

Big Data



Big Data Engineering with Hadoop & Spark

Assignment on Introduction to Big Data

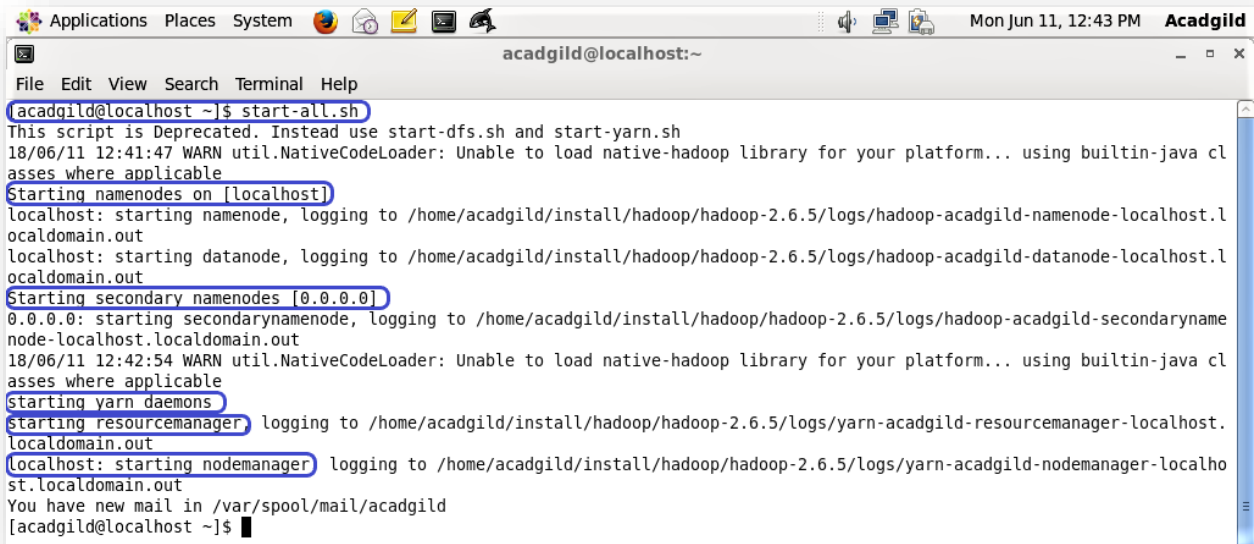


Session 1: Assignment 1.1

This assignment is aimed at consolidating the concepts that was learned during the opening session of the course.

Task 1:

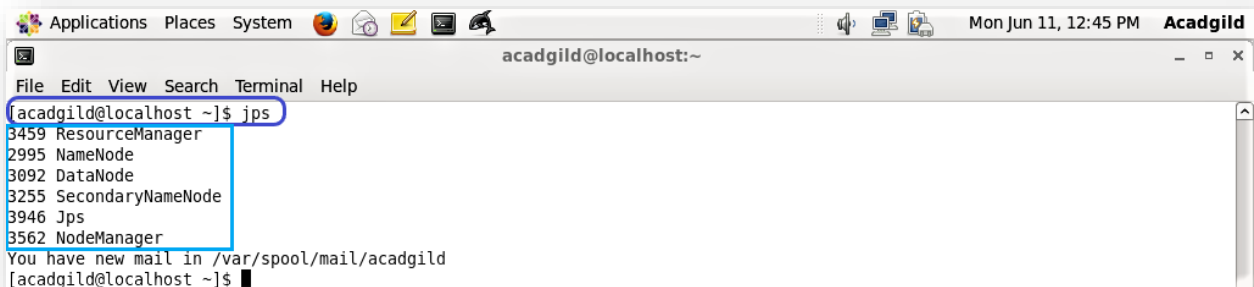
Start Hadoop single node on AcadGild VM. The command is “*start-all.sh*”.

A terminal window titled 'acadgild@localhost:~' showing the execution of the 'start-all.sh' script. The script starts with a deprecation warning and then proceeds to start the Hadoop NameNode, DataNode, SecondaryNameNode, Yarn daemons, ResourceManager, and NodeManager. Each component's startup is logged to a specific file in the /home/acadgild/install/hadoop/hadoop-2.6.5/logs directory. The terminal output is as follows:

```
acadgild@localhost ~]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/06/11 12:41:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.l
ocaldomain.out
localhost: starting datanode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.l
ocaldomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondaryname
node-localhost.localdomain.out
18/06/11 12:42:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
starting yarn daemons
starting resourcemanager logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.
localdomain.out
localhost: starting nodemanager logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 2:

Run a *JPS* command to see if all Hadoop daemons are running.

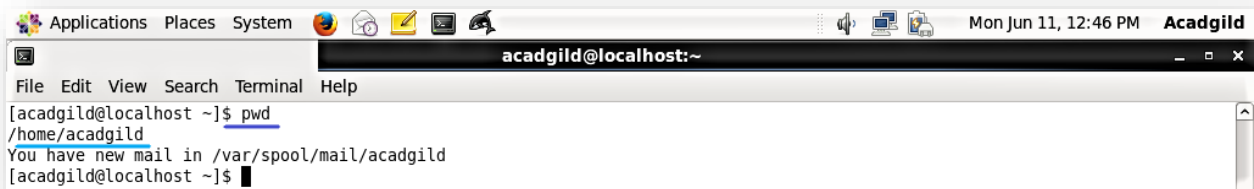
A terminal window titled 'acadgild@localhost:~' showing the output of the 'jps' command. The output lists five processes: ResourceManager (PID 3459), NameNode (PID 2995), DataNode (PID 3092), SecondaryNameNode (PID 3255), and Jps (PID 3946). The terminal output is as follows:

