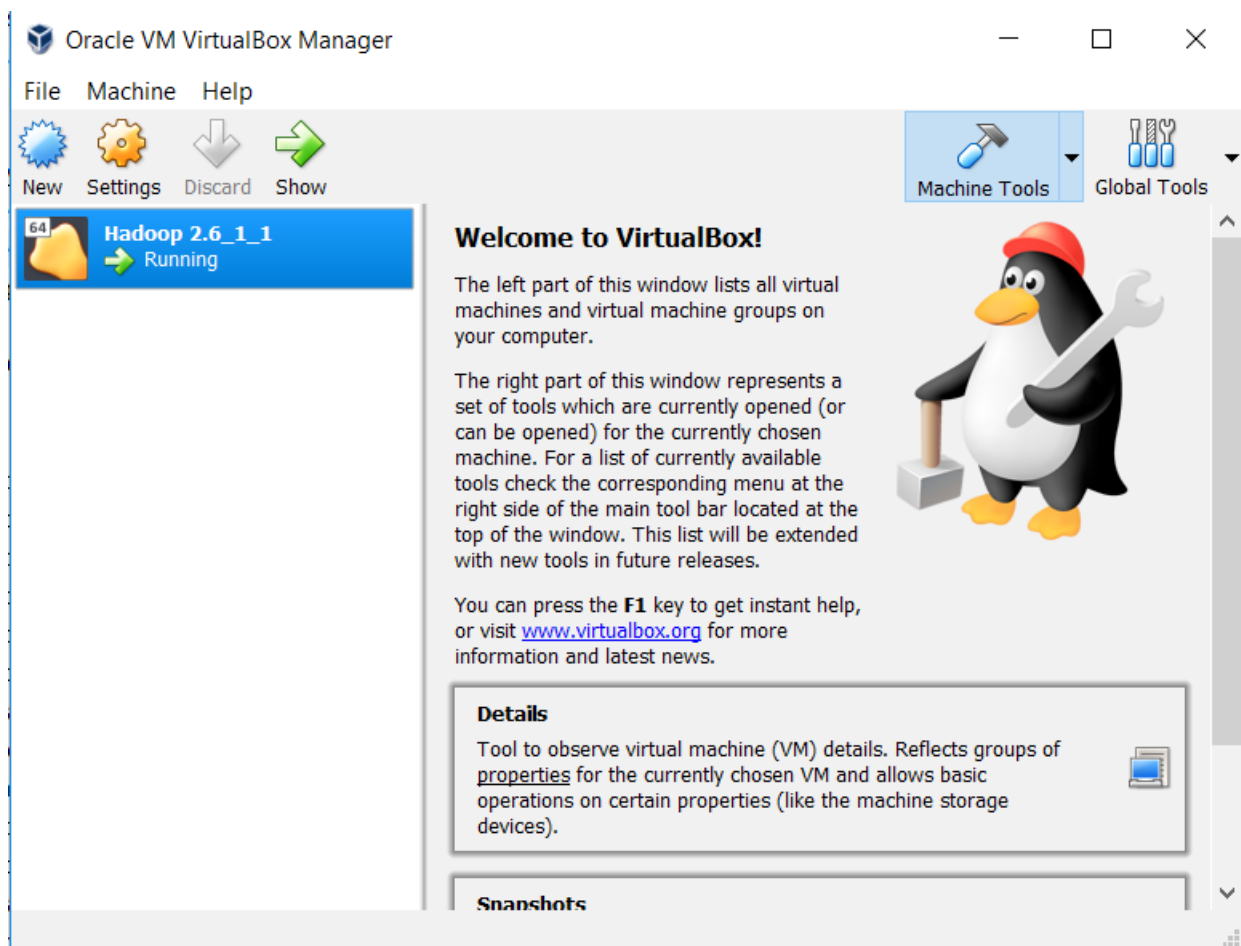


## BIG DATA ENGINEERING FOR HADOOP & SPARK TRAINING

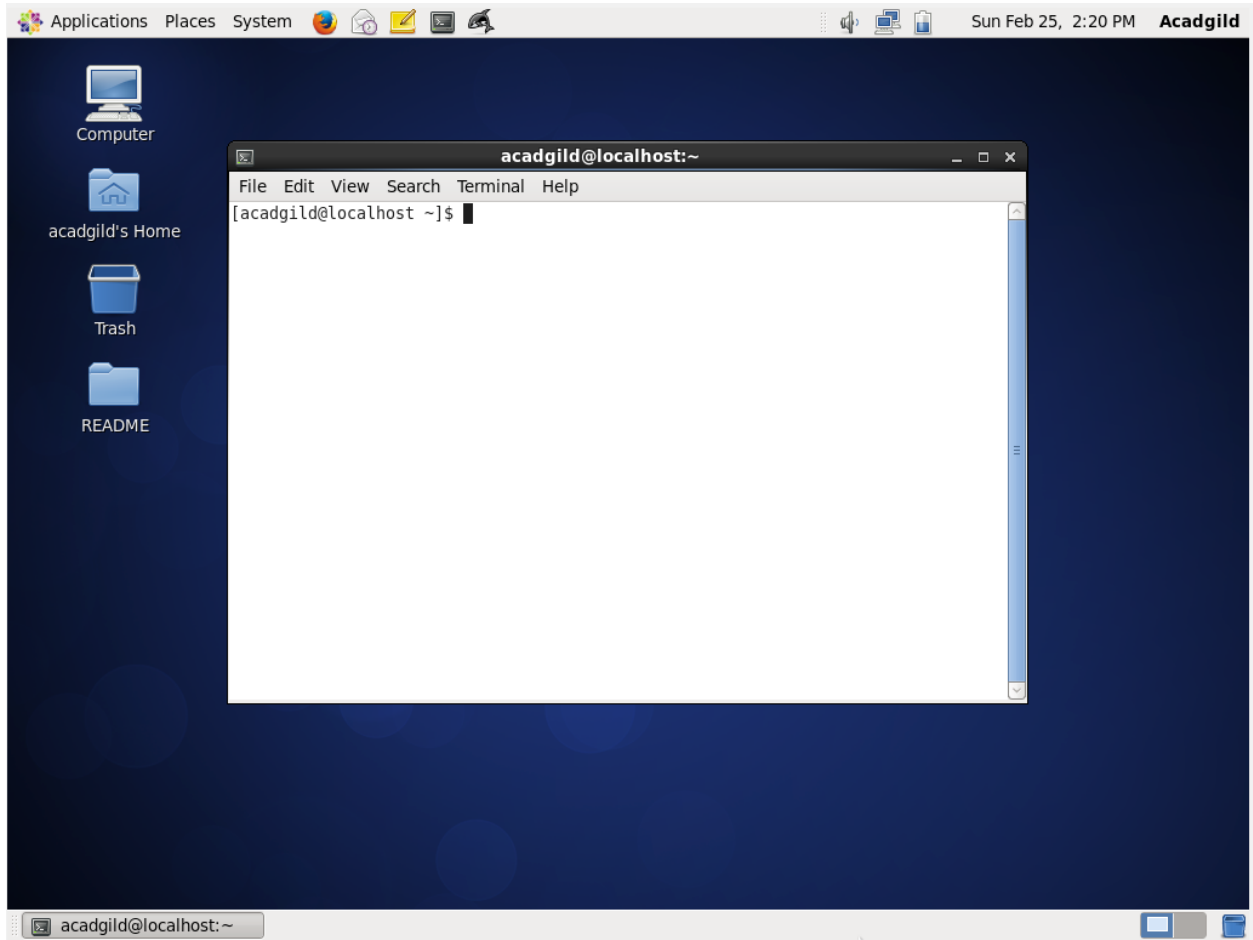
### ASSIGNMENT 1

#### 1) Start Hadoop Single Node VM

i) First open Virtual Box and then Import Acadgild appliance and then start the VM using default settings



