65474
Up time 0:00:06:54

Base speed: 2.00 GHz
Sockets: 1
Cores: 2
Logical processors: 4
Virtualization: Enabled
L1 cache: 128 KB
L2 cache: 512 KB
L3 cache: 3.0 MB

7. While installing Cloudera VM, I am getting an error as Failed to open a session for virtual machine. In details it’s written - not in hypervisor partition and AMD -V is disabled in BIOS (or host OS)



Failed to open a session for the virtual machine **cloudera-quickstart-vm-5.13.0-0-virtualbox 1**.

Ans: It appears that virtualization is not enabled in your Windows computer. Some Windows computers have the virtualization disabled by default at the BIOS level and it needs to be enabled to set up a new virtual machine.

1.Close the virtual box

2. To open Bios follow this link to find the key used to enter BIOS

https://www.lifewire.com/bios-setup-utility-access-keys-for-popular-computer-systems-2624463

3. Follow below steps

- Turn ON the System.
- Press F2(Step 2 will help you identify this key for your machine) key at startup BIOS Setup.
- Press the right arrow key to select Advanced tab → Select Virtualization→ press the Enter key.
- Select Enabled →press the Enter key.
- Press the F10 key→ select Yes → press the Enter key to save changes→ Reboot into Windows.

Note: If your Laptop is very old then it might not support virtualization.

8. I am unable to download virtual box and vm. I didn’t find a link to download from the pdf given.

   Ans: Scroll down under the video and it has both the links.

   1. Oracle VirtualBox Download: https://www.virtualbox.org/wiki/Downloads
   2. Cloudera quickstart VM Download: http://shorturl.at/aMW56

9. I am facing issues importing Cloudera quickstart VM to Oracle Virtual Box. On clicking Import it shows progress for some time and then gives error:

   Failed to import appliance

   /home/rsi/cloudera-quickstart-vm-5.13.0-0-virtualbox/cloudera-quickstart-vm-5.13.0-0-v irtualbox.ovf.

   Result Code: NS_ERROR_INVALID_ARG (0x80070057)

   My host machine is Ubuntu 16.04. If anyone else is trying to install on Ubuntu machine and is able to import successfully.

   Ans: Open your virtual box →unzip VM → import a new VM→ select the ovf

10. As I ‘ m working in a cloudxlab I do not know how to copy files from local to directory. On doing it is throwing an error on the gedit command?

    Ans: - gedit won’t work in Cloudxlab.

    You have to create a File/Folder using Hue.

    also remember you have to create it inside your home Directory.

11. My cloudera installation does not have /user/cloudera
    Ans: You can see /user/cloudera in HDFS. For local file system it is /home/cloudera
12. John is facing an issue with the download of Cloudera Quickstart VM. His network connection is good however download is failing with network error. Please let me know what the issue is?

    Ans: You need a good stable internet connection as the file size is too large (5.7GB). Check your internet download speed it should be in MB/s. I will not recommend to use a mobile hot spot as it is more prone to fail. Better try in night when the network traffic is less. The best way to download is to use a broadband cable connection.

13. I am unable to access www.google.com despite enabling System eth0. I disconnected my VM from the internet and connected back using System eth0, but can't access google.com. Is it a dns issue?

Ans: Restart the VM and check it again.

14. John have deleted the 'data' folder on cloudxlab where sample files are kept by cloudx team. How to get them back? I am trying to change the replication factor dynamically for a file. Sumit sir has shown demo on a file which is already made available by cloud x. So, how to get it back?

    Ans: You have to be very careful in these things because all users will get affected by this.

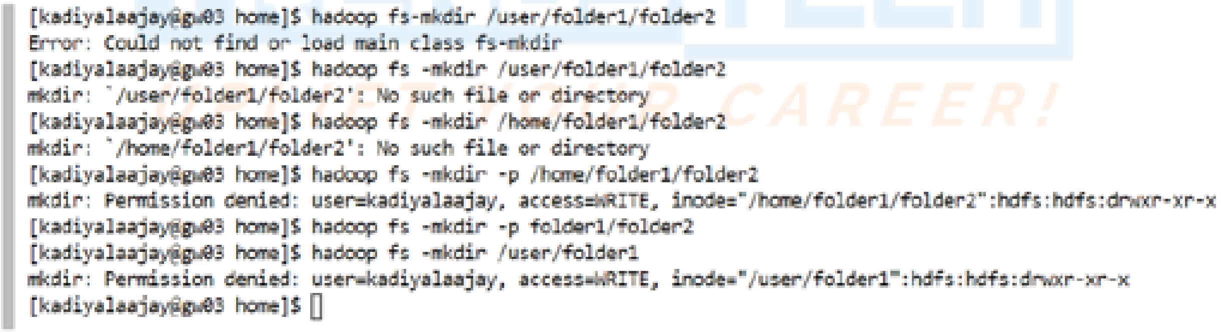
    I think you will not be able to get it back. Only the Cloudxlab team admin can do it.

15. Which env (cloudera or cloudxlab) is used in this course?

    Ans: If your laptop has minimum 8 GB Ram then choose Cloudera env, if not then go for Cloudxlab or itversity lab.

16. How to practice using Itversity labs
    Ans: You can connect to itversity using putty or MobaXterm.
    Reference link: https://www.youtube.com/watch?v=rUSkqrqUU6Q
17. How to disable Hyper V?
    Ans: Goto Control Panel→Programs→Turn windows feature ON or OFF→ disable Hyper V and Windows Hypervisor platform

