

What is umask in linux?

The user file-creation mode mask (umask) is used to determine the file permissions for newly created files or directories. It can be used to control the default file or directory permissions for new files. It is a four-digit octal number. The umask value for normal user is 0002 and the umask value for root user is 0022.

So, the effected file permissions for normal users =  $666 - 002 = 664$ .

The effected directory permissions for normal users =  $777 - 002 = 775$ .

The effected file permissions for root user =  $666 - 022 = 644$

The effected directory permissions for root user =  $777 - 022 = 755$

# umask <value> (to change the umask value temporarily)

# vim /etc/bashrc (open this file and change the umask value to effect the whole system)

# source /etc/bashrc (to updated the source file)

# vim .bashrc (open this file in user's home directory and at last type as follows)

umask <value>(save and exit the file)

# source .bashrc or logout and login again (to the system to effect that umask value)

\* If the/etc/login.defs file is corrupted then new users will be added and can be assigned the passwords but users cannot login.

\* If the /etc/login.defs file is deleted then new users cannot be added.

Ps Arshad: There are 7 types of files.

1. - ----> regular file
2. d ----> directory
3. c ----> character device file (Ex. console file, open and close terminals, ...etc.,)
4. b ----> block device file (Ex. device blocks like hard disks, CD/DVD disks)
5. s ----> socket file (programmers will deal this file)
6. p ----> pipe file (programmers will deal this file)
7. l ----> linked file (nothing but short cut file)

Ps Arshad: \*FILE SYSTEM ?\*

It is a method of storing the data in an organized fashion on the disk. Every partition on the disk except MBR don't have file system, but Extended partition should be assigned with some file system in order to make them to store the data. File system is applied on the partition by formatting it with a particular type of file system.

[09/03, 08:11] Ps Arshad: \*TYPES OF FILE SYSTEM\*

ext2 ---->Second extended file system (default in RHEL - 3 & 4)

ext3 ---->Third extended file system (default in RHEL - 5)  
ext4 ----> Fourth extended file system (default in RHEL - 6)  
xfs ----> Extended file system (default in RHEL - 7)  
ufs ----> Unix file system (default in Solaris)  
jfs ----> Journal file system (default in IBM-AIX)  
hfs ----> High performance file system (default in HP-UX)  
vxfs ----> Veritas file system  
procfs ----> Process file system (temporary)  
tempfs ----> Temporary file system (temporary)  
cdfs ----> Compact disk file system  
hdfs ----> DVD file system  
iso9660 ----> To read the CD/DVD.iso image format files in Linux

Ps Arshad: \*How to troubleshoot if the boot disk is not available?\*

(i)First check the hard disk is present in the system or not. If not present connect the hard disk and restart the system.

(ii)If the hard disk is present, then go to BIOS and find the location of the hard disk.

(iii) Check the boot priority in the BIOS. If boot priority is not the hard disk then change it to hard disk and restart the system.

(iv)Even though the system is not started then boot the system with CDROM in single user mode and open the `*/boot/grub/grub*.conf` file and see the hard disk name and partition number. Normally it should be `/dev/hda1` (if the hard disk is IDE hard disk) or `/dev/sda1` (if the hard disk is SATA or SCSI). If the hard disk name and partition number is different instead of the original then change them and reboot the system with hard disk.

(v)If the GRUB is corrupted then reboot the system with CDROM in single user mode and restore the grub information from the recent backup and then restart the system with hard disk.

Ps Arshad: \*How to reboot the production server?\*

(i)In general the production servers will not be rebooted frequently because the end users will suffer if the productions server are in down state. If any changes made to the system like grub, selinux policy, default run level is changed and if kernel patches are applied the system reboot is required.

(ii) If any inconsistency in root ( / ) file system, then take the business approval from higher authorities,make a plan for proper schedule and also inform to the different teams like application team to stop the application, databaseteam to stop the databases, QC team to stop the testing, monitoring people to ignore the alerts from this server and other teams if any and then reboot the system with CDROM in single user mode and then run `#fsck` command on that file system.