```
acadgild@localhost ~]$ jps
3459 ResourceManager
2995 NameNode
3092 DataNode
3255 SecondaryNameNode
3946 Jps
3562 NodeManager
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

Task 3:

Run few Unix commands.

1. **pwd:** When you first open the terminal, you are in the home directory of your user. To know which directory you are in, you can use the “*pwd*” command. It gives us the absolute path, which means the path that starts from the root. The root is the base of the Linux file system. It is denoted by a forward slash(/). The user directory is usually something like “/home/username”. Here the username is “acadgild”.

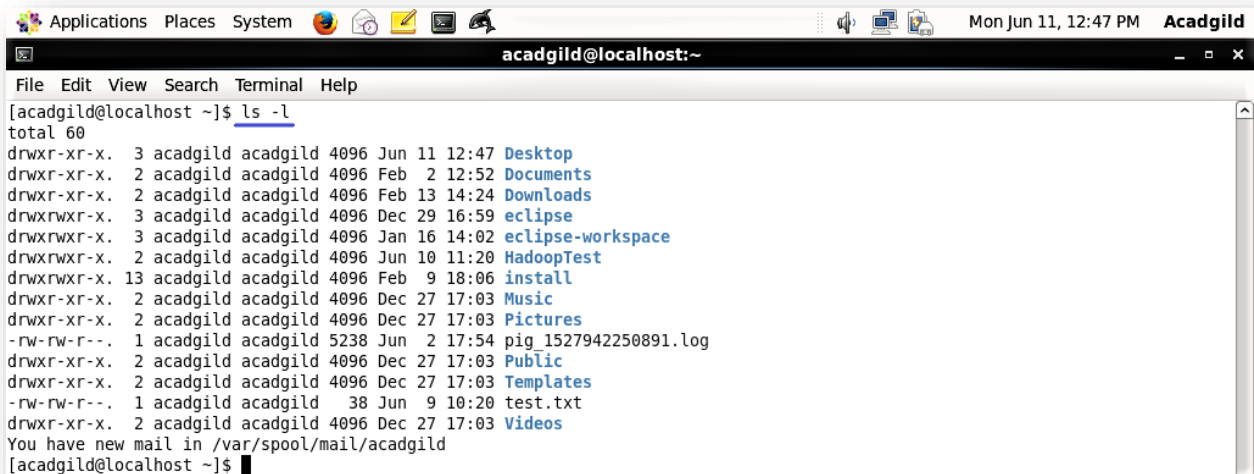


```

Applications Places System  Mon Jun 11, 12:46 PM  Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ pwd
/home/acadgild
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

2. **ls:** Use the “*ls*” command to know what files are in the directory you are in. “*-l*” is a quick and easy way to list a file's permissions are with the long listing option of the *ls* command.

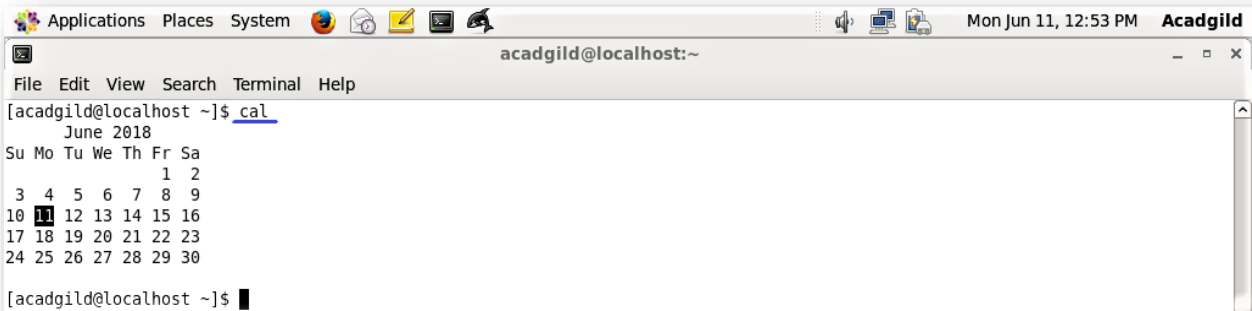