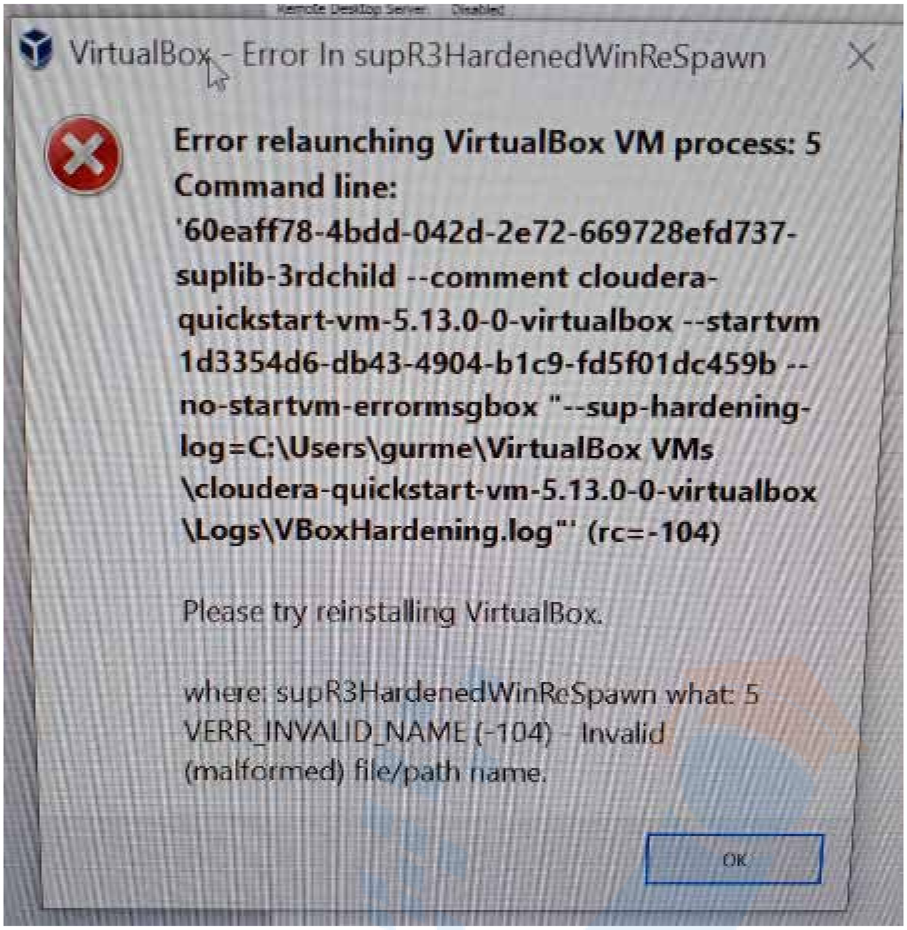
18. John is unable to create hierarchy of directories in itversity, single directory in terminal, as it throws permission error.



```
[kadiyalaajay@gw03 home]$ hadoop fs-mkdir /user/folder1/folder2
Error: Could not find or load main class fs-mkdir
[kadiyalaajay@gw03 home]$ hadoop fs -mkdir /user/folder1/folder2
mkdir: `/user/folder1/folder2': No such file or directory
[kadiyalaajay@gw03 home]$ hadoop fs -mkdir /home/folder1/folder2
mkdir: `/home/folder1/folder2': No such file or directory
[kadiyalaajay@gw03 home]$ hadoop fs -mkdir -p /home/folder1/folder2
mkdir: Permission denied: user=kadiyalaajay, access=WRITE, inode="/home/folder1/folder2":hdfs:hdfs:drwxr-xr-x
[kadiyalaajay@gw03 home]$ hadoop fs -mkdir -p folder1/folder2
[kadiyalaajay@gw03 home]$ hadoop fs -mkdir /user/folder1
mkdir: Permission denied: user=kadiyalaajay, access=WRITE, inode="/user/folder1":hdfs:hdfs:drwxr-xr-x
[kadiyalaajay@gw03 home]$ []
```

    Ans: Try this command
    Hadoop fs -ls /user/kadiyalaajay
    Hadoop fs -mkdir /user/kadiyalaajay/folder1
    It should work

    In itversity you have access only /user/kadiyalaajay hence work only in this path, root is not accessible to user
19. Help John to solve below error

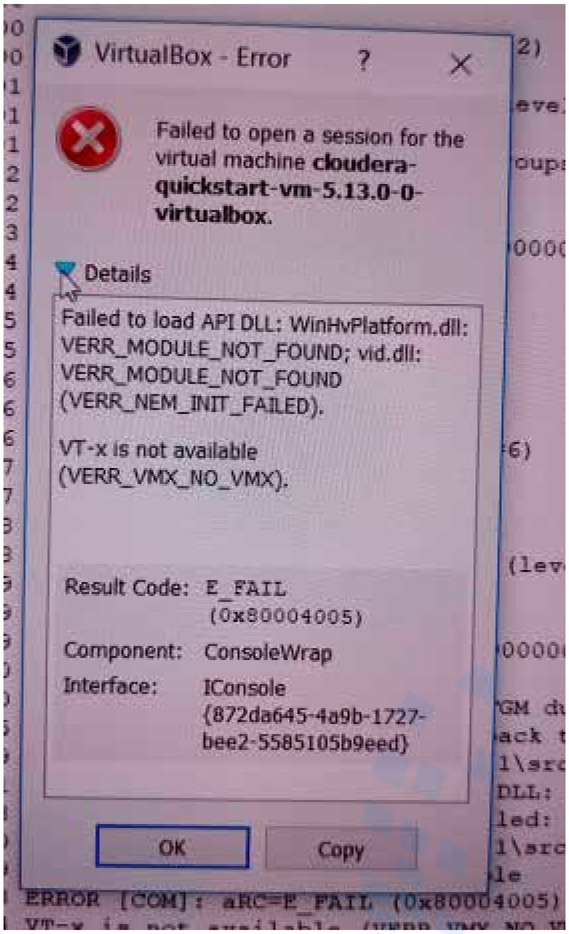


Remote Desktop Server Disabled
VirtualBox - Error In supR3HardenedWinReSpawn
OK

```
Error relaunching VirtualBox VM process: 5
Command line:
'60eaff78-4bdd-042d-2e72-669728efd737-
suplib-3rdchild --comment cloudera-
quickstart-vm-5.13.0-0-virtualbox --startvm
1d3354d6-db43-4904-b1c9-fd5f01dc459b --
no-startvm-errormsgbox "--sup-hardening-
log=C:\Users\gurme\VirtualBox VMs
\cloudera-quickstart-vm-5.13.0-0-virtualbox
\Logs\VBoxHardening.log"' (rc=-104)

Please try reinstalling VirtualBox.

where: supR3HardenedWinReSpawn what: 5
VERR_INVALID_NAME (-104) - Invalid
(malformed) file/path name.
```

Ans: This issue generally comes if you will not start the machine after installation and you re-started the computer. Now the solution is you have to completely uninstall the virtual box and start from the scratch.

20. Help John to solve below error



Failed to open a session for the virtual machine **cloudera-quickstart-vm-5.13.0-0-virtualbox.**

Details

Failed to load API DLL: WinHvPlatform.dll: VERR_MODULE_NOT_FOUND; vid.dll: VERR_MODULE_NOT_FOUND (VERR_NEM_INIT_FAILED).

VT-x is not available (VERR_VMX_NO_VMX).