(iii) If O/S disk is corrupted or damaged then, reboot the system temporarily with the mirror disk then fix that problem and again boot the system with original disk.

Ps Arshad: \*System is continuously rebooting. How to troubleshoot it?\*

(i) Connect the system through console port through putty by providing IP address, user name and password.

(ii) At console prompt boot with CDROM in single user mode and mount the root ( / ---> O/S) file system on temporary mount point.

(iii) Check any wrong entries in the cron jobs i.e., crontab editor see any reboot scripts are there or not. If found remove those entries and reboot the system.

(iv) If the above is not resolved, then check the memory (RAM).

(v) If RAM module is not working the system get panic and it may continuously reboots.

(vi) If the RAM module is working then check the RAM size whether the sufficient RAM that requires to run the application is available or not. If not there then increasing the RAM size may resolve this issue.

(vii) Check "/var/log/messages" file for any messages regarding continuous rebooting.

(viii) Even though there is a sufficient RAM may be swap space is not sufficient to run all the services and applications then system get panic and may continuously reboots. If so, then increasing the swap size may resolve this issue.

Ps Arshad: \*What are the differences between a daemon and a process?\*

(i) Daemon is a service to provide some services to the users, whereas a process is to do some particular tasks.

(ii) We can enable or disable the daemon, but we cannot disable or enable the process.

(iii) We can do start or stop the daemon, but we cannot start or stop the process. We only kill the process.

(iv) We can enable or disable to start the daemons at boot time as per our requirement, i.e., on demand is possible, but it is not possible if it is a process.

(v) Daemon is a background process whereas process is a foreground process.

Ps Arshad: \*What is process and explain it?\*

A process is a set of instructions which executes in the memory. It is created in the memory when a program or command is executed. Every process is identified by a unique no. i.e., PID (Process ID). Several processes are started at boot time and which are running at background called daemons. The Linux kernel is used to communicate with the processes by their process ID's (PID's). Daemon is a process running in the background. These are handled by the system and process are handled by the users.

The first process in RHEL - 6 is \*initd\* and it starts at boot time. Its process ID is 1 whereas in RHEL - 7 the first process is \*systemd\* and it starts at boot time

Ps Arshad: \*What is service or daemon?\*

Service or daemon is a program that starts at background and continuously runs in the background. The service or daemon is ready for input or monitors the changes in our computer and responds to them. For example, the Apache server has a daemon called `httpd` that listens on port no. 80 on our computer and when it receives a request for a page, it sends the appropriate data back to the client machine.

Ps Arshad: What is a user?

In Linux, a user is one who uses the system.

How many types of users are available in Linux?

There are 5 types of users available in Linux.

(i) System user (Admin user who controls the whole system, nothing but root user).

(ii) Normal user (Created by the Super user. In RHEL - 7, the user IDs are from 1000 - 60000).

(iii) System user (Created when application or software is installed).

(In RHEL - 7, the System users are Static system user IDs from 1 - 200 and

(ii) Dynamic system user user IDs from 201 - 999).

(iv) Network user (Nothing but remote user, i.e., who are logged in to the system through network, created in Windows Active Directory or in Linux LDAP or NIS).

(v) Sudo user (The normal users who are having admin or Super user privileges)

Ps Arshad: \*What is the journaling filesystem?\*

A journaling filesystem is a filesystem that maintains a special file called a journal that is used to repair any inconsistencies that occur as the result of an improper shutdown of a computer.

In journaling file systems, every time GFS2 writes metadata, the metadata is committed to the journal before it is put into place.

This ensures that if the system crashes or loses power, you will recover all of the metadata when the journal is automatically replayed at mount time.