```

Applications Places System  Mon Jun 11, 12:47 PM  Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ ls -l
total 60
drwxr-xr-x. 3 acadgild acadgild 4096 Jun 11 12:47 Desktop
drwxr-xr-x. 2 acadgild acadgild 4096 Feb  2 12:52 Documents
drwxr-xr-x. 2 acadgild acadgild 4096 Feb 13 14:24 Downloads
drwxrwxr-x. 3 acadgild acadgild 4096 Dec 29 16:59 eclipse
drwxrwxr-x. 3 acadgild acadgild 4096 Jan 16 14:02 eclipse-workspace
drwxrwxr-x. 2 acadgild acadgild 4096 Jun 10 11:20 HadoopTest
drwxrwxr-x. 13 acadgild acadgild 4096 Feb  9 18:06 install
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Music
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Pictures
-rw-rw-r--. 1 acadgild acadgild 5238 Jun  2 17:54 pig_1527942250891.log
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Public
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Templates
-rw-rw-r--. 1 acadgild acadgild  38 Jun  9 10:20 test.txt
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Videos
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

3. `cal`: The "`cal`" command displays a simple, formatted calendar in your terminal.

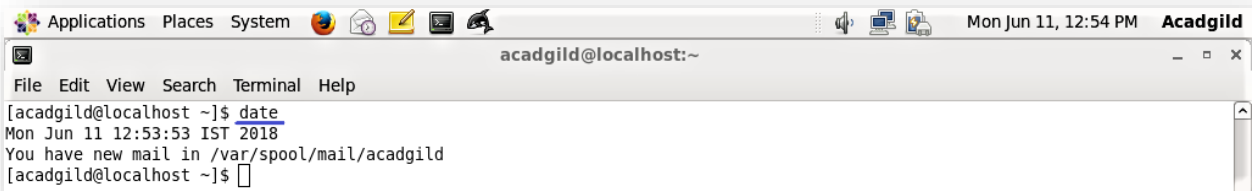


A terminal window titled 'acadgild@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar. The command `cal` has been executed, displaying a calendar for June 2018. The calendar shows days of the week (Su, Mo, Tu, We, Th, Fr, Sa) and dates (1-30). The current date, June 11, is highlighted. The prompt `[acadgild@localhost ~]$` is visible at the bottom.

```
[acadgild@localhost ~]$ cal
      June 2018
Su Mo Tu We Th Fr Sa
                1  2
 3  4  5  6  7  8  9
10 11 12 13 14 15 16
17 18 19 20 21 22 23
24 25 26 27 28 29 30

[acadgild@localhost ~]$
```

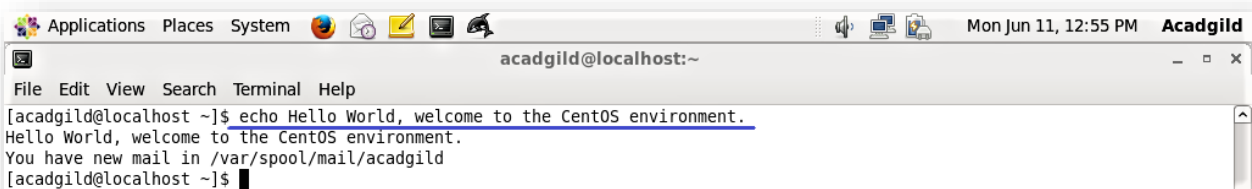
4. `date`: The "`date`" command is used to display the system date and time.