Result Code: E_FAIL (0x80004005)
Component: ConsoleWrap
Interface: IConsole {872da645-4a9b-1727-bee2-5585105b9eed}

```
ERROR [COM]: aRC=E_FAIL (0x80004005)
```

Ans: Check the following steps

1. Check Virtualization is enabled
2. Disable Hyper V
3. Reference link: Solved: VT-x is not available (VERR_VMX_NO_VMX) Error for ... - Cloudera Community
4. If all above does not work then try importing new ovf file

Hadoop Architecture

21. What would be the maximum and the minimum limit to the block size which can be configured?
Ans- There is no as such limit set by Hadoop to bound the user having a certain block size. Hadoop has fixed reasonable block size considering Big data situations hence there should not be a need to minimize the size. Increasing size depends solely on your data

22. What are edge nodes and can it be configured in the configuration file?
Ans: Yes, it can be configured. Generally, the cluster will not be accessible/open to all. So here edge nodes act as an interface/gateway between the cluster and outside network. It has various clients installed and various configurations to connect to different systems like HDFS, hive in cluster. It provides a safer way to connect to clusters.
This is a job of administrator.

23. In HDFS Architecture. The file of 500Mb is divided into blocks of 128MB. The last block gets

116Mb. As explained the space from the last block does not get wasted. Does it mean that when we take another file for processing it would occupy the remaining space first from the last block.

Ans- Yes, the last data block of the file can be of varying size. Remaining space is freed up.

NO, second file will occupy new data blocks as per requirement.

Note: One data block cannot be shared by two files.

24. I have a doubt in HDFS architecture. Consider our file size is 1GB and we have 4 nodes. Then how the file is divided into blocks with how much memory size??
Ans: Considering 1GB file, it will be divided into 8 blocks (considering the default block size as 128MB) and with the replication factor 3, A total number of blocks will be 8*3 = 24 blocks. These 24 blocks will be distributed among 4 nodes.

25. What is the difference between DataNode being Corrupt and DataNode being dead? I see DataNode is corrupt we can know by BlockReport and DataNode is dead we can know by Heartbeat. But is there any difference?
Ans: Yes, when you say DataNode is corrupt, it means the data stored as a block in the node is corrupted and you have to reload the data. Whereas DataNode is dead if it is not communicating back to NameNode which happens if the DataNode service running on the node is stopped or there is any network issue

26. What happens if edit log is corrupted, how it handled? Is there any scenario where edit log is deleted or missing, how to retrieve or debug to get the same?

Ans: If admin as configured copies of edit log then it can be easily recovery. Another option is manual recovery

```
./bin/Hadoop NameNode -recover
```

This is where HA comes in picture where multiple standby NameNodes are maintained

27. The data file gets divided in the form of 128 MB blocks. Who does the work of dividing the data into these blocks?

Ans: Framework will take care of block division and storing their meta data in the NameNode. By default, it is 128. It can be configured in HDFS-site.xml.

28. We have 3 blocks in a data node, if one block fails how name node come to know? and is it replicate that block again to maintain the default replication (3)

Ans: Using heartbeat you can identify dead DN... By default, every 3 seconds Data nodes will send its information to Name Node. If Name Node does not receive heart beats from any data nodes for a certain period then it will consider that data node is dead.

Once, a data node is dead. Name node will try to create the replicas of the blocks in dead data node in other data nodes.

Ans 2: If a block in the data node is dead, Then NN gets to know this through the block report.

Block Report: Each DN sends a block report to the name node at a fixed frequency indicating if any blocks are corrupted.

29. If the meta data created by name node is in memory, then where is it permanently stored?

Ans: Name node stores a snapshot of entire metadata states into the local disk as fsimage file, whenever we start NN this copy will be brought into main memory, this helps in overcoming metadata lost problem.

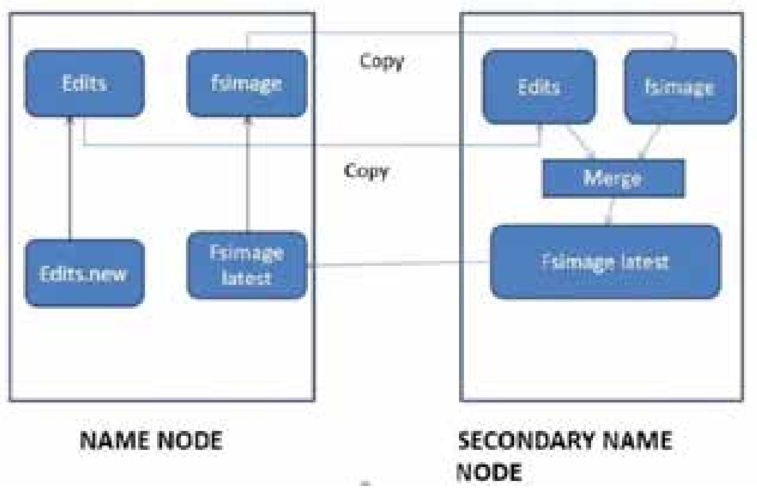
30. In the check point mechanism, the initial FS image is provided by name node then subsequently the updated FS image from secondary name node overwrites. So, does the name node only provide FS image only once during the first time?

Ans: Yes, NameNode provides the initial fsimage only once at the start of the NameNode. And the updates are merged by the secondary name node, which then overwrites this initial fsimage.

31. Why is Hadoop not considered a good fit for large no. of small files?
Ans:
1. A small file is the one which is significantly smaller than the HDFS block size (default 128MB)
If you have a million of files each requiring 1 block, leading to millions of entries to the NameNode causes a potential memory overhead on the NameNode. Considering a 20 million smaller files each of which occupies 150 bytes would use 6GB of memory.
2. Smaller files also degrade the performance of MapReduce processing. If we have 10k files each containing 10 MB of data, a MapReduce job will schedule 10k map tasks which will have the overhead of spinning up and tearing down 10k JVMs

32. Where I can find the edit logs?
Ans: It's stored by NameNode's directory configured in dfs.NameNode.edits.dir property

33. On HDFS Architecture side:
Active node writing FSimage and Edit logs in shared folder and Secondary Name Node (SNN) merge those and create New FSImage.
My question is, every 30 seconds those 2 files write in shared folder by Active node and while writing is that merging (Reading and create new file) can be possible same time by SNN? Or SNN needs to wait till writing part is complete?
What happen if writing part takes more than 30 sec?
Ans: NameNode instead of copying to edit log, it uses edit.new till SNN informs name node that it has merged the data and the all the edit and edit.new is combined for the next merge.
Below pic can help you understand



34. In Video, it explained what happens when NameNode or DataNode fails but What happens

when NameNode and DataNode is up then how they regain the functionality
Ans: If NameNode fails SNN is made active, when NameNode is up it can act as SNN. With new HA NameNode can be added as standby NameNode.