GFS2 requires one journal for each node in the cluster that needs to mount the file system. For example, if you have a 16-node cluster but need to mount only the file system from two nodes, you need only two journals. If you need to mount from a third node, you can always add a journal with the `gfs2_jadd` command.

Ps Arshad: \*How to replace the failed hard disk?\*

(a) Check whether the disk is failed or not by `# iostat -En | grep -i hard/soft` command.

(b) If hard errors are above 20, then we will go for replacement of the disk.

(c) If the disk is from SAN, people then we will inform them about the replacement of the disk. If it is an internal disk, then we raise the CRQ to replace the disk.

(d) For this, we will consider two things.

(i) whether the system is within the warranty.

(ii) without warranty.

(e) We will directly call to the toll free no. of the system vendor and raise the ticket. They will issue the case no. This is the no. we have to mention in all correspondences to vendor regarding this issue.

(f) If it is having warranty they ask rack no. system no. and other details and replace the hard disk with co-ordinate of the data centre people.

(g) If it is not having warranty, we have to solve the problem by our own or re-agreement to extend the warranty and solve that problem.

Ps Arshad: \*What is syntax of the usermod command with full options?\*

```
# usermod <options><user name>
```

\* The options are,

-L -----> lock the password

-U -----> unlock the password

-o -----> creates duplicate user modify the user's id same as other user

-u -----> modify user id

-g -----> modify group id

-G -----> modify or add the secondary group

-c -----> modify comment

-d -----> modify home directory

-s -----> modify user's login shell

-l -----> modify user's login name

-md ----> modify the users home directory and the old home directory

Ps Arshad: \*What is the syntax to assign read and write permissions to particular user, group and other at a time?\*

```
# setfacl -m u : <user name> : <permissions>, g : <user name> : <permissions>, o : <user name> :  
      <permissions><file or directory>
```

\*Useful commands :\*

```
# setfacl -x u : <user name><file or directory name> (to remove the ACL permissions from the  
user)
```

```
# setfacl -x g : <user name><file or directory name>(to remove the ACL permissions from group)
```

```
# setfacl -x o : <user name><file or directory name> (to remove the ACL permissions from other)
```

```
# setfacl -b <file or directory>      (to remove all the ACL permissions on that file  directory)
```

Ps Arshad: \*What is Access Control List (ACL)?\*

Define more access rights nothing but permissions to files and directories. Using Access Control list we assign the permissions to some particular users to access the files and directories.

ACL can be applied on ACL enabled partition that means you need to enable ACL while mounting

the partition.

=====

How to implement ACLs?\*

Create a partition and format it with ext4 file system.

Mount the file system with ACL.

Apply ACL on it.

Create a partition using `# fdisk` command.

Format the above partition with ext4 file system using `# mkfs.ext4 <partition name>` command.

Create the mount point using `# mkdir /<mount point>` command.

Mount that file system on the mount point using `# mount -o acl <partition name><mount point>` command.

Mount the partition permanently using `# vim /etc/fstab` (open this file and make an entry as below)

```
<partition name><mount point><file system type> defaults, acl 0 0
```

Save and exit this file.

If the partition is already mounted then just add `acl` after `defaults` in `/etc/fstab` file and execute the below command `# mount -o remount <partition name>`

=====

\*How to check the ACL permissions?\*

```
# getfacl <options><file or directory name>
```

The options are, `-d` -----> Display the default ACLs.

`-R` -----> Recurses into subdirectories.

=====

\*How to assign ACL permissions?\*

```
# setfacl <options><argument> : <username>: <permissions><file or directory name>
```

The options are, `-m` -----> Modifies an ACL.

`-x` -----> Removes an ACL.

-b -----> Remove all the ACL permissions on that directory.

-R -----> Recurses into subdirectories.