A terminal window titled 'acadgild@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar. The command `date` has been executed, displaying the system date and time: 'Mon Jun 11 12:53:53 IST 2018'. A system message 'You have new mail in /var/spool/mail/acadgild' is also visible. The prompt `[acadgild@localhost ~]$` is visible at the bottom.

```
[acadgild@localhost ~]$ date
Mon Jun 11 12:53:53 IST 2018
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

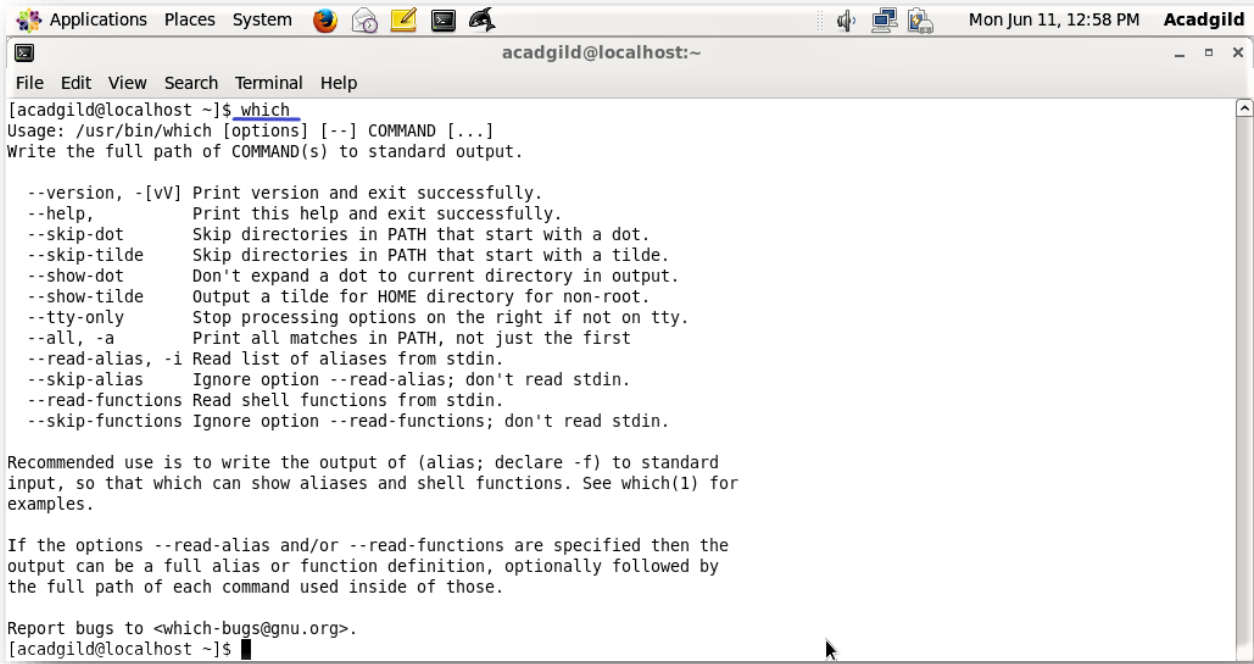
5. `echo`: The "`echo`" command is used to input a line of text and display on standard output.



A terminal window titled 'acadgild@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help) and a toolbar. The command `echo Hello World, welcome to the CentOS environment.` has been executed, displaying the same text on the next line. A system message 'You have new mail in /var/spool/mail/acadgild' is also visible. The prompt `[acadgild@localhost ~]$` is visible at the bottom.

```
[acadgild@localhost ~]$ echo Hello World, welcome to the CentOS environment.
Hello World, welcome to the CentOS environment.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

6. **which:** The “*which*” command is used to identify the location of executables. The command takes one or more arguments; for each of these arguments, it prints the full path of the executable to standard output that would have been executed if this argument had been entered into the shell.



```

Applications Places System [Icons] [System Status] Mon Jun 11, 12:58 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ which
Usage: /usr/bin/which [options] [--] COMMAND [...]
Write the full path of COMMAND(s) to standard output.

--version, -[vV] Print version and exit successfully.
--help,          Print this help and exit successfully.
--skip-dot       Skip directories in PATH that start with a dot.
--skip-tilde     Skip directories in PATH that start with a tilde.
--show-dot       Don't expand a dot to current directory in output.
--show-tilde     Output a tilde for HOME directory for non-root.
--tty-only       Stop processing options on the right if not on tty.
--all, -a        Print all matches in PATH, not just the first
--read-alias, -i Read list of aliases from stdin.
--skip-alias     Ignore option --read-alias; don't read stdin.
--read-functions Read shell functions from stdin.
--skip-functions Ignore option --read-functions; don't read stdin.

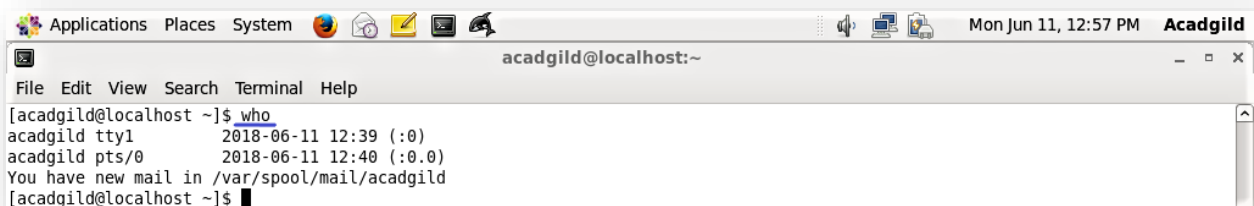
Recommended use is to write the output of (alias; declare -f) to standard
input, so that which can show aliases and shell functions. See which(1) for
examples.

If the options --read-alias and/or --read-functions are specified then the
output can be a full alias or function definition, optionally followed by
the full path of each command used inside of those.

Report bugs to <which-bugs@gnu.org>.
[acadgild@localhost ~]$

```

7. **who:** The “*who*” command is used to get information about currently logged in user on to system.



```

Applications Places System [Icons] [System Status] Mon Jun 11, 12:57 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ who
acadgild tty1      2018-06-11 12:39 (:0)
acadgild pts/0    2018-06-11 12:40 (:0.0)
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```


8. ping: The “ping” command is used to check your connection to a server. When you type in, for example, “**ping google.com**”, it checks if it can connect to the server and come back. It measures this round-trip time and gives you the details about it. The use of this command for simple users like us is to check your internet connection. If it pings the Google server (in this case), you can confirm that your internet connection is active!