When DataNode is up, it acts as a new node, available for new data. If previous data is retained then that is erased because already 3 copies of replication is maintained.

35. When you create a new directory on HDFS, where will it be created? On all data nodes or just one DataNode?
Ans: It will be created on all DataNode as they all have same directory structure but all DataNodes might not have data. NameNode knows where the data is and accordingly the particular DataNode is connected.

36. We know that in Hadoop 2, each block size is 128 Mb. And there will be data nodes that store data in the form of blocks and we have multiple copies of each block in case of data node failure. Now, what if we reach the maximum limit of the all the data node? Will we add new nodes and then name nodes will configure it accordingly?
Ans: Hadoop is horizontally scalable, you can have any number of nodes added to it at any point of time depending on the volume of your data.
Coming to data node failure, If any of the DataNodes goes down, NameNode creates another copy of all the blocks in the available data nodes to maintain the replication factor.
Refer Q no 57 answer for more detail

37. I can see a lot of properties are present in HDFS-site.xml under the cloudxlab (cloud), but in the cloudera quickstart VM (local) very few properties are present/configured. For ex, “df.block size”, “dfs.replication” & etc. Did I miss anything?

Ans -cloudxlab uses Hortonworks and we are using Cloudera VM. Cloudera and Hortonworks are two different distribution of Hadoop. If you can use it practically then no need to worry about these files.

Also, these properties file comes under administration part.

38. What is a Hadoop file system and how is it different from a local file system?
Ans:
1. HDFS is a distributed file system. Local file system is not.
2. HDFS have a replication factor . Local file system is does not have replication.
3. HDFS has large Block Size, Local file system has small balck size which cause multiple seeks to read large file
4. HDFS is master-slave architecture. LFS uses tree format to store data.

39. Will there be both standby NN and Secondary NN in every HDFS architecture?
Ans: No - either Secondary NN is enabled or Standby NameNode is enabled. In Hadoop 1.x & 2.x SNN is used but after Hadoop 2.x with HA Standby NameNode is used.
40. Explain High latency and High Throughput?

Ans:

High throughput:
Throughput is the amount of work done in a unit time.

- In Hadoop, the task is divided among different blocks, the processing is done parallel and independent to each other. So, because of parallel processing, HDFS has high throughput.

- The HDFS is based on Write Once and Read Many Model, it simplifies the data coherency issues as the data is written once can't be modified and therefore, provides high throughput data access.

- Apache Hadoop works on Data Locality principle which states that move computation to data instead of data to computation, this reduces network congestion and therefore, enhances the overall system throughput.

High latency:- Latency is the time that passes between a user action and the resulting response.
since the request first goes to NameNode and then goes to DataNodes, there is a delay in getting the first byte of data. Therefore, there is high latency in accessing data from HDFS.

41. Explain the Write once and read many model?
Ans: "Write once and read many" is a term we use for the systems those are not efficient for writing / updating (or updating not available), it means once we have written some data we are not going to modify it. but we are going to retrieve / read it many times for processing or doing analytical work on it.
42. Difference between Heartbeat and Block Report in HDFS?
Ans: Heartbeat message is sent by DataNode to NameNode saying it is active it is in stable state (No hardware failure). dfs.heartbeat.interval (in HDFS-site.xml). By default, this is set to 3 seconds.
Block Report: Block reports says that the data stored in that DataNode is not corrupted. dfs.blockreport.intervalMsec (in HDFS-site.xml). By default, this is set to 21600000 milliseconds.
43. What If Name Node and Secondary Name Node fails at the same time, how it will be tackled, I understand that the shared folder will be having the latest image but how fast Hadoop will identify the third node and make it as Name node?

    Ans:First of all, NameNode and Secondary NN are not cheap hardware. These are high-end machines and are less prone to failure. There are very rare chances that both the high-end devices fail together. Also note that SNN is always passive which makes it even less prone to failure.

    If incase both the NN and SNN fails simultaneously (in case of natural calamities) then Hadoop admin manually restart the cluster by providing a backup node.

    However, in Hadoop 3.0, it allows the user to run many standby NameNodes. For example, configuring five JournalNodes and three NameNode. As a result Hadoop cluster is able to tolerate the failure of two name nodes rather than one.

44. In QJM does only the edit log get shared in JNs or fsimage as well? And in checkpointing we saw the new fsimage getting stored again in the shared folder, does the same happen in QJM? Does the new fsimage again get stored in QJM? If yes, does it get shared in all JNs?

    Ans: Yes, in QJM only edit logs are shared in JNs. FSimage is always available with both active and standby NN - so no need for any shared location.

45. Is it possible to set up fully distributed mode by installing each software and tools on every node?

    Ans 1: Not required: Fully distributed mode is for production environments. It is not possible in a single system.

Ans 2: Yes, You can try if you have n(4) machines connected through LAN and configure it yourself.

46. John knows horizontal scalability can be achieved using Hadoop. But at the same time, vertical scalability can also be achieved by increasing the resources of the nodes. So, on the generic note can we say both horizontal and vertical scalability can be achieved through Hadoop. If not please explain why?

Ans: Yes, correct you can achieve both if you want too but vertical scaling doesn't make any sense in Hadoop cluster.

- Vertically scaling needs downtime, greater risk of hardware failure and is tedious.
- With horizontal-scaling it is often easier to scale dynamically by adding more machines into the existing pool. Also, you can downscale it whenever required by removing machines which is not the case in case of vertical scaling.

47. What is the difference between Secondary NameNode and Standby NameNode? Why is the QJM standby NameNode used instead of the secondary NameNode?
Ans: SNN can't replace the NN automatically on NN's failure. But Standby NN automatically takes the control and becomes active. So avoiding Single Point Of Failure.

48. What are the situations when a node gets fail?

Ans:

1.Network issue between DataNode and NameNode.
2.Power failure
3.JVM processes running on DataNode failed.
4.Reboot of system.
5. etc….