The arguments are, u -----> user, g -----> group, o -----> other

=====

\*What is the syntax to assign read and write permissions to particular user, group and other?\*

```
# setfacl -m u : <user name> : <permissions><file or directory>
```

```
# setfacl -m g : <user name> : <permissions><file or directory>
```

```
# setfacl -m o : <user name> : <permissions><file or directory>
```

Ps Arshad: \*What is parent process?\*

The process which starts or creates another process is called the parent process. Every process will be having a parent process except initd/systemd process. The initd/systemd process is the parent process to all the remaining processes in

Linux system because it is the first process which gets started by the kernel at the time of booting and it's PID is 1. Only after initd/systemd process gets started, the remaining processes are called by it, and hence it is responsible for all the remaining processes in the system.

The parent process is identified by PPID (parent process ID).

=====

\*What is child process?\*

A process which started or created by the parent process is called child process and it is identified by PID.

=====

\*Useful # ps commands :\*

```
# ps -a (it displays all the terminals processes information)
```

```
# ps -au (it displays all the terminals processes information with user names)
```

```
# ps -aux (it displays all the terminals processes information including background processes with user names)
```

\* ? (question mark) if it is appeared at tty column, it indicates that is a background process.

```
# ps -ef (it displays the total processes information with parent process ID (PPID))
```

```
# ps -P <process id> (it displays the process name if we know the process ID (pid))
```

```
# pidof<process name> (to see the process ID of the specified process)
```

```
# pidof initd (to see the process ID of the initd process)
```

```
# pstree      (to display the parent and child processes structure in tree format)
# ps -u <user name> (to display all the processes of the specified user)
# ps -u raj    (to display all the processes of the user raj)
# ps -G <group name>(to display all the processes that are running by a particular group)
```

Ps Arshad: \*~What is Orphan process?~\*

The processes which are running without parent processes are called Orphan processes. Sometimes parent process closed without knowing the child processes. But the child processes are running at that time. These child processes are called Orphan processes.

=====

~\*What is Zombie process?~\*

When we start parent process, it will start some child processes. After some time the child processes will died because of not knowing the parent processes. These parent processes (which are running without child processes) are called Zombie processes. These are also called as defaunct processes.

=====

\*How to set the priority for a process?\*

Processes priority means managing processor time. The processor or CPU will perform multiple tasks at the same time. Sometimes we can have enough room to take on multiple projects and sometimes we can only focus on one thing at a time. Other times something important pops up and we want to devote all of our energy into solving that problem while putting less important tasks on the back burner.

In Linux we can set guidelines for the CPU to follow when it is looking at all the tasks it has to do. These guidelines are called \*niceness or nice value\*. The Linux \*niceness scale goes from -20 to 19\*. The lower the number the more priority that task gets. If the niceness value is higher number like 19 the task will be set to the lowest priority and the CPU will process it whenever it gets a chance. The default nice value is 0 (zero).

By using this scale we can allocate our CPU resources more appropriately. Lower priority programs that are not important can be set to a higher nice value, while the higher priority programs like deamons and services can be set to receive more of the CPU's focus. We can even give a specific user a lower nice value for all his/her processes so we can limit their ability to slow down the computer's core services.

There are two options to reduce/increase the value of a process. We can either do it using the nice or renice commands.

\*Examples :\*

```
# nice -n <nice value range from -20 to 19><command> (to set a priority to a process before starting it)
```

# nice -n 5 cat > raju (to set the medium priority to cat command)

# ps -elf (to check the nice value for that command)

\*To reschedule the nice value of existing process, first check the PID of that process by # ps -elf command

and then change the niceness of that command by

# renice <nice value (-20 to 19)><PID> command.

# renice 10 1560 (to reschedule the PID 1560)

Ps Arshad: \*How to recover if a file system is corrupted or crashed?\*

If the normal or not related to O/S file system is \*corrupted\* first unmount that file system and run fsck command on that file system and if the O/S related file system is corrupted then boot the system with CDROM in single user mode and run the fsck command.

If the normal or not related to O/S file system is \*crashed\* then restore it from the recent backup and if the O/S related file system is crashed then boot the system with CDROM in single user mode and restore it from the recent backup.