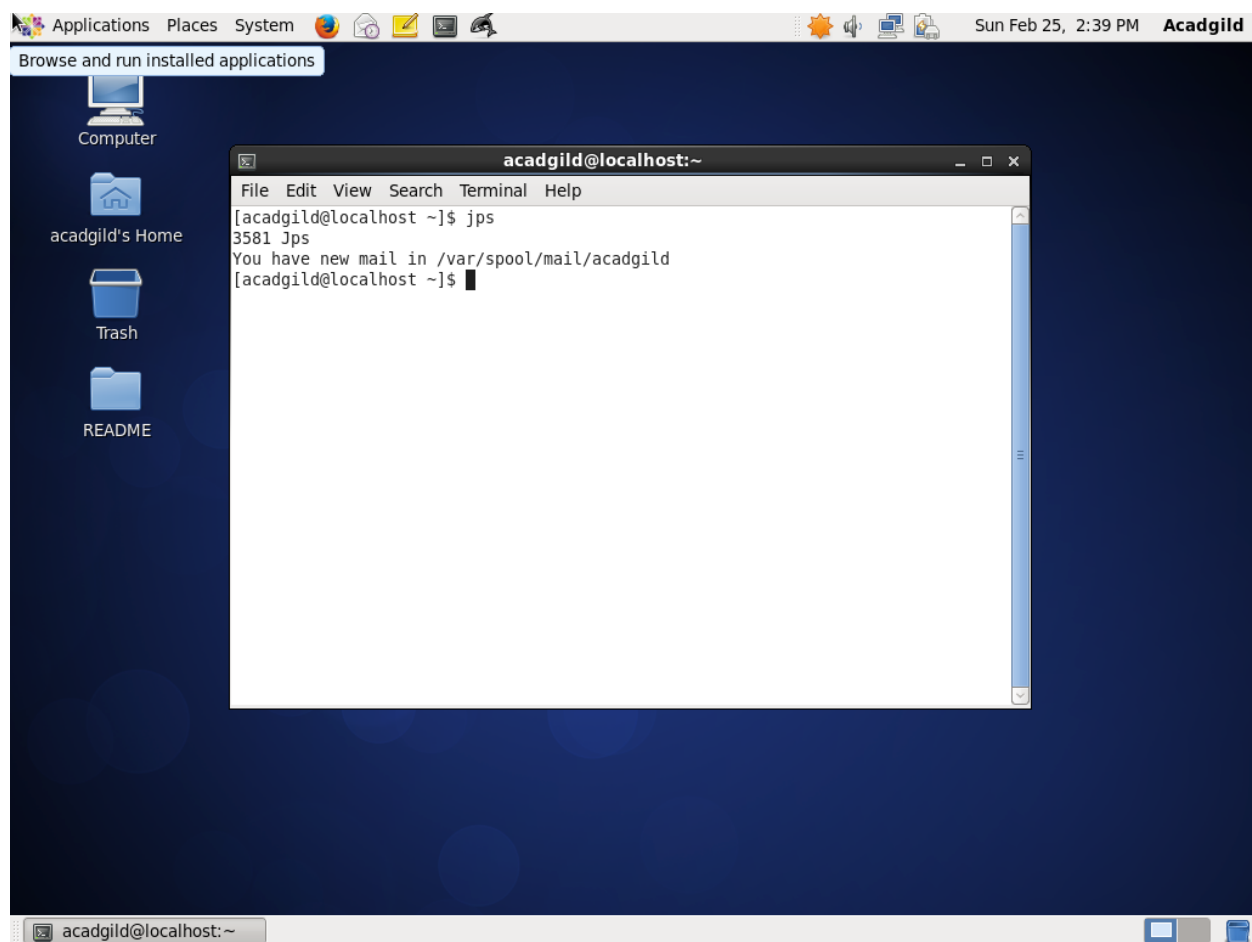
ii) Now Start the VM and provide password for Acadgild user and then open terminal