49. What if for a 4 node cluster 2 DN goes down at same time then how the replication will happen in that situation?
Ans: In real time, we never run 4 node clusters. If we consider this situation then remaining 2 DN will store 3 copies of each block (default replication :3)

50. Is heartbeat time configurable? Also is it configurable how many times it tries for a dead or slow data node?
Ans:Yes, you can. Refer below option in HDFS-default.xml

1. dfs.heartbeat.interval : To set heartbeat

2. dfs.NameNode.heartbeat.recheck-interval : decides the interval to check for expired DataNodes

51. Is there any chance to see the metadata which is available in the name node?

Ans: Yes, you can but you can't read it directly as they're in binary format. You have to convert it to XML or txt format. This comes under admins profile.

52. I have 8GB RAM on my laptop Install the Oracle VM default configuration: 4GB RAM, CPU core 1 but run a simple Hadoop command(ls) it will take 1-2 min to get the output, what is the reason behind this? Do I need to increase the CPU core from 1 to 2?

Ans: It depends on which processor you have, background processes running simultaneously. Make it 2 cores and see if it works well.

53. If one of the data nodes goes down then the data is replicated in the remaining nodes to maintain the default replication. But what happens when the faulty data node is up and running? Or the name node doesn't expect anymore heartbeat once it marks it as faulty and uses remaining nodes only.

    Ans: Faulty nodes will be acting like a fresh machine and all data will be wiped up. Refer Q no 57 answer for more details

54. If replication can be changed using the below command then does that mean each file can have a different replication? *HDFS dfs -setrep [-R] [-w] <numReplicas> <path>*

    Ans:Yes, it is possible. The default replication size is 3 copies, all files stored in HDFS follow the default replication size. However, you can specify a file/folder which does not use the default replication size but the replication size which you specify.

55. In QJM, the journal Node has just editlogs or it contains Fsimage as well?
    Ans: Only Edit Logs

56. Consider a 4-node cluster and if one data node goes down then it becomes a three-node cluster, so the replication happens in all the three nodes, what happens if the 4th node which was failed is replaced?
    Ans: If the failed data node comes back with data then NameNode will remove extra replication. If the failed data node comes back without data then it's consider new and use for newer jobs. Also it's very rare that you will have 4 node cluster.
    Refer Q no 57

57. What happens when the Data Node 1 is up again and running? Are the data block's (B1, B3, B4) retained or deleted from Data Node 1. If the data block's are retained, then the replication factor will be 4 because B1, B3 and B4 are already present in different Data Node's. If the data block's are not retained, then it clearly identifies non-optimal resource utilization.
    Ans: If DataNode1 failed because of disk failure then the DataBlock is deleted else its retained. When the failed data node comes back HDFS re-activates data blocks (except incase of disk failure) and will automatically delete the excess replicas (particularly replicas from the failed node which is now active) to maintain default replication factor. Hadoop will always make sure to fulfill exact replication factor.

Block
1
2
3

Node 1	Node 2	Node 3	Node 4	Node 5
1		1		1
	2	2	2	
3	3	3		

Node 2 Fails - DB replicated on another nodes

Node 1	Node 2	Node 3	Node 4	Node 5
1		1		1
	2	2	2	**2**
3	3	3	**3**	

Node 2 comes back with data then it will be deleted

Node 1	Node 2	Node 3	Node 4	Node 5
1		1		1
	~~2~~	2	2	2
3	~~3~~	3	3	

Node 2 comes back without data

Node 1	Node 2	Node 3	Node 4	Node 5
1		1		1
		2	2	2
3		3	3	

58. How a single data node of 128 mb manages to form 3 replicas.?

Ans:- 128 mb is the block size. A data node can have many blocks according to capacity (generally 1TB)

For example a 1 TB data node can have around 8192 blocks of 128mb.

Normally In local mode, replication is set to 1. Even if you set to 3 then all 3 copies will be stored on same machine. If the machines goes down all copies are lost so it does not mae sense having replication on single node cluster.

59. Is there any way to modify the files directly in HDFS without copying them in local, modify them and again overwrite them in HDFS

Ans:- HDFS is designed for read only service. It is not recommended to modify files in HDFS.

60. Is HDFS always on Linux only?
Ans: NO, IT can be on windows also, Hadoop on windows has lot of compatibility issues and error. Hence Linux is recommended

61. Is there any specific calculation to calculate how many Data Nodes require for a specific set of data like we have 10TB data so how many Data node clusters we require?

Ans: Company first designs the cluster and the admins decide the number of nodes based on the overall needs of the company. Many teams can use the cluster. Cluster depends on how many resources it has, not just number of DataNode.

62. John is not able to understand whether the no of mappers equals to no of blocks to which the data has been divided into or the no of data nodes into which the data has been placed?

Ans:- Number of mappers will be equal to number of DataBlocks.

For e.g : 500MB input file, 4 blocks are generated as per 128MB block size , so 4 mappers will be required.

One mapper is given to one container inside DataNode. A container is a task execution template allocated by the Resource Manager on any of the data node in order to execute the Map/Reduce tasks

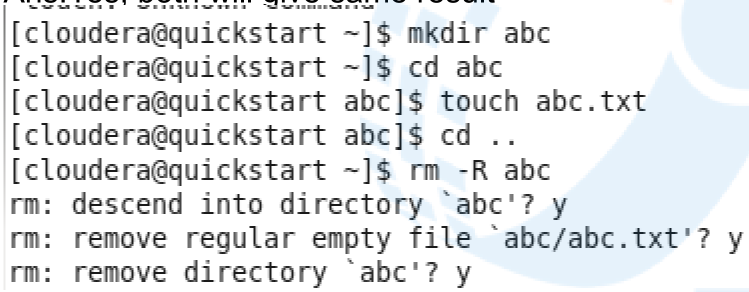
63. If any data node is failed, after that entry is deleted from the NameNode, then how is the NameNode going to identify which block copy is kept where?

Ans: Based on the block reports the NameNode receives from the DataNodes periodically, it knows the details about blocks and NameNode also knows the block mapping information. If data node fails, NameNode knows which data node failed, as it will stop getting the heartbeat from that DataNode. It also comes to know which blocks were residing in the failed DataNode. So, it then decides on re-replicating the under-replicated blocks, to other available data nodes in cluster. Rack awareness algo will also be taken into consideration.