Ps Arshad: \*What is a link file and how many types?\*

Link file is a short cut file to the original file. Creating and removing (deleting) inks between two files is known as managing links. There are two types of links files available in Linux.

(i) Soft link

(ii) Hard link

=====

\*What is soft link and how to create it?\*

\*~Soft link~\* is nothing but a short cut file. If original file is deleted, no use of short cut file. ie., we cannot access the original data by selecting the link file. Soft link can be applied on both directories and files. These files can be stored in any of the file system. ie., the original file may be in one file system and the link file may be on another file system. If we edit any file, the link files are also updated automatically. When we create a soft link file, the permissions are full permissions. The soft link file and the original file inode no's are different. The size of the soft link file is same as the length of the original file name. The soft link can be created by

# ln -s <original file or directory><link file or directorywith path> (to create a soft link)

# ln -s /root/script /root/Desktop/script (to create a link file for the script and stored on root Desktop)

=====

\*What is hard link and how to create it?\*

\*~Hard link~\* is nothing but a backup file. If the original file is deleted, there is no effect on hard link file. i.e., we can access the original file data even though the link file is deleted. Hard links can be applied on files only not on directories. Hard link files can be stored in the same file system. i.e., original and hard link files both should be in the same file system not on different file systems. The inode no's are same for original and hard link files. If the original is edited, the updations are applied on both original and hard link files. The size of the hard link file is same as the size of the original file.

[09/03, 08:12] Ps Arshad: \*What are the commands to search files and directories?\*

To search files and directories there are two commands.

- (i) # locate
- (ii) # find

=====

\*Explain the locate command and how to use it?\*

locate always looks the locate database and not in a specific location. The data of the locate is stored in `*/var/lib/mlocate/mlocate.db` file.\* If the data is not updated in locate database or the locate database is not available or locate database is deleted, we cannot locate the files and directories. `## updatedb*` is the command to update the locate database. locate database cannot be find the newly created files and directories. It is *\*not recommended to use on production servers because it impacts on performance of the servers.\** So, to overcome this problem we normally use `## find*` command on production servers.

`## updatedb*` (to update the locate database)

`## locate*` <file name/directory name> (to search the specified file or directory)

=====~~\*\*\*\*\*~~=====~~\*\*\*\*\*~~=====~~\*\*\*\*\*~~=====~~\*\*\*\*\*~~=====

\*Explain the find command and how to use it?\*

Find command require the specific location. Without specific location we cannot find the files or directories.

`# find <location><options><file or directory>` (to find the specific file or directory)

The options are,

- name -----> search files and directories
- perm -----> search for permissions
- size -----> search for sizes
- user -----> search for the owner
- uid -----> search for files/directories of uid)

-gid -----> search for files/directories of gid)  
-group -----> search for group owner  
-empty -----> search for empty files  
-amin -----> search for access time  
-atime -----> search for access day (access day, minutes, hrs, ...etc)  
-mtime -----> search for modify day (change the content)  
-ctime -----> search for change day (permissions, .....etc)

~\*Examples :\*~

# find / -name <file name> (to search for file names in / directory)  
# find / -name <file name> -type f (to find file names only)  
# find / -name <directory name> -type d (to find directories with small letters only)  
# find / -iname <file/directory name> -t d (to search for small or capital letter files/directories)  
# find / -empty (to search empty files or directories)  
# find / -empty -type f (to search for empty files only)  
# find / -empty -type d (to search for empty directories only)