```

[acadgild@localhost ~]$ ping google.com
PING google.com (172.217.31.206) 56(84) bytes of data.
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=1 ttl=53 time=111 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=2 ttl=53 time=106 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=3 ttl=53 time=136 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=4 ttl=53 time=123 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=5 ttl=53 time=122 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=6 ttl=53 time=120 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=7 ttl=53 time=120 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=8 ttl=53 time=133 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=9 ttl=53 time=116 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=10 ttl=53 time=145 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=11 ttl=53 time=122 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=12 ttl=53 time=121 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=13 ttl=53 time=117 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=14 ttl=53 time=115 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=15 ttl=53 time=123 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=16 ttl=53 time=111 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=17 ttl=53 time=334 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=18 ttl=53 time=117 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=19 ttl=53 time=124 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=20 ttl=53 time=122 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=21 ttl=53 time=121 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=22 ttl=53 time=184 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=23 ttl=53 time=129 ms
64 bytes from maa03s28-in-f14.1e100.net (172.217.31.206): icmp_seq=24 ttl=53 time=128 ms
^C
--- google.com ping statistics ---
24 packets transmitted, 24 received, 0% packet loss, time 23254ms
rtt min/avg/max/mdev = 106.301/133.859/334.369/44.384 ms
[acadgild@localhost ~]$

```

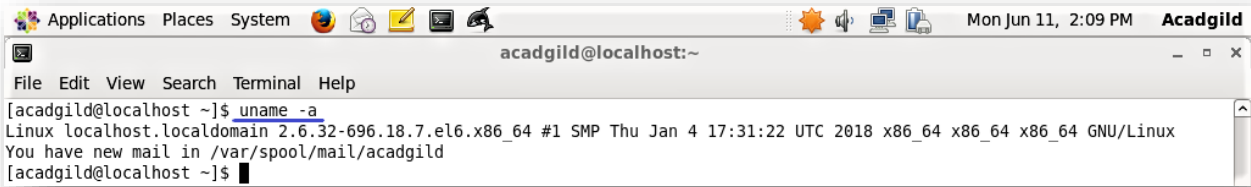
9. hostname: The “hostname” command is used to know your name in your host or network.

```

[acadgild@localhost ~]$ hostname
localhost.localdomain
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

- 10.uname:** The “*uname*” command shows the information about the system your Linux distro is running. Using the command “*uname -a*” prints most of the information about the system. This prints the kernel release date, version, processor type, etc.

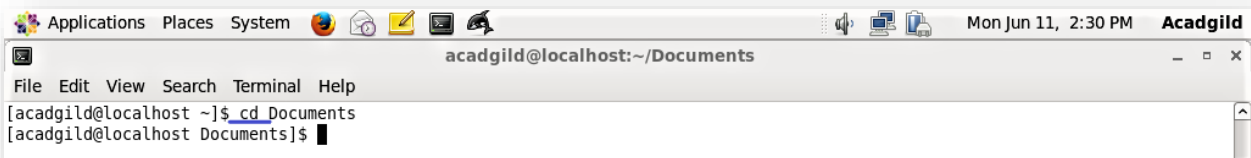


```

Applications Places System acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ uname -a
Linux localhost.localdomain 2.6.32-696.18.7.el6.x86_64 #1 SMP Thu Jan 4 17:31:22 UTC 2018 x86_64 x86_64 x86_64 GNU/Linux
[acadgild@localhost ~]$

```

- 11.cd:** The “*cd*” command is used to go to a directory.

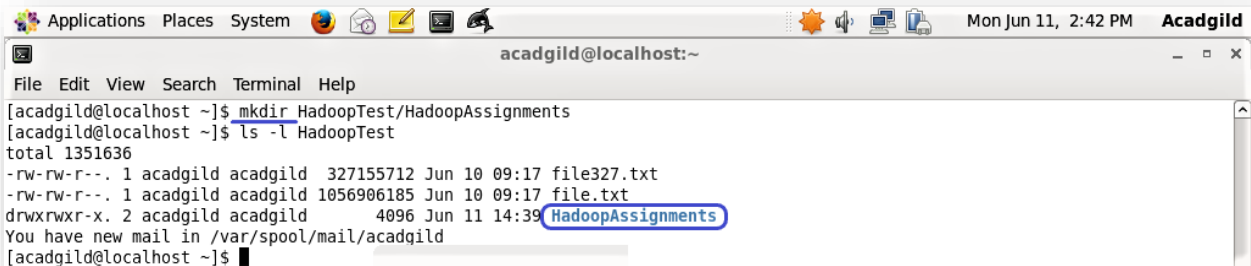


```

Applications Places System acadgild@localhost:~/Documents
File Edit View Search Terminal Help
[acadgild@localhost ~]$ cd Documents
[acadgild@localhost Documents]$

```

- 12.mkdir:** The “*mkdir*” command is used to create a folder or a directory.