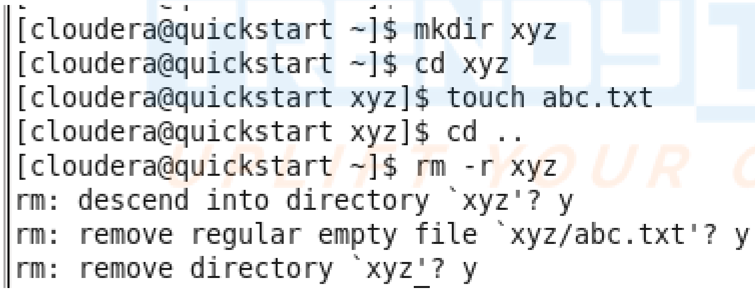
Linux Commands

64. After executing command “rm file1”, it asks rm:remove regular empty file ‘file1’? and press the enter button. But after checking the ls command still file is not deleted.
Ans: Its asking you for confirmation, you need to type Y and then only it will delete the file. You can use -f option to delete the file forcefully (does not ask for confirmation)

65. During removing the directory using this command rm -R in a recursive way as shown in video, Can we type rm -r to get the same result?
Ans:Yes, both will give same result



```
[cloudera@quickstart ~]$ mkdir abc
[cloudera@quickstart ~]$ cd abc
[cloudera@quickstart abc]$ touch abc.txt
[cloudera@quickstart abc]$ cd ..
[cloudera@quickstart ~]$ rm -R abc
rm: descend into directory `abc'? y
rm: remove regular empty file `abc/abc.txt'? y
rm: remove directory `abc'? y
```



```
[cloudera@quickstart ~]$ mkdir xyz
[cloudera@quickstart ~]$ cd xyz
[cloudera@quickstart xyz]$ touch abc.txt
[cloudera@quickstart xyz]$ cd ..
[cloudera@quickstart ~]$ rm -r xyz
rm: descend into directory `xyz'? y
rm: remove regular empty file `xyz/abc.txt'? y
rm: remove directory `xyz'? y
```

66. What is the difference between creating file using touch and cat commands.
Ans:

- cat command: It is used to create the file with content. Cat commands is used in many ways
  To show content
  cat abc.TXT
  To write to the file
  cat > abc.TXT
  To append to the file
  cat >> abc.TXT
- touch command: It is used to create a file without any content. The file created using touch command is empty.

67. John current working directory is Desktop. He is createatinf two files, one using touch and

another using cat. Where these two files will be created.
Ans: it will create the files in the PWD only. Whenever you create any file or directory without specifying path then the file/directory will be created in the directory you are currently in.
For example, if you are in /user/cloudera/download path and if you try to create file/directory then it will be created in download folder.

68. I am getting this error while running only “Hadoop fs -ls” command[cloudera@quickstart ~]$ Hadoop fs -ls

```
21/02/16 07:13:50 WARN ipc.Client: Failed to connect to server: quickstart.cloudera/10.0.2.15:8020: try once and fail.
java.net.ConnectException: Connection refused
at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:744)
```

Ans: Run this command : *service --status-all*

If the services are dead then you need to reinstall cloudera

69. I have observed in first video Linux commands are done in Cloudera terminal and in the second video it’s done in Linux platform window. So I use windows laptop So where do I need to execute these commands explained in the second video of Linux ?

Ans: It’s correct that it was done using the Linux platform.

On mac system (which is Linux based) can use Linux window as well as cloudera.

On Windows, you cannot execute these commands on terminal, you have to use cloudera terminal

Hadoop Commands

70. When I execute Hadoop fs -ls command I don’t see any output for this. So, do I need to set home directory /user/cloudera before executing this above command?

Ans: When you first time configure Hadoop, HDFS has no files or directories. You need to create it first

Hadoop fs -mkdir firstdir

This will create directory

Hadoop fs -ls

Now it should show you one directory firstdir

71. How to surf directories in HDFS?
Ans: You cannot surf directories like cd in linux. We can only check the available files or directories using ls command and full path
Hadoop fs -ls /user/userName/

Another way is to see in Web UI: Here you can click and enter the directory and see the directories and files inside it.

72. Give an example of getmerge command

Ans: getmerge is used to merge the multiple HDFS files into single file to the local directory.
Consider you have 4 files in HDFS under
/user/cloudera/data/file1.txt
/user/cloudera/data/file2.txt
/user/cloudera/data/file3.txt
/user/cloudera/data/file4.txtYou need to merge all these files into single file. Then getmerge is used.
Syntax:
Hadoop fs -getmerge <source_path> <destination_path>Hadoop fs -getmerge /user/cloudera/data/* /home/cloudera/Desktop/
Note: The destination path should be always local path.

73. As we do pwd in the cloudera terminal which displays the present working directory, So is there a way to display working directory in HDFS commands ?

Ans- pwd is a Linux command, not HDFS command.

We always access HDFS from outside (using terminal) also you can't change directory in HDFS like cd in Linux. So, it is meaningless to have commands like pwd in HDFS.

74. While moving file2.txt from local to HDFS, John is facing the checksum error. Previously he moved file1.txt without any issue. Please help me where am I getting it wrong.

Ans: CheckSumException can be caused because of various reason

1. Hadoop libraries issue : since your file1 was copied successfully hence this reason is not applicable

2. File is not closed properly after write.

Solution :

1. If you have written file then make sure you have closed it properly.

2. Copy the contents of file2.txt to a new file say xyz.txt and then try to move it.

75. getmerge command in HDFS is for merging two HDFS files into a single file in local. Can we merge two files in HDFS into a single file in HDFS itself?

Ans: There is no direct way to do it.

However a trick can do it by redirecting cat command output stream to the put command and output the result to an HDFS file.

Hadoop fs -cat /user/sumit/myfiles/* | Hadoop fs -put - /user/sumit/dest

Here all files from myfiles directory will be added to a file in dest directory

76. How do I view the contents of a file in HDFS? Does cat work with the HDFS commands. Also,can we use CD to navigate to a folder/directory in HDFS using Hadoop fs - ?

Ans:- Hadoop has cat command but not cd command because we run HDFS from linux terminal. Everytime you need to specify full path in HDFS.

Hadoop fs -cat <HDFS file path>

77. John is not able to see copied file in the directory.

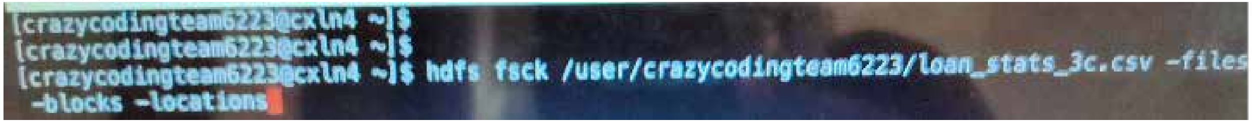
```
[cloudera@quickstart ~]$ Hadoop fs -copyFromLocal Desktop/file.txt /testr
[cloudera@quickstart ~]$ Hadoop fs -ls /testr
-rw-r--r--   1 cloudera supergroup        10 2021-02-19 22:42 /testr
[cloudera@quickstart ~]$ Hadoop fs -ls /testr/
-rw-r--r--   1 cloudera supergroup        10 2021-02-19 22:42 /testr
```

Ans: The file has been copied but the file name here is /testr because you didnt create a directory in HDFS before running copyFromLocal command.