# find / -name "\*.mp3" (to search for .mp3 files only)  
# find / -size 10M (to search for exact 10M size file/directories)  
# find / -size -10M (to search for less than 10M size files/directories)  
# find / -size +10M (to search for greater than 10M size files/directories)  
# find / -user student (to search for student user files/directories)  
# find / -group student (to search for student group files/directories)  
# find / -user student -not -group student (to search for student user files and not student group files)  
# find / -user student -o -group student (to search for student user and student group files/directories)  
# find / -uid <uid no.> (to search for files/directories which belongs to the user having the specified user id)  
# find / -gid <gid no.> (to search for files/directories which belongs to the group having the specified group id)  
# find / -perm 755 (to search file/directories which are having the permissions 755)  
# find / -perm -755 (to search file/directories which are having the

permissions below 755 and also at least one match also)

```
# find / -mmin 20
```

(to search for files/directories which are modified within 20 minutes,  
+20 ----> above 20 minutes and

-20 -----> below 20 minutes)

Ps Arshad: \*How to solve the issue if the CPU utilization is 99% ?\*

(i)First check which process and who executed that process is consuming more CPU utilization or memory utilization by executing # top command.

(ii)Then inform to those users who executed that process though mail, message or raising the ticket.

(iii)If those users are not available or not responding to our mail then we have to change the priority of that process using # renice command.

(iv) Before changing the process priority level , we have to get or take approval from our team lead or project manager.

Ps Arshad: [2/19, 12:14 PM] +1 (510) 764-6871: What is the use of Zookeeper in Storm?

Storm uses Zookeeper for coordinating the cluster. Zookeeper is not used for message passing. If you want failover or are deploying large Storm clusters you may want larger Zookeeper clusters.

It's critical that you run Zookeeper under supervision, it is a fail-fast type and will exit the process if it encounters any error case. You set up a cron to compact Zookeeper's data and transaction logs. The Zookeeper daemon does not do this on its own, and without cron, Zookeeper will quickly run out of disk space.

[2/19, 12:18 PM] +1 (510) 764-6871: Apache Kafka: It is a distributed and robust messaging system that can handle huge amount of data and allows passage of messages from one end-point to another.

Apache Storm: It is a real time message processing system, and you can edit or manipulate data in real time. Apache storm pulls the data from Kafka and applies some required manipulation.

[2/20, 4:13 AM] +1 (510) 764-6871: 4)

What is the fundamental difference between a MapReduce Split and a HDFS block?

MapReduce split is a logical piece of data fed to the mapper. It basically does not contain any data but is just a pointer to the data. HDFS block is a physical piece of data.

5) When is it not recommended to use MapReduce paradigm for large scale data processing?

It is not suggested to use MapReduce for iterative processing use cases, as it is not cost effective, instead Apache Pig can be used for the same.

6) What happens when a DataNode fails during the write process?

When a DataNode fails during the write process, a new replication pipeline that contains the other DataNodes opens up and the write process resumes from there until the file is closed. NameNode observes that one of the blocks is under-replicated and creates a new replica asynchronously.

7) List the configuration parameters that have to be specified when running a MapReduce job.

Input and Output location of the MapReduce job in HDFS.

Input and Output Format.

Classes containing the Map and Reduce functions.

JAR file that contains driver classes and mapper, reducer classes.

8) Is it possible to split 100 lines of input as a single split in MapReduce?

Yes this can be achieved using Class NLineInputFormat

9) Where is Mapper output stored?

The intermediate key value data of the mapper output will be stored on local file system of the mapper nodes. This directory location is set in the config file by the Hadoop Admin. Once the Hadoop job completes execution, the intermediate will be cleaned up.

10) Explain the differences between a combiner and reducer.

Combiner can be considered as a mini reducer that performs local reduce task. It runs on the Map output and produces the output to reducers input. It is usually used for network optimization when the map generates greater number of outputs.

Unlike a reducer, the combiner has a constraint that the input or output key and value types must match the output types of the Mapper.

Combiners can operate only on a subset of keys and values i.e. combiners can be executed on functions that are commutative.

Combiner functions get their input from a single mapper whereas reducers can get data from multiple mappers as a result of partitioning.

11) When is it suggested to use a combiner in a MapReduce job?