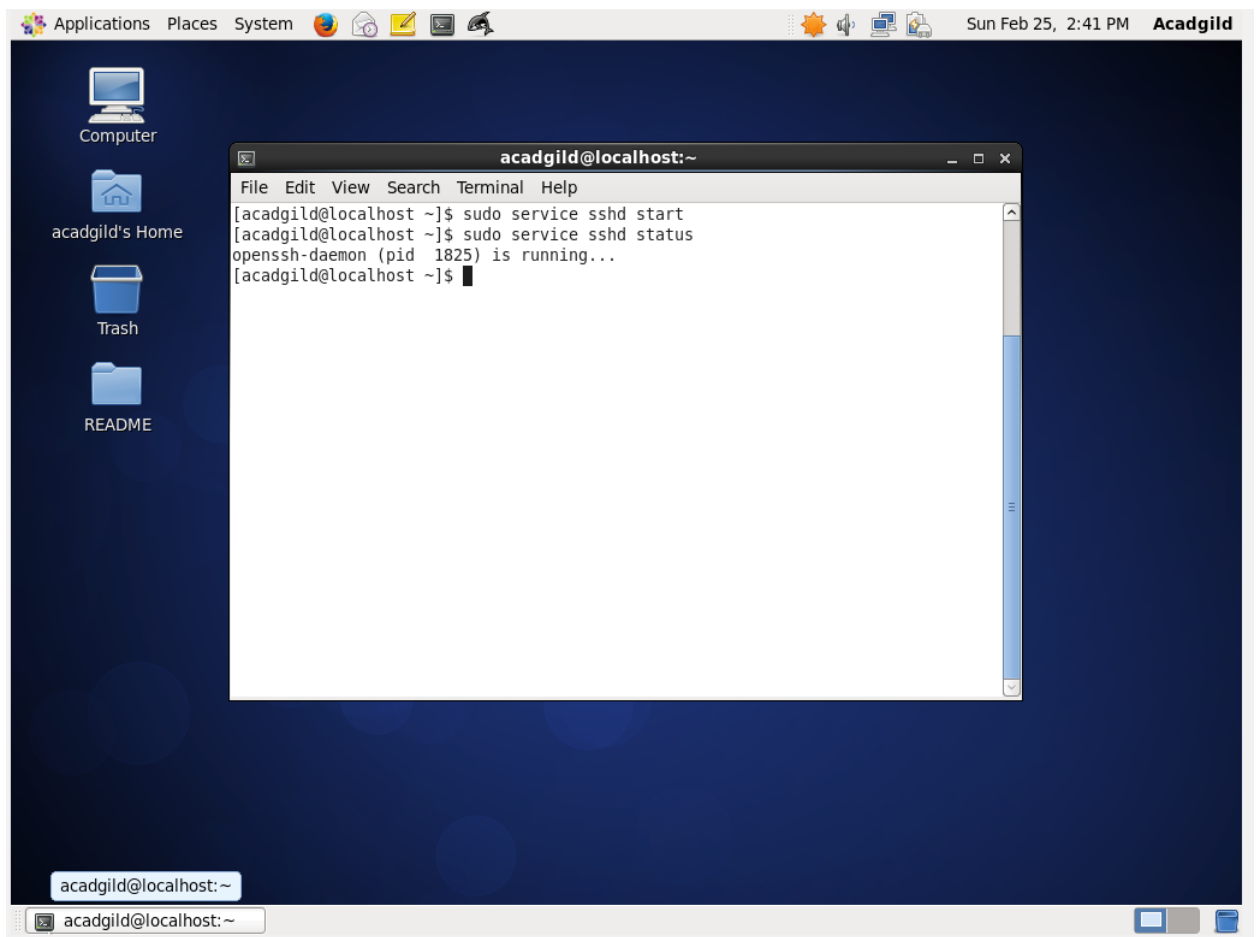


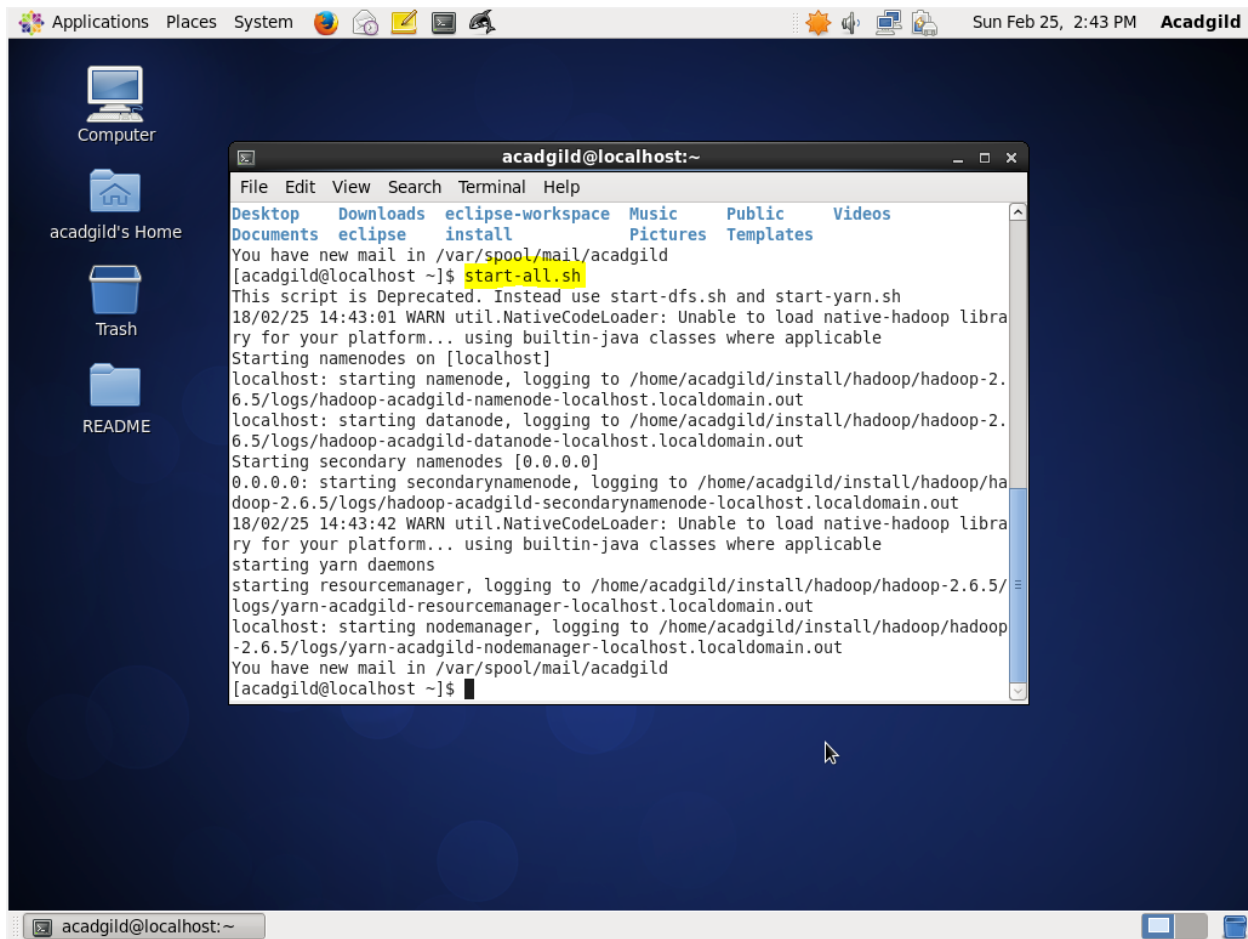
iii) Type “jps” command to check any deamons are running if not use “start-all.sh” to start the deamons

jps(Java Virtual Machine Process Status ) is a command which dispays all the hadoop processes that are running on JVM

start-all.sh script will start all the Hadoop deamons



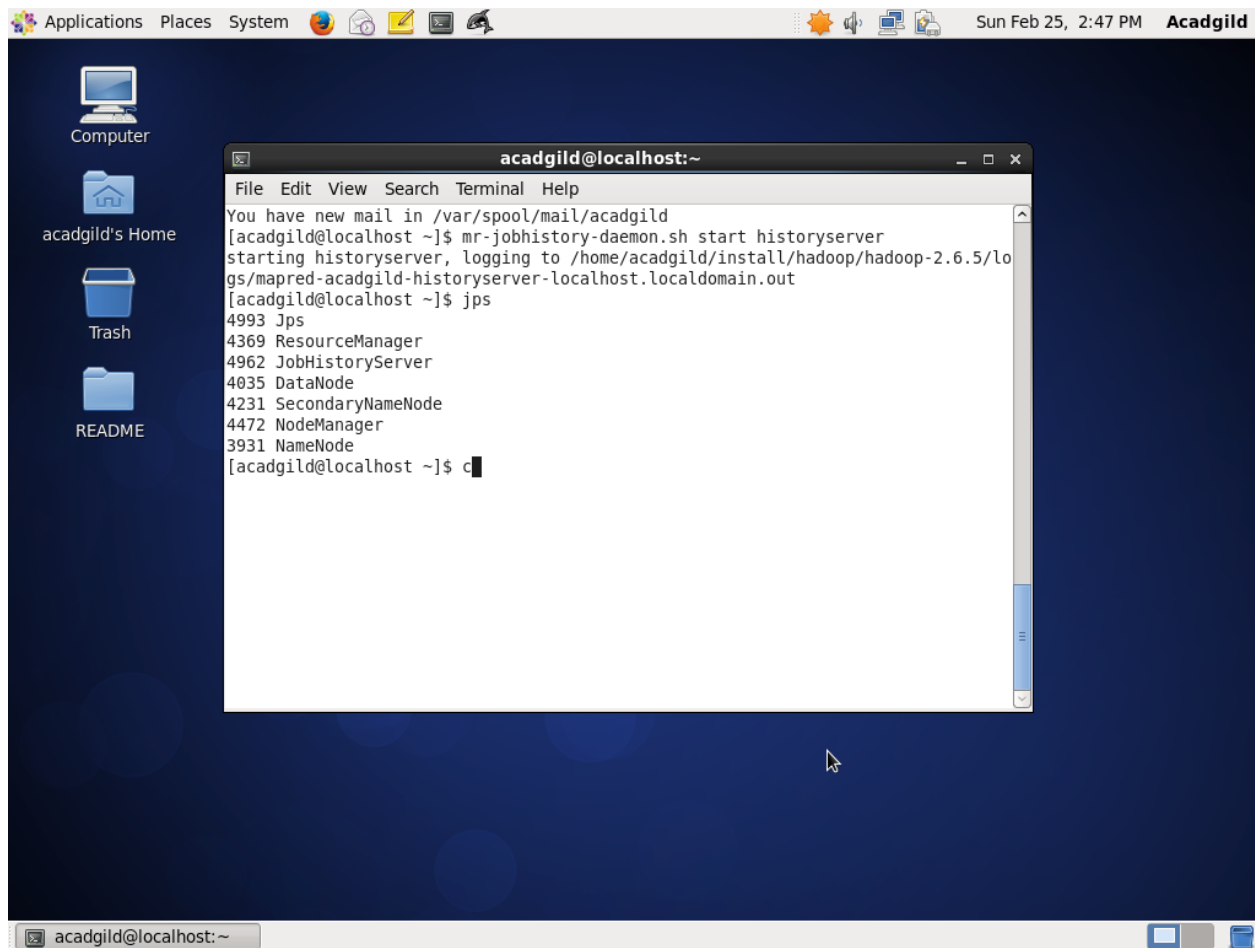




start-all.sh script will start **namenodes,secondary namenodes, resource manager and node manager** hadoop deamons.