First create a directory testr in HDFS location and then run the copyFromLocal command.

Hadoop fs -mkdir -p /user/cloudera/testr

78. Why Hadoop fs is not written infront of this command as this command runs in HDFS to show metadata?



```
[crazycodingteam6223@cxln4 ~]$
[crazycodingteam6223@cxln4 ~]$
[crazycodingteam6223@cxln4 ~]$ hdfs fsck /user/crazycodingteam6223/loan_stats_3c.csv -files
 -blocks -locations
```

Ans: “Hadoop fs” is more generic but “HDFS” points to specifically HDFS file system. when you say “Hadoop fs” it covers multiple file systems for example local file system and HDFS. But when you say “HDFS” it points to HDFS only.

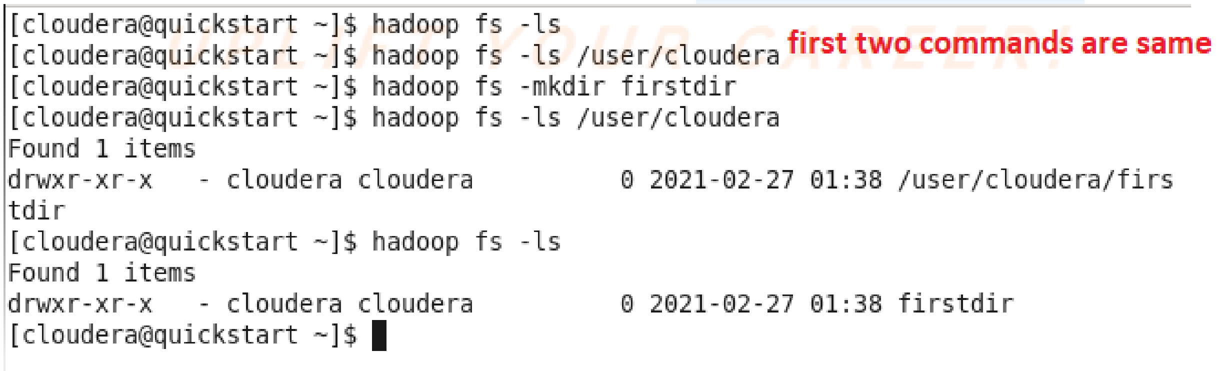
79. I logged on cloudxlab with my credential. I worked on Linux commands. But how can I work HDFS commands in cloudxlab? What path do I need to redirect?

Ans: When you first time use Hadoop, HDFS has no files or directories. You need to create it first

Hadoop fs -mkdir firstdir //This will create directory

Hadoop fs -ls   //Now it should show you one directory firstdir

Note : Your HDFS has one directory firstdir now, then you must specify the path whichever you created in all commands



```
[cloudera@quickstart ~]$ hadoop fs -ls
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera
[cloudera@quickstart ~]$ hadoop fs -mkdir firstdir
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2021-02-27 01:38 /user/cloudera/firs
tdir
[cloudera@quickstart ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2021-02-27 01:38 firstdir
[cloudera@quickstart ~]$ 
```