Combiners are generally used to enhance the efficiency of a MapReduce program by aggregating the intermediate map output locally on specific mapper outputs. This helps reduce the volume of data that needs to be transferred to reducers. Reducer code can be used as a combiner, only if the operation performed is commutative. However, the execution of a combiner is not assured.

12) What is the relationship between Job and Task in Hadoop?

A single job can be broken down into one or many tasks in Hadoop.

13) Is it important for Hadoop MapReduce jobs to be written in Java?

It is not necessary to write Hadoop MapReduce jobs in Java but users can write MapReduce jobs in any desired programming language like Ruby, Perl, Python, R, Awk, etc. through the Hadoop Streaming API.

14) What is the process of changing the split size if there is limited storage space on Commodity Hardware?

If there is limited storage space on commodity hardware, the split size can be changed by implementing the "Custom Splitter". The call to Custom Splitter can be made from the main method.

15) What are the primary phases of a Reducer?

The 3 primary phases of a reducer are –

1) Shuffle 2) Sort 3) Reduce

16) What is a TaskInstance?

The actual Hadoop MapReduce jobs that run on each slave node are referred to as Task instances. Every task instance has its own JVM process. For every new task instance, a JVM process is spawned by default for a task.

17) Can reducers communicate with each other?

Reducers always run in isolation and they can never communicate with each other as per the Hadoop MapReduce programming paradigm.

18) What is the difference between Hadoop and RDBMS?

In RDBMS, data needs to be pre-processed being stored, whereas Hadoop requires no pre-processing.

RDBMS is generally used for OLTP processing whereas Hadoop is used for analytical requirements on huge volumes of data.

Database cluster in RDBMS uses the same data files in shared storage whereas in Hadoop the storage is independent of each processing node.

19) Can we search files using wildcards?

Yes, it is possible to search for file through wildcards.

20) How is reporting controlled in hadoop?

The file `hadoop-metrics.properties` file controls reporting.

21) What is the default input type in MapReduce?

Text

22) Is it possible to rename the output file?

Yes, this can be done by implementing the multiple format output class.

23) What do you understand by compute and storage nodes?

Storage node is the system, where the file system resides to store the data for processing.

Compute node is the system where the actual business logic is executed.

24) When should you use a reducer?

It is possible to process the data without a reducer but when there is a need to combine the output from multiple mappers – reducers are used. Reducers are generally used when shuffle and sort are required.

25) What is the role of a MapReduce partitioner?

MapReduce is responsible for ensuring that the map output is evenly distributed over the reducers. By identifying the reducer for a particular key, mapper output is redirected accordingly to the respective reducer.

26) What is identity Mapper and identity reducer?

IdentityMapper is the default Mapper class in Hadoop. This mapper is executed when no mapper class is defined in the MapReduce job.

IdentityReducer is the default Reducer class in Hadoop. This mapper is executed when no reducer class is defined in the MapReduce job. This class merely passes the input key value pairs into the output directory.

27) What do you understand by the term Straggler ?

A map or reduce task that takes unusually long time to finish is referred to as straggler.

Please share your interview experience on mapreduce questions asked in your interview in the comments below to help the big data community.

2) Explain about the basic parameters of mapper and reducer function.

Mapper Function Parameters

The basic parameters of a mapper function are LongWritable, text, text and IntWritable.

LongWritable, text- Input Parameters

Text, IntWritable- Intermediate Output Parameters

Here is a sample code on the usage of Mapper function with basic parameters –

```
public static class Map extends MapReduceBase implements Mapper {  
private final static IntWritable one = new IntWritable (1);  
private Text word = new Text () ;}
```

Reducer Function Parameters

The basic parameters of a reducer function are text, IntWritable, text, IntWritable

First two parameters Text, IntWritable represent Intermediate Output Parameters

The next two parameters Text, IntWritable represent Final Output Parameters

How data is spilt in Hadoop?