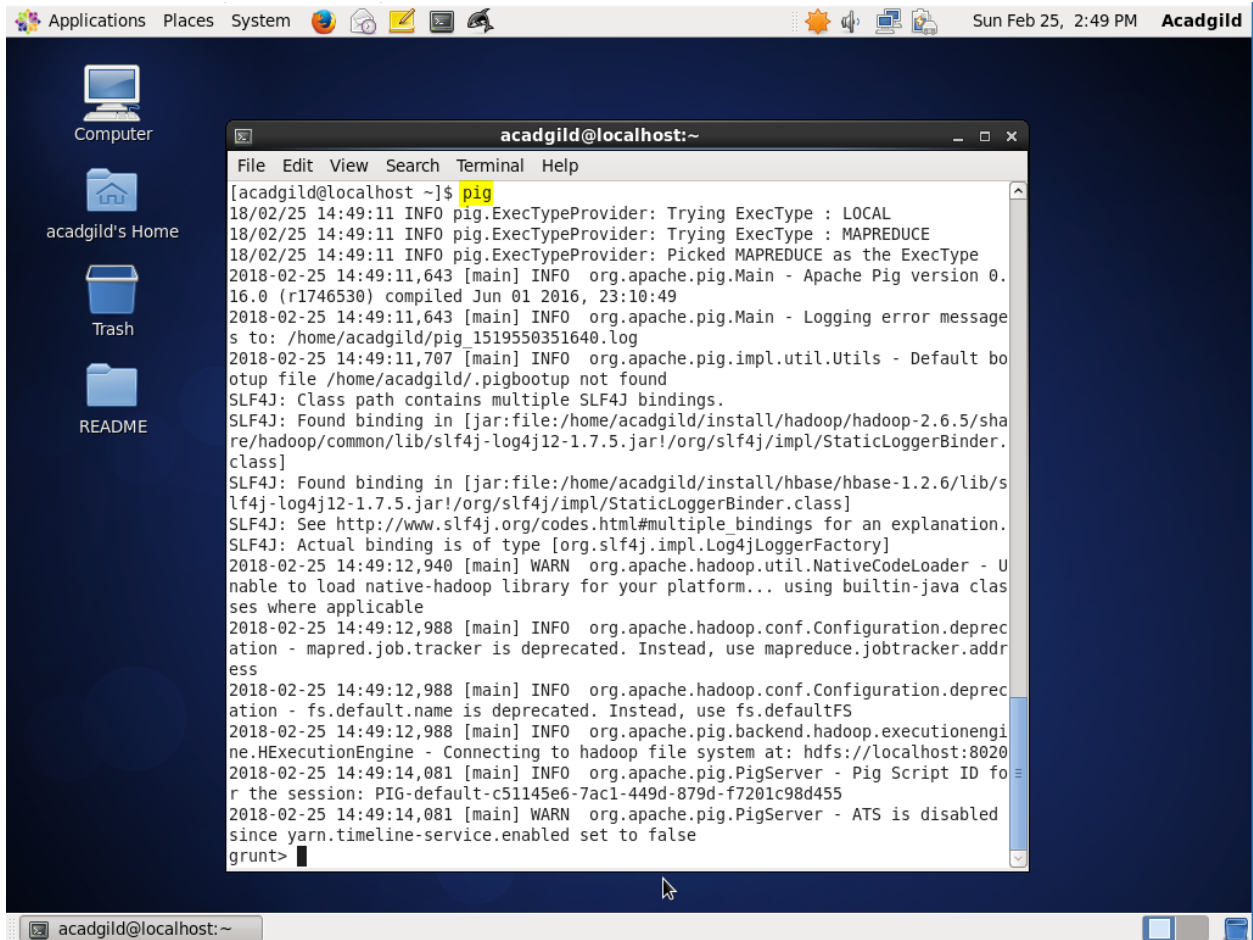
iv )for starting historyserver use the command “mr-jobhistory-daemon.sh start historyserver”

Now type “jps” command to check all the deamons that are started in previous command



v) **Pig** is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java.

### Starting Pig in HDFS mode

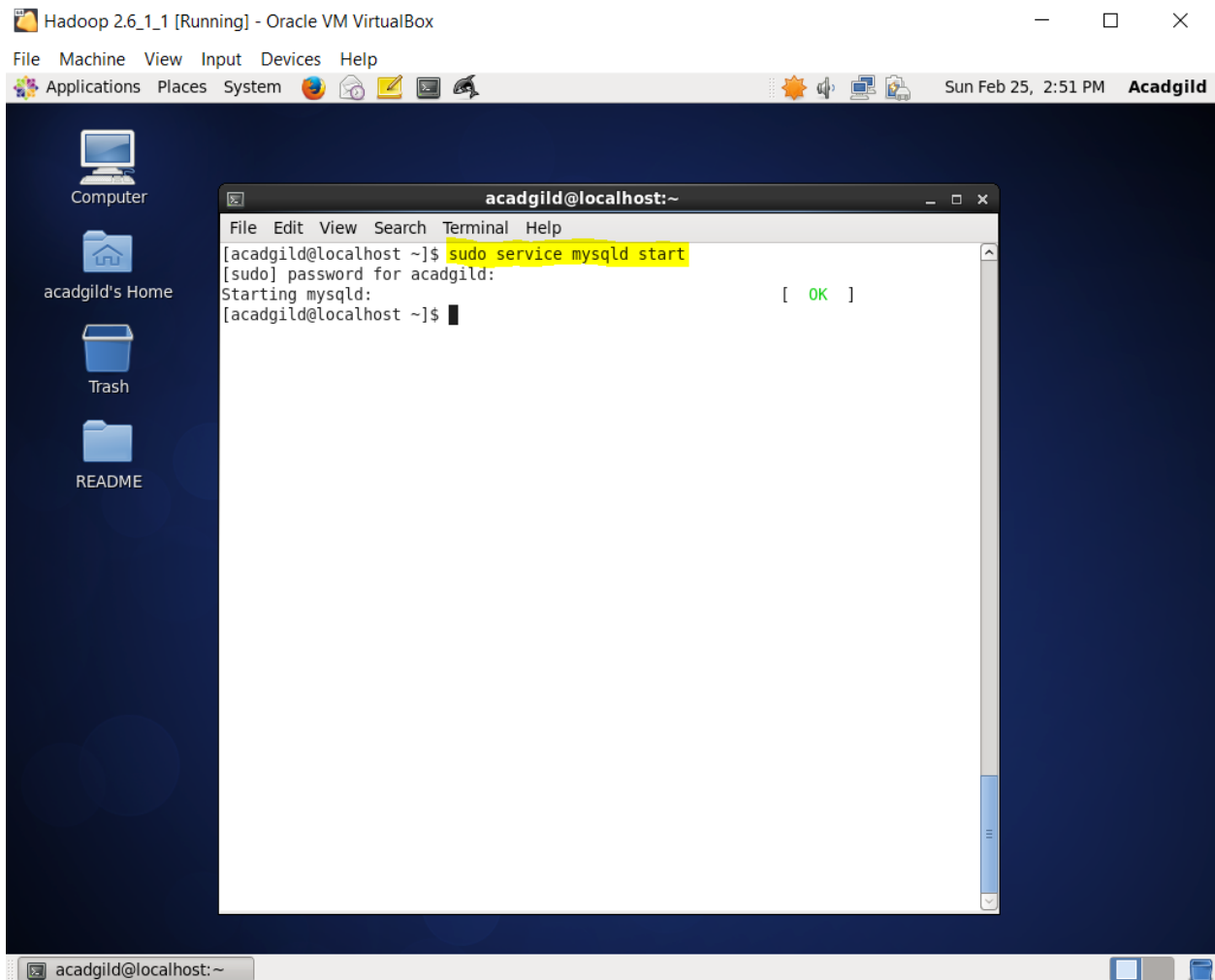


The screenshot shows a Linux desktop with a dark blue background. On the left sidebar, there are icons for 'Computer', 'acadgild's Home', 'Trash', and 'README'. The top panel displays 'Applications', 'Places', 'System', and system status 'Sun Feb 25, 2:49 PM Acadgild'. A terminal window titled 'acadgild@localhost:~' is open, showing the command 'pig' being executed. The terminal output displays various log messages from the Pig startup process, including version information, logging configuration, and warnings about deprecated settings.

```
acadgild@localhost:~$ pig
18/02/25 14:49:11 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/02/25 14:49:11 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/02/25 14:49:11 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-02-25 14:49:11,643 [main] INFO org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-02-25 14:49:11,643 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/acadgild/pig_1519550351640.log
2018-02-25 14:49:11,707 [main] INFO org.apache.pig.impl.util.Utils - Default bo
otup file /home/acadgild/.pigbootup not found
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/sha
re/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.
class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/s
lf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-02-25 14:49:12,940 [main] WARN org.apache.hadoop.util.NativeCodeLoader - U
nable to load native-hadoop library for your platform... using builtin-java clas
ses where applicable
2018-02-25 14:49:12,988 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2018-02-25 14:49:12,988 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2018-02-25 14:49:12,988 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:8020
2018-02-25 14:49:14,081 [main] INFO org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-c51145e6-7ac1-449d-879d-f7201c98d455
2018-02-25 14:49:14,081 [main] WARN org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt>
```