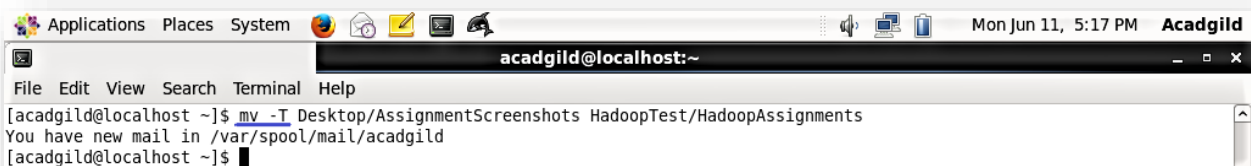


```

Applications Places System acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ mkdir HadoopTest/HadoopAssignments
[acadgild@localhost ~]$ ls -l HadoopTest
total 1351636
-rw-rw-r--. 1 acadgild acadgild 327155712 Jun 10 09:17 file327.txt
-rw-rw-r--. 1 acadgild acadgild 1056906185 Jun 10 09:17 file.txt
drwxrwxr-x. 2 acadgild acadgild 4096 Jun 11 14:39 HadoopAssignments
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

```

- 13.mv:** The “*mv*” command is used to move files through the command line.



```

Applications Places System acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ mv -T Desktop/AssignmentScreenshots HadoopTest/HadoopAssignments
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$

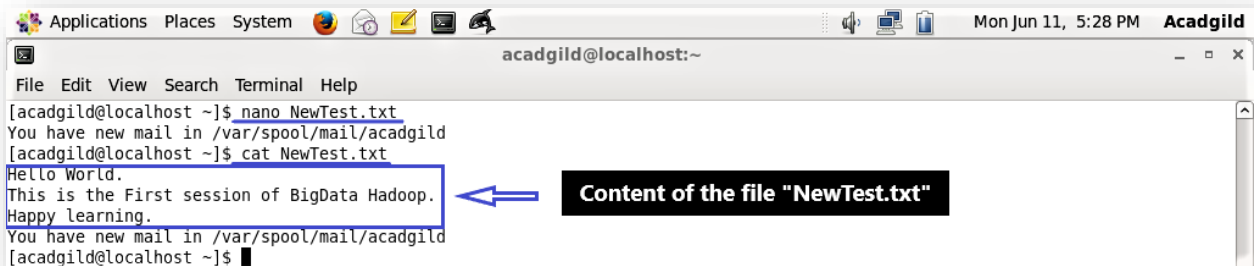
```


Task 4:

Create a file from the terminal using “nano editor” (example: nano test.txt) and add some content in it. “Cat” it to see if the content is saved.

Solution:

- A file named “NewTest.txt” was created using “nano editor” with the help of “nano” command on command line.
- Text content was added on to the editor and was saved.
- From command line “cat” command was used to check in the text content was saved on the file or not.