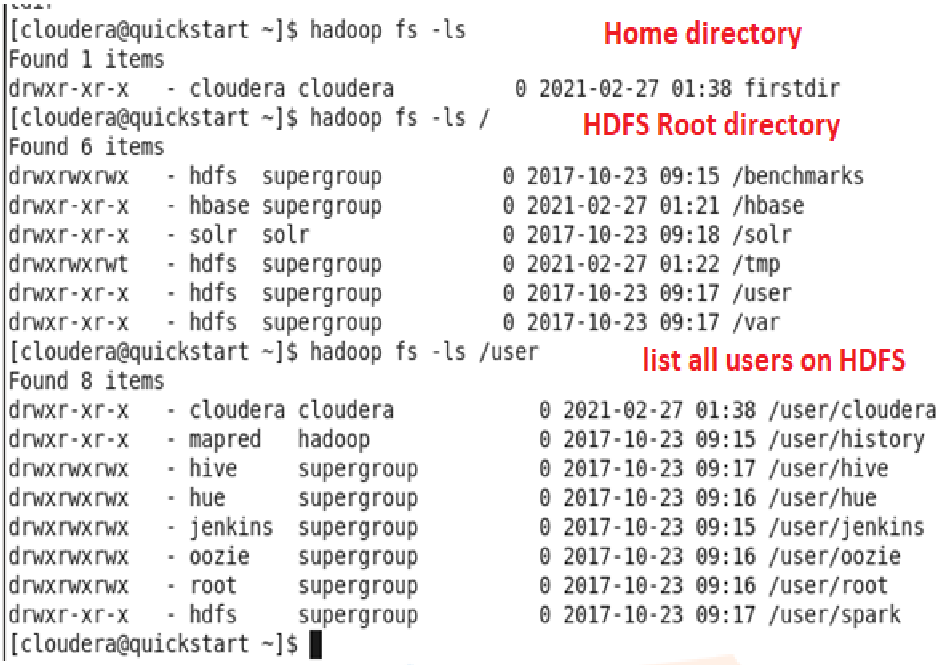
first two commands are same

80. What is the main difference between these two commands

Hadoop fs -ls and Hadoop fs -ls /

Ans: HDFS root directory which is referred to as (/) is the topmost drive

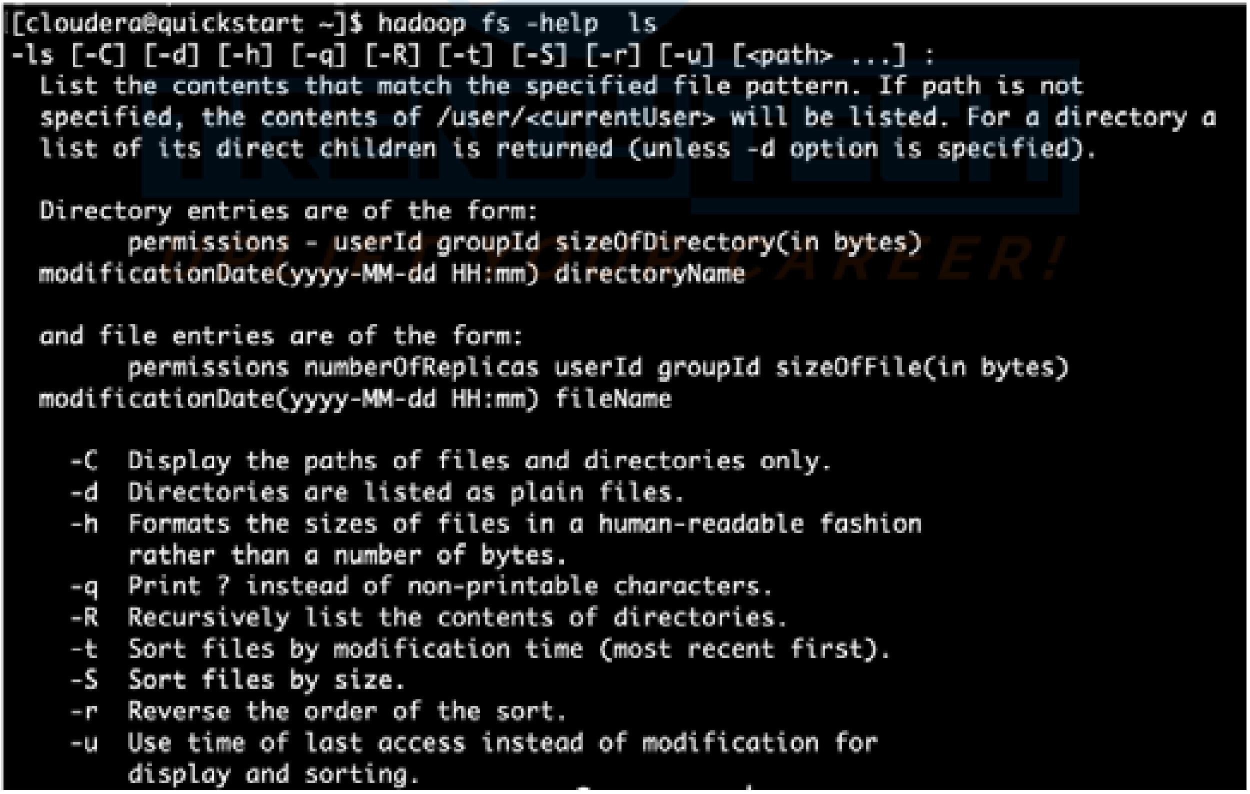
HDFS Home directory which is /user/userName comes under the root directory.



```
[cloudera@quickstart ~]$ hadoop fs -ls
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2021-02-27 01:38 firstdir
[cloudera@quickstart ~]$ hadoop fs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2021-02-27 01:21 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2021-02-27 01:22 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hadoop fs -ls /user
Found 8 items
drwxr-xr-x   - cloudera cloudera            0 2021-02-27 01:38 /user/cloudera
drwxr-xr-x   - mapred   hadoop              0 2017-10-23 09:15 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-10-23 09:17 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-10-23 09:16 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-10-23 09:15 /user/jenkins
drwxrwxrwx   - oozie    supergroup          0 2017-10-23 09:16 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-10-23 09:16 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user/spark
[cloudera@quickstart ~]$ ▌
```

The admin has the access to the root directory for any changes particularly the configuration settings, whereas the user has only access to home directory. You can clearly see in below picture root directory shows files of all users where as home directory shows for particular user (Here cloudera is a username)

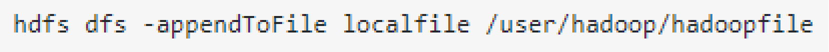
81. If we do ls -lrt in local, it shows the list of all files/folders in local but if I run the same command in Hadoop as : Hadoop fs -ls -lrt, it doesn't work, why is it so?
Ans: Hadoop ls command is little different from linux ls command in arguments .
Hadoop fs -ls -r -t /user/cloudera/



```
[cloudera@quickstart ~]$ hadoop fs -help  ls
-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...] :
  List the contents that match the specified file pattern. If path is not
  specified, the contents of /user/<currentUser> will be listed. For a directory a
  list of its direct children is returned (unless -d option is specified).

  Directory entries are of the form:
        permissions - userId groupId sizeOfDirectory(in bytes)
  modificationDate(yyyy-MM-dd HH:mm) directoryName

  and file entries are of the form:
        permissions numberOfReplicas userId groupId sizeOfFile(in bytes)
  modificationDate(yyyy-MM-dd HH:mm) fileName

    -C  Display the paths of files and directories only.
    -d  Directories are listed as plain files.
    -h  Formats the sizes of files in a human-readable fashion
        rather than a number of bytes.
    -q  Print ? instead of non-printable characters.
    -R  Recursively list the contents of directories.
    -t  Sort files by modification time (most recent first).
    -S  Sort files by size.
    -r  Reverse the order of the sort.
    -u  Use time of last access instead of modification for
        display and sorting.
```

82. Why 'cd' and 'pwd' commands not included in HDFS ?
Ans: There is no working directory concept in Hadoop Distributed Filesystem and hence pwd and cd commands are not included.

83. how do we edit a file in HDFS?
Ans: HDFS is of write once read many model hence editing is not supported.
You can use



```
hdfs dfs -appendToFile localfile /user/hadoop/hadoopfile
```

appendToFile is generally used when you want to merge multiple small files into one and put it in HDFS.

84. I'm using Hadoop fs -ls command but it gives the same files and Directories that are in local, is that expected?

Ans: Normally the files won't be same.

He got different results because he has lots of directories already present in his HDFS cluster.

Just practice 2_3 times and it will be clear for you.

Just try to distinguish what is local and what is HDFS. When you communicate with HDFS u need to prefix Hadoop fs but for local the commands will run directly.

Ex:

For local: ls

For HDFS: Hadoop fs -ls

85. John has a concern on -create snapshot command, why we need SNAPSHOT command even-though we have Fault tolerance mechanism in HDFS architecture. snapshot also making a recovery for disaster management same replication do?

Ans: - Snapshot does not copy or replicate actual data. It is just pointers to original data.

The snapshot mechanism lets administrators persistently save the current state of the filesystem, so that if the upgrade results in data loss or corruption it is possible to rollback the upgrade and return HDFS to the namespace and storage state as they were at the time of the snapshot.

Some replications, especially those that require a long time to finish, can fail because source files are modified during the replication process. You can prevent such failures by using Snapshots in conjunction with Replication.