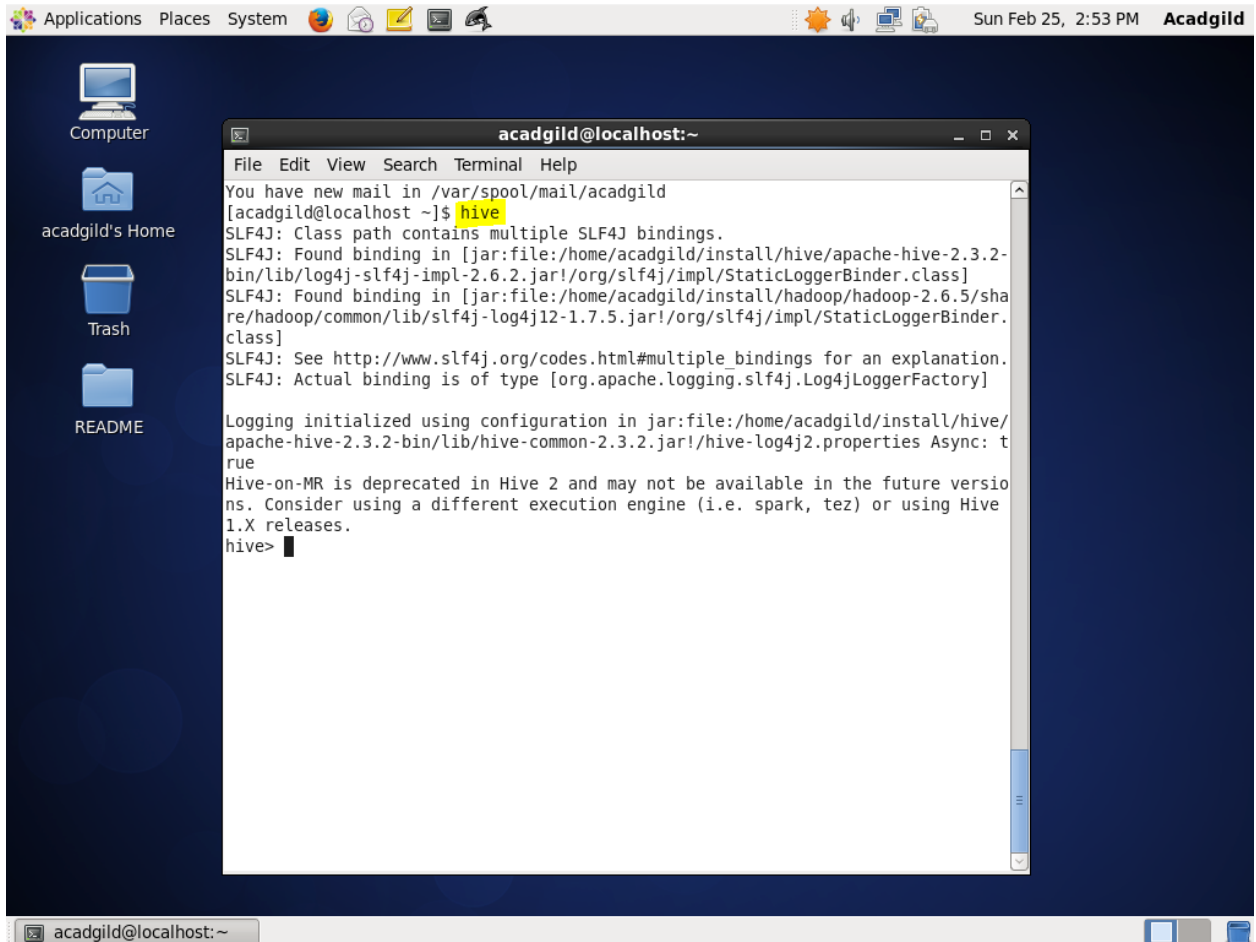
vi) Starting hive service

first start mysql service and then we have to start hive



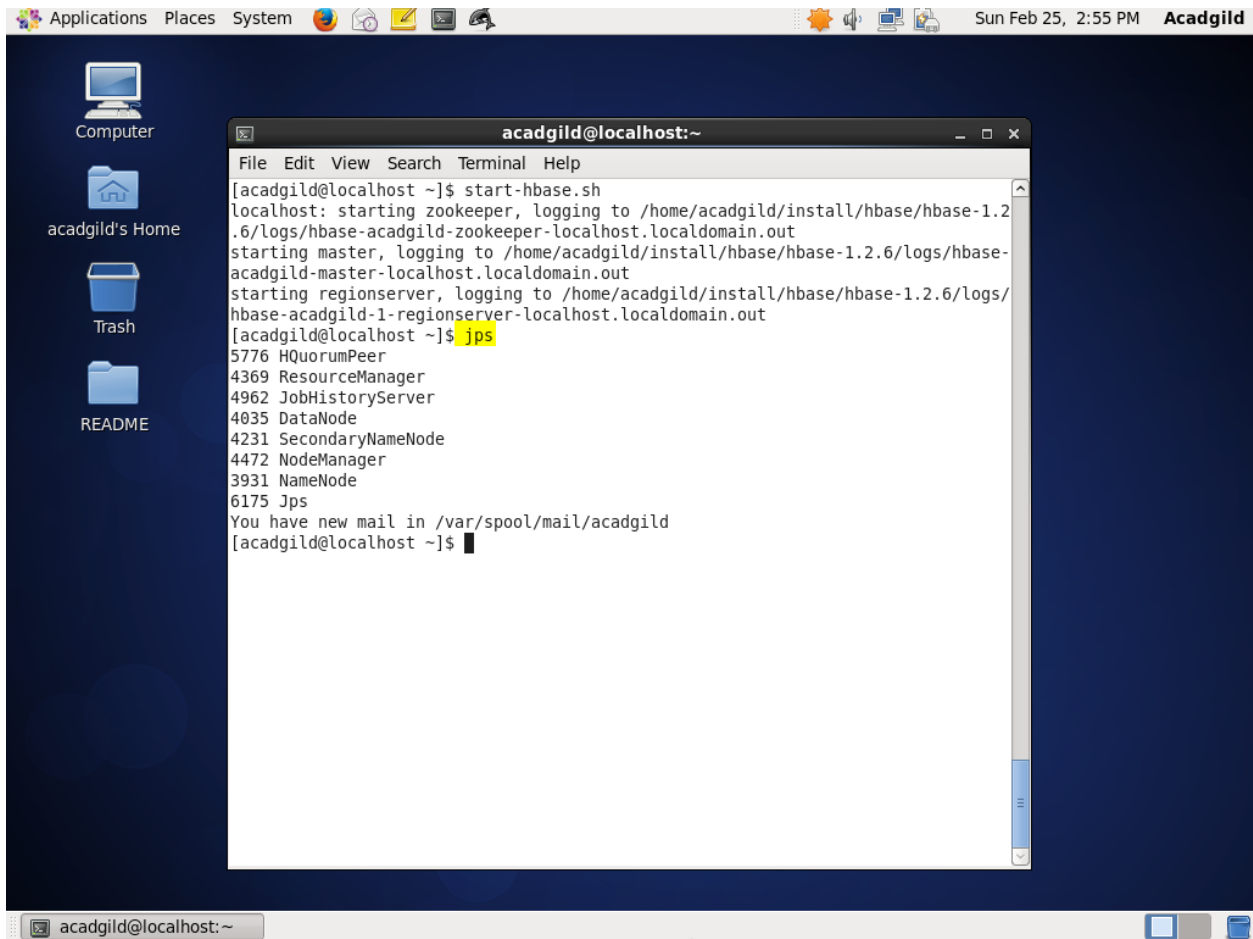
The Apache Hive distributed storage. By using Hive, we can access files stored in Hadoop Distributed File System (HDFS is used to querying and managing large datasets residing in) or in other data storage systems such as Apache HBase





vii) Start hbase using “start-hbase.sh” script

now type jps command to check all the hadoop deamons that we started are running.