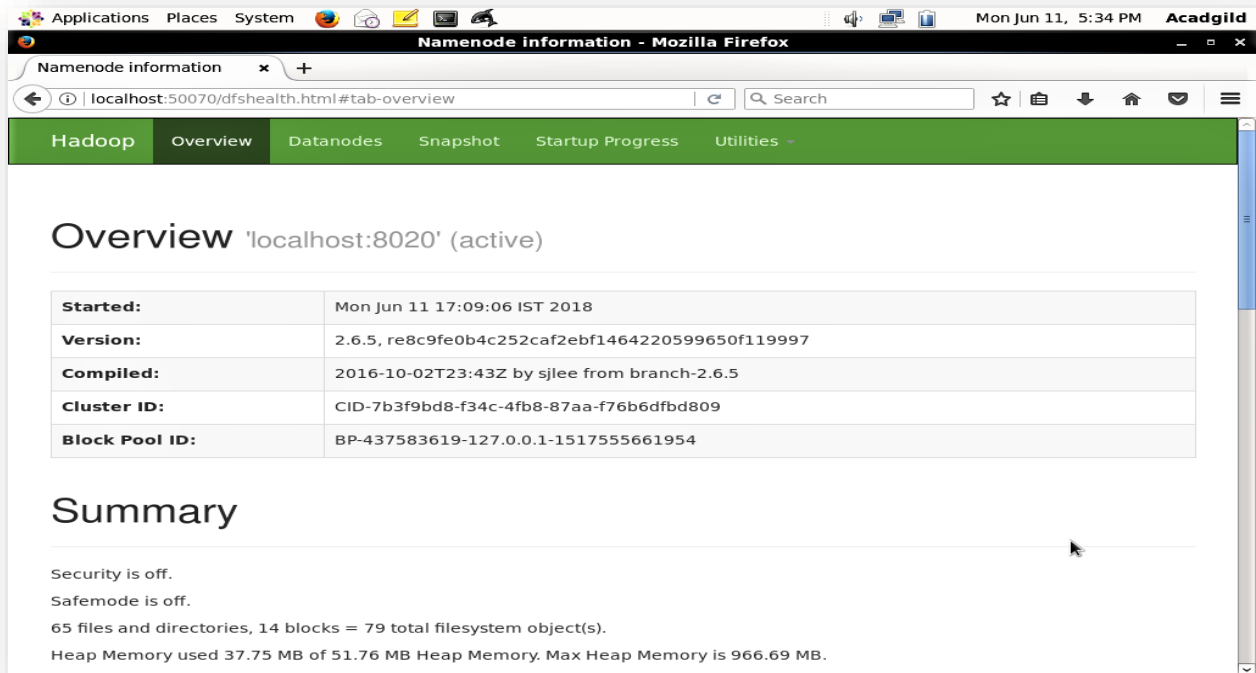
The screenshot shows a terminal window titled "acadgild@localhost:~". The terminal output is as follows:

```
[acadgild@localhost ~]$ nano NewTest.txt
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat NewTest.txt
Hello World.
This is the First session of BigData Hadoop.
Happy learning.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

A blue box highlights the content of the file, and a blue arrow points from a black box labeled "Content of the file 'NewTest.txt'" to the highlighted text.

Task 5:

Open the hdfs web page by typing *localhost:50070* in the browser. Check all the details of the HDFS.

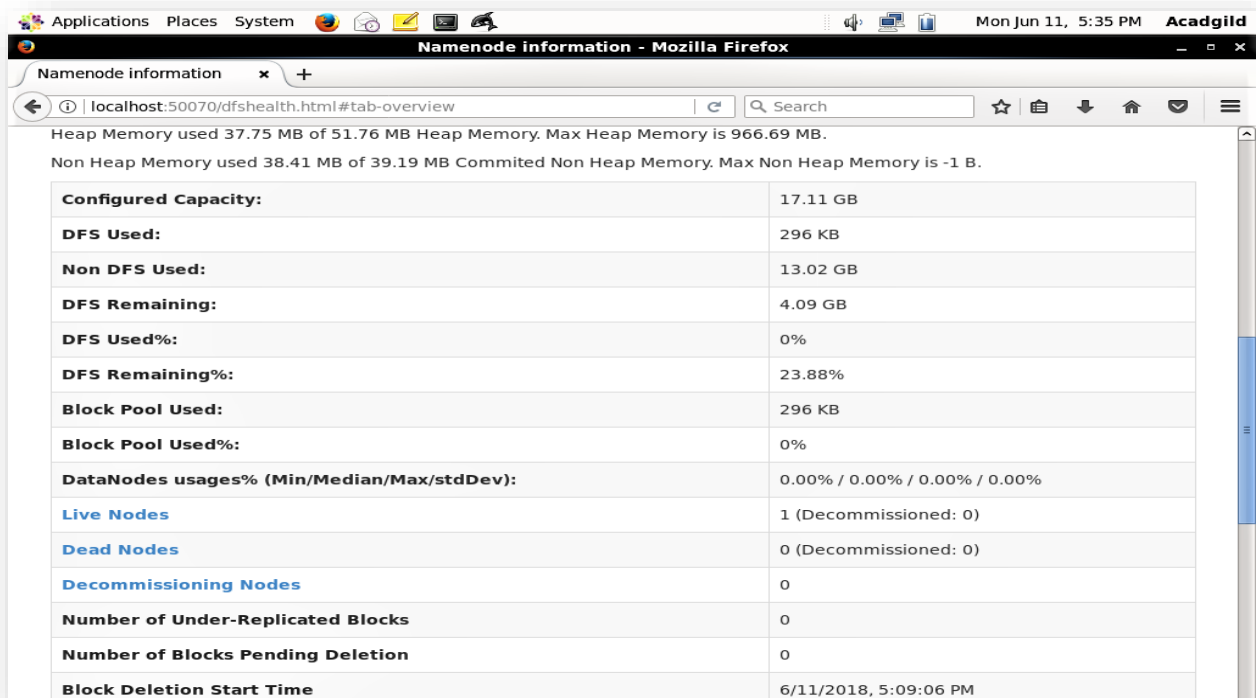


The screenshot shows the 'Overview' tab of the Hadoop NameNode web interface. The browser address bar shows `localhost:50070/dfshealth.html#tab-overview`. The page title is 'Namenode information - Mozilla Firefox'. The navigation bar includes 'Hadoop', 'Overview', 'Datanodes', 'Snapshot', 'Startup Progress', and 'Utilities'. The main content area is titled 'Overview 'localhost:8020' (active)'. Below the title is a table with the following information:

Started:	Mon Jun 11 17:09:06 IST 2018
Version:	2.6.5, re8c9fe0b4c252caf2ebf1464220599650f119997
Compiled:	2016-10-02T23:43Z by sjlee from branch-2.6.5
Cluster ID:	CID-7b3f9bd8-f34c-4fb8-87aa-f76b6dfbd809
Block Pool ID:	BP-437583619-127.0.0.1-1517555661954

Below the table is a 'Summary' section with the following text:

Security is off.
 Safemode is off.
 65 files and directories, 14 blocks = 79 total filesystem object(s).
 Heap Memory used 37.75 MB of 51.76 MB Heap Memory. Max Heap Memory is 966.69 MB.



The screenshot shows the 'Overview' tab of the Hadoop NameNode web interface, displaying detailed memory and capacity information. The browser address bar shows `localhost:50070/dfshealth.html#tab-overview`. The page title is 'Namenode information - Mozilla Firefox'. The navigation bar includes 'Hadoop', 'Overview', 'Datanodes', 'Snapshot', 'Startup Progress', and 'Utilities'. The main content area is titled 'Overview 'localhost:8020' (active)'. Below the title is a table with the following information:

Configured Capacity:	17.11 GB
DFS Used:	296 KB
Non DFS Used:	13.02 GB
DFS Remaining:	4.09 GB
DFS Used%:	0%
DFS Remaining%:	23.88%
Block Pool Used:	296 KB
Block Pool Used%:	0%
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	6/11/2018, 5:09:06 PM

Applications Places System Mon Jun 11, 5:35 PM Acadgild

Namenode information - Mozilla Firefox

localhost:50070/dfshealth.html#tab-overview

Number of Blocks Pending Deletion	0
Block Deletion Start Time	6/11/2018, 5:09:06 PM

NameNode Journal Status

Current transaction ID: 470

Journal Manager	State
FileJournalManager(root=/home/acadgild/install/data/dfs/name)	EditLogFileOutputStream(/home/acadgild/install/data/dfs/name/current/edits_inprogress_00000000000000000470)

NameNode Storage