The InputFormat used in the MapReduce job create the splits. The number of mappers are then decided based on the number of splits. Splits are not always created based on the HDFS block size. It all depends on the programming logic within the getSplits () method of InputFormat.

Ps Arshad: \*What is loopback address?\*

A special IP number (127.0.0.1) is designated for the software loopback interface of a machine. 127.0.0.0 and 127.255.255.255 is also reserved for loopback and is used for internal testing on local machines.

=====

\*What is multicasting?\*

Multicasting allows a single message to be sent to a group of recipients. Emailing and Teleconferencing are examples of multicasting. It uses the network infrastructure and standards to send messages.

=====

\*What is subnet mask?\*

A subnet mask allows the users to identify which part of an IP address is reserved for the network and which part is available for host use.

=====

\*What is Gateway?\*

A Gateway is the network point that provides entrance into another network. On the internet a node or stopping point can be either gateway node or a host (end point) node. Both the computers of internet users and the computer that serve the pages to users are host nodes. The computer that control traffic within your company's network or at our local internet service provider (ISP) are the gateway nodes.

Ps Arshad: \*Explain about set uid (suid)?\*

If we plan to allow all the users to execute the root users command then we go for set uid (suid).

It can be applied for user level and is applicable for files only.

# chmod u+s <file name> (to set the suid on that file)

# chmod u-s <file name> (to remove the suid from that file)

# ls -l (if 'x' is replaced with 's' in owner's level permissions that means suid is applied on that file)

- r w s r w x r w x <file name> (here 's' is called set uid or suid)

Example : # chmod u+s /usr/sbin/init 6 (then any user can restart the system using this command #init 6)

# chmod u+s /sbin/fdisk (then any user can run the fdisk command)

# strings <command name> (to read the binary language of the command ie., the string command converts the binary language into human readable language)

# strings mkfs (to read the mkfs command's binary language into human readable language)

\* Normally set uid (suid) permission will be given on scripting files only.

=====

**\*Explain about set gid (sgid)?\***

If we plan to allow all the users of one group to get the group ownership permissions then we go for set gid (sgid). It can be applied for group level and is applicable on directories only.

Example: # chmod g+s <directory name> (to set the sgid on that directory)

# chmod g-s <directory name> (to remove the sgid from that directory)

=====

**\*Explain about sticky bit?\***

It protects the data from other users when all the users having full permissions on one directory.

It can be applied on others level and applicable for directories only.

Example : # chmod o+t <directory name> (to set the sticky bit permission on that directory)

# ls -ld <directory name> rwxrwxrwt <directory name> (where 't' is called the sticky bit)

Ps Arshad: **\*How many process are run generally on Linux and explain them?\***

There are generally three types of processes that run on Linux. They are,

**\*(i) Interactive Processes\***

**\*(ii) System Process or daemon\***

**\*(iii) Automatic or batch\***

**\*Interactive Processes :\***

Interactive processes are those processes that are invoked by a user and can interact with the user.

For example #vi or #vim are the interactive processes. Interactive processes may be run in foreground or background. The foreground process is the process that we are currently interacting with and is using the terminal as its stdin (standard input) and stdout (standard output). The background process is not interacting with the user and can be in one of two states, ie., paused or running.

**\*System Processes or daemons :\***

Daemon is refer to processes that are running on the computer and provides services but do not interact with the console. Most server software is implemented as a daemon.

For example:- Apache, samba, sshd are the daemons. Any process can become a daemon as long as it is run in the background and does not interact with the user.

\*Automatic processes :\*

Automatic processes are not connected to a terminal and these are queued into a spooler area where they wait to be executed on a FIFO (First In - First Out) basis. Such tasks can be executed using one of two criteria.

At certain date and time : done using the "at" command.

When the total system load is low enough to accept extra jobs : done using the "cron" command. By default tasks are put in a queue where they wait to be executed until the system load is lower than 0.8 and cron job processing is also used for optimizing system performance