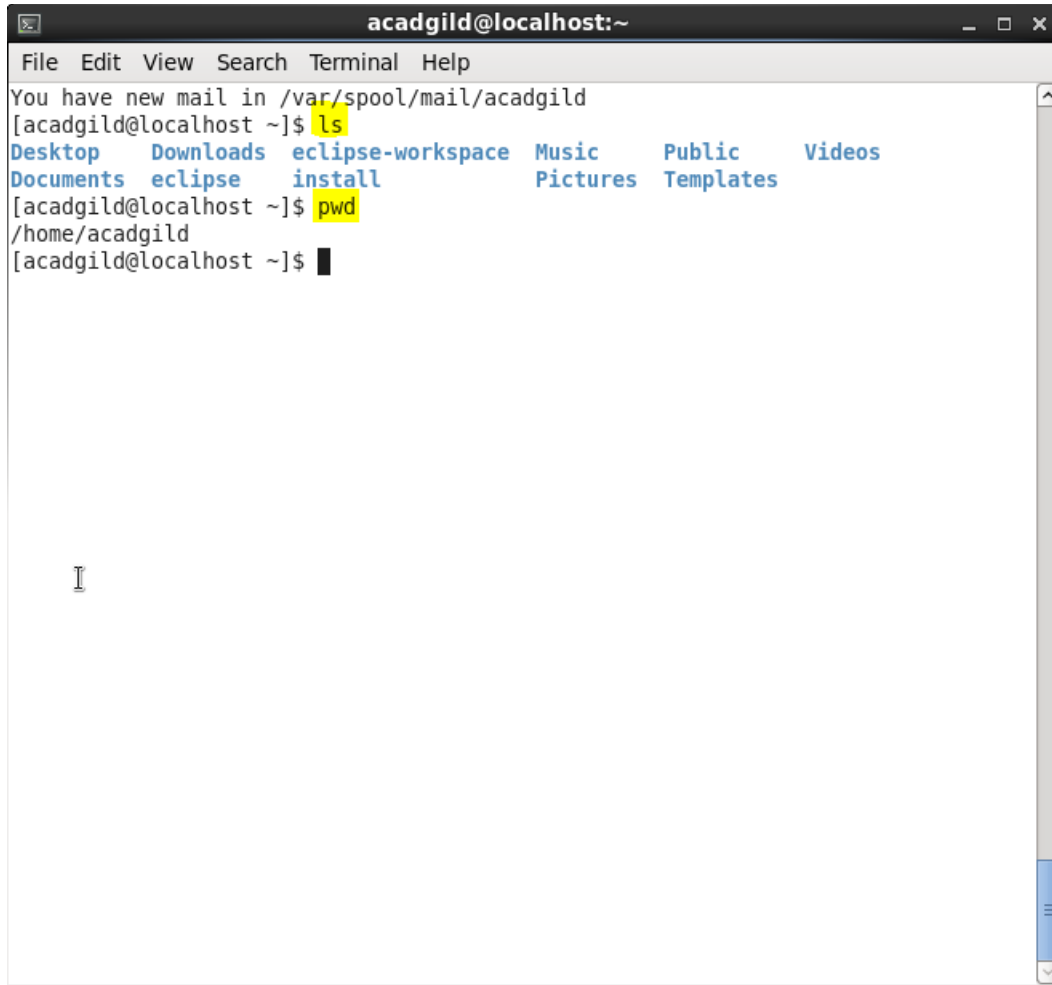
The screenshot shows a Linux desktop environment with a dark blue background. On the left side, there is a sidebar with icons for 'Computer', 'acadgild's Home', 'Trash', and 'README'. The top of the window features a menu bar with 'Applications', 'Places', and 'System', along with system status icons and the date 'Sun Feb 25, 2:55 PM' and the username 'Acadgild'. A terminal window titled 'acadgild@localhost:~' is open in the center. It displays the output of the 'start-hbase.sh' script, which includes starting ZooKeeper, HBase Master, and RegionServer. Below this, the 'jps' command is executed, showing a list of running Java processes: HQuorumPeer, ResourceManager, JobHistoryServer, DataNode, SecondaryNameNode, NodeManager, NameNode, and Jps. A system message about new mail is also visible.

```
acadgild@localhost:~$ start-hbase.sh
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper-localhost.localdomain.out
starting master, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-master-localhost.localdomain.out
starting regionserver, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-1-regionserver-localhost.localdomain.out
[acadgild@localhost ~]$ jps
5776 HQuorumPeer
4369 ResourceManager
4962 JobHistoryServer
4035 DataNode
4231 SecondaryNameNode
4472 NodeManager
3931 NameNode
6175 Jps
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

## 2) Linux Command

ls : lists all the files in the present working directory

pwd : displays the present working directory



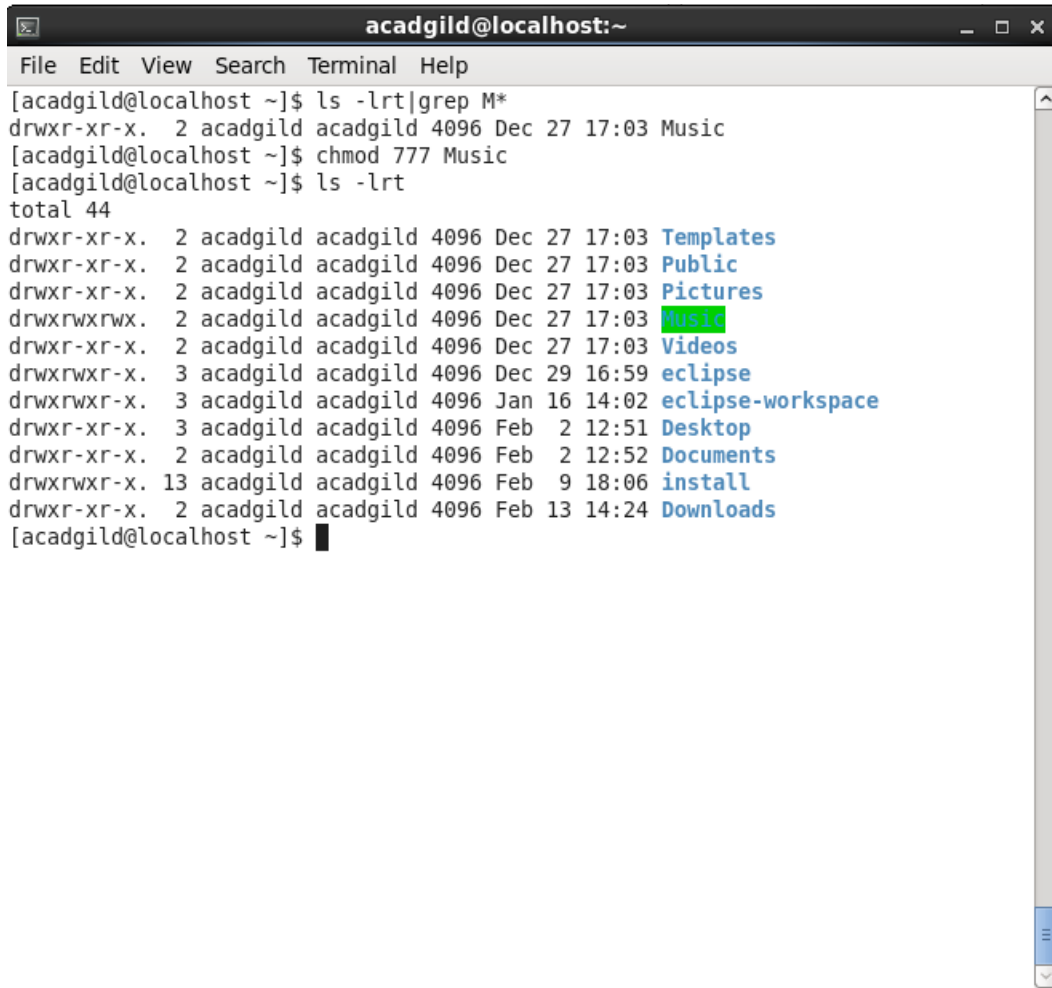
```
acadgild@localhost:~  
File Edit View Search Terminal Help  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$ ls  
Desktop  Downloads  eclipse-workspace  Music  Public  Videos  
Documents  eclipse  install  Pictures  Templates  
[acadgild@localhost ~]$ pwd  
/home/acadgild  
[acadgild@localhost ~]$
```

Ps -ef : displays all the processes running

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ ps -ef  
UID      PID  PPID  C  STIME TTY          TIME CMD  
root         1      0  0  14:17 ?        00:00:02 /sbin/init  
root         2      0  0  14:17 ?        00:00:00 [kthreadd]  
root         3      2  0  14:17 ?        00:00:00 [migration/0]  
root         4      2  0  14:17 ?        00:00:00 [ksoftirqd/0]  
root         5      2  0  14:17 ?        00:00:00 [stopper/0]  
root         6      2  0  14:17 ?        00:00:00 [watchdog/0]  
root         7      2  0  14:17 ?        00:00:21 [events/0]  
root         8      2  0  14:17 ?        00:00:00 [events/0]  
root         9      2  0  14:17 ?        00:00:00 [events_long/0]  
root        10      2  0  14:17 ?        00:00:00 [events_power_ef]  
root        11      2  0  14:17 ?        00:00:00 [cgroup]  
root        12      2  0  14:17 ?        00:00:00 [khelper]  
root        13      2  0  14:17 ?        00:00:00 [netns]  
root        14      2  0  14:17 ?        00:00:00 [async/mgr]  
root        15      2  0  14:17 ?        00:00:00 [pm]  
root        16      2  0  14:17 ?        00:00:00 [sync_supers]  
root        17      2  0  14:17 ?        00:00:00 [bdi-default]  
root        18      2  0  14:17 ?        00:00:00 [kintegrityd/0]  
root        19      2  0  14:17 ?        00:00:02 [kblockd/0]  
root        20      2  0  14:17 ?        00:00:00 [kacpid]  
root        21      2  0  14:17 ?        00:00:00 [kacpi_notify]  
root        22      2  0  14:17 ?        00:00:00 [kacpi_hotplug]  
root        23      2  0  14:17 ?        00:00:00 [ata_aux]  
root        24      2  0  14:17 ?        00:00:00 [ata_sff/0]  
root        25      2  0  14:17 ?        00:00:00 [ksuspend_usbd]  
root        26      2  0  14:17 ?        00:00:00 [khubd]  
root        27      2  0  14:17 ?        00:00:00 [kseriod]  
root        28      2  0  14:17 ?        00:00:00 [md/0]  
root        29      2  0  14:17 ?        00:00:00 [md_misc/0]  
root        30      2  0  14:17 ?        00:00:00 [linkwatch]  
root        33      2  0  14:17 ?        00:00:00 [khungtaskd]
```