Storage Directory	Type	State
/home/acadgild/install/data/dfs/name	IMAGE_AND_EDITS	Active

Hadoop, 2016. Legacy

Applications Places System Mon Jun 11, 5:36 PM Acadgild

Namenode information - Mozilla Firefox

localhost:50070/dfshealth.html#tab-datanode

Hadoop Overview Datanodes Snapshot Startup Progress Utilities






Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
localhost (127.0.0.1:50010)	1	In Service	17.11 GB	296 KB	13.02 GB	4.09 GB	14	296 KB (0%)	0	2.6.5

Decomissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
------	--------------	-------------------------	------------------------------	---

Applications Places System      Mon Jun 11, 5:37 PM Acadgild

Namenode information - Mozilla Firefox

Namenode information x +

localhost:50070/dfshealth.html#tab-snapshot






Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Snapshot Summary

Snapshottable directories: 0

Path	Snapshot Number	Snapshot Quota	Modification Time	Permission	Owner	Group
Snapshotted directories: 0						
Snapshot ID	Snapshot Directory	Modification Time				

Hadoop, 2016.

Applications Places System      Mon Jun 11, 5:38 PM Acadgild

Namenode information - Mozilla Firefox

Namenode information x +

localhost:50070/dfshealth.html#tab-startup-progress

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Startup Progress

Elapsed Time: 54 sec, Percent Complete: 100%

Phase	Completion	Elapsed Time
Loading fsimage /home/acadgild/install/data/dfs/name/current/fsimage_00000000000000000468 4.76 KB	100%	4 sec
inodes (0/0)	100%	
delegation tokens (0/0)	100%	
cache pools (0/0)	100%	
Loading edits	100%	0 sec
/home/acadgild/install/data/dfs/name/current/edits_00000000000000000469-00000000000000000469 1 MB (1/1)	100%	
Saving checkpoint	100%	1 sec
inodes /home/acadgild/install/data/dfs/name/current/fsimage.ckpt_00000000000000000469 (0/0)	100%	

Namenode information

fsimage_0000000000000000468 4.76 KB

inodes (0/0)	100%
delegation tokens (0/0)	100%
cache pools (0/0)	100%
Loading edits	100% 0 sec
/home/acadgild/install/data/dfs/name/current /edits_0000000000000000469-0000000000000000469 1 MB (1/1)	100%
Saving checkpoint	100% 1 sec
inodes /home/acadgild/install/data/dfs/name/current/fsimage.ckpt_0000000000000000469 (0/0)	100%
delegation tokens /home/acadgild/install/data/dfs/name/current /fsimage.ckpt_0000000000000000469 (0/0)	100%
cache pools /home/acadgild/install/data/dfs/name/current /fsimage.ckpt_0000000000000000469 (0/0)	100%
Safe mode	100% 41 sec
awaiting reported blocks (14/14)	100%

Hadoop, 2016.

Legacy

Browsing HDFS

localhost:50070/explorer.html#/

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

Search

Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	acadgild	supergroup	0 B	0	0 B	hbase
drwxr-xr-x	acadgild	supergroup	0 B	0	0 B	sqoopout111
-rw-r--r--	acadgild	supergroup	36 B	1	128 MB	test.txt
drwxrwx---	acadgild	supergroup	0 B	0	0 B	tmp
drwxr-xr-x	acadgild	supergroup	0 B	0	0 B	user

Hadoop, 2016.