ls -lrt : displays files and their permission set for user,group,others

chmod 777 <file name> changes the permissions set for user,group,others

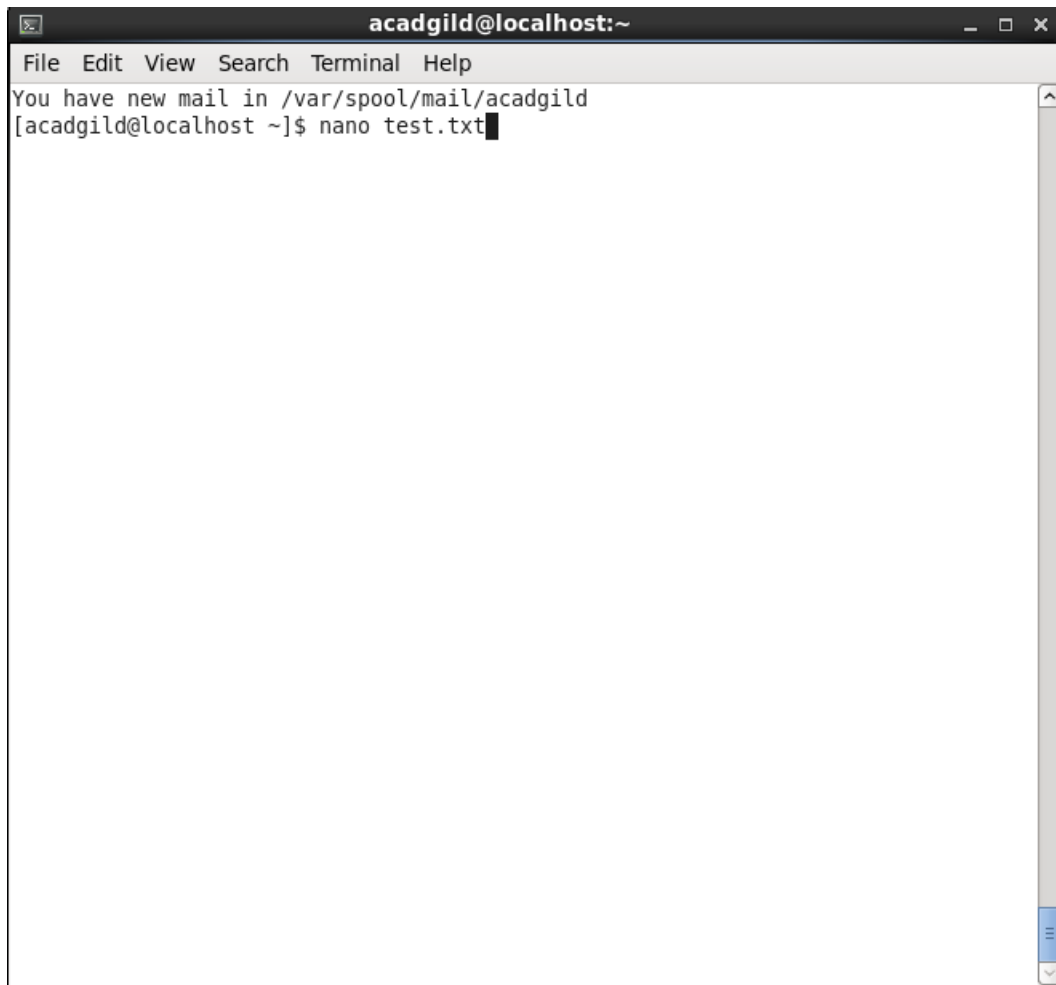
A terminal window titled 'acadgild@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
[acadgild@localhost ~]$ ls -lrt|grep M*
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Music
[acadgild@localhost ~]$ chmod 777 Music
[acadgild@localhost ~]$ ls -lrt
total 44
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Templates
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Public
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Pictures
drwxrwxrwx. 2 acadgild acadgild 4096 Dec 27 17:03 Music
drwxr-xr-x. 2 acadgild acadgild 4096 Dec 27 17:03 Videos
drwxrwxr-x. 3 acadgild acadgild 4096 Dec 29 16:59 eclipse
drwxrwxr-x. 3 acadgild acadgild 4096 Jan 16 14:02 eclipse-workspace
drwxr-xr-x. 3 acadgild acadgild 4096 Feb  2 12:51 Desktop
drwxr-xr-x. 2 acadgild acadgild 4096 Feb  2 12:52 Documents
drwxrwxr-x. 13 acadgild acadgild 4096 Feb  9 18:06 install
drwxr-xr-x. 2 acadgild acadgild 4096 Feb 13 14:24 Downloads
[acadgild@localhost ~]$
```

df : displays the free space on the disk

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ df  
Filesystem            1K-blocks      Used Available Use% Mounted on  
/dev/mapper/VolGroup-lv_root  
17938864 11689028    5331924   69% /  
tmpfs                  961076       14412    946664    2% /dev/shm  
/dev/sda1              487652      105781    356271   23% /boot  
[acadgild@localhost ~]$
```

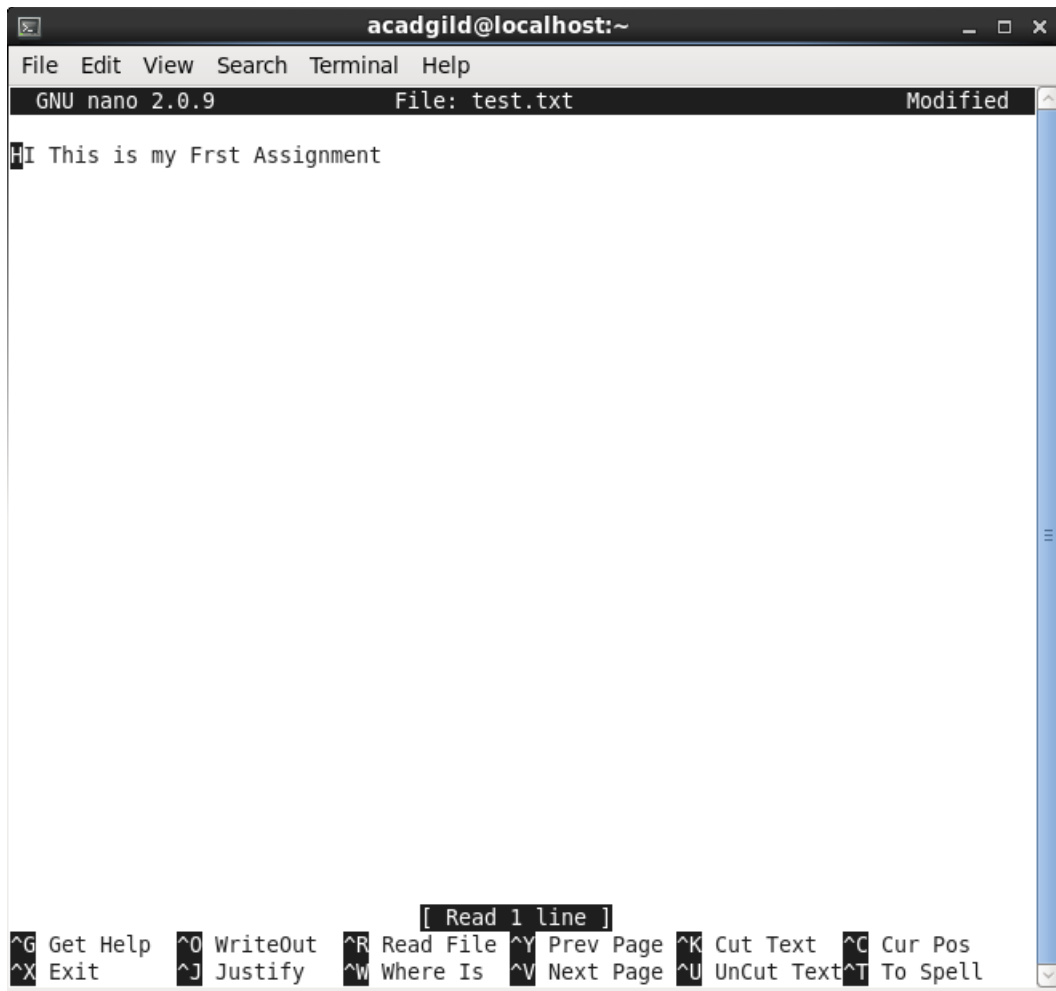
Creating a file using command nano



```
acadgild@localhost:~  
File Edit View Search Terminal Help  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$ nano test.txt
```

The image shows a terminal window titled "acadgild@localhost:~". The window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The main area of the terminal displays a notification "You have new mail in /var/spool/mail/acadgild" and the command prompt "[acadgild@localhost ~]\$ nano test.txt" with a cursor at the end of the line. The terminal window has a scrollbar on the right side.

Add some text in the file the save the file using ctrl+x and then type Y and then provide a name.



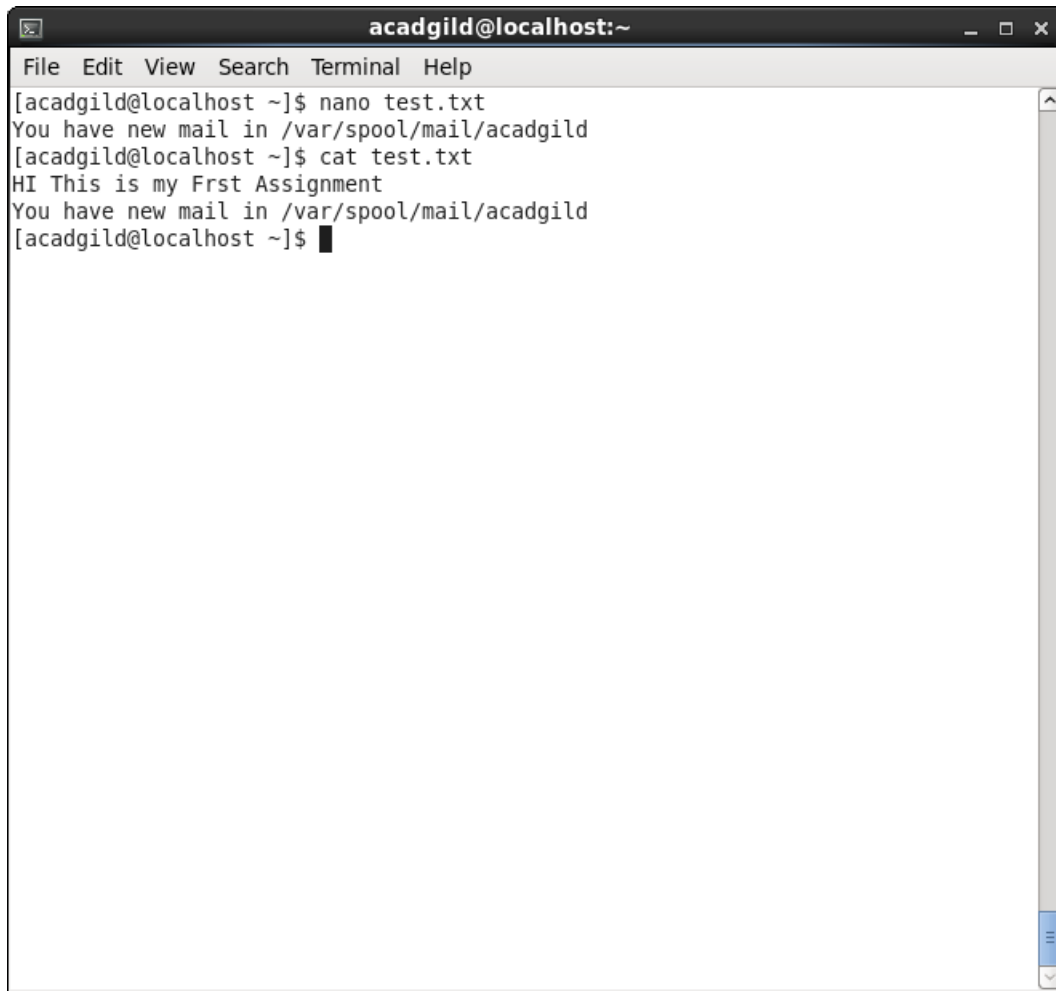
The image shows a terminal window with the nano text editor open. The window title is 'acadgild@localhost:~'. The nano editor's menu bar includes 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. Below the menu bar, it says 'GNU nano 2.0.9' and 'File: test.txt'. The main editing area contains the text 'HI This is my Frst Assignment'. At the bottom, there is a status bar with the text '[ Read 1 line ]' and a list of keyboard shortcuts: '^G Get Help', '^O WriteOut', '^R Read File', '^Y Prev Page', '^K Cut Text', '^C Cur Pos', '^X Exit', '^J Justify', '^W Where Is', '^V Next Page', '^U UnCut Text', and '^T To Spell'.

```
acadgild@localhost:~
File Edit View Search Terminal Help
GNU nano 2.0.9 File: test.txt Modified
HI This is my Frst Assignment

[ Read 1 line ]
^G Get Help ^O WriteOut ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where Is ^V Next Page ^U UnCut Text ^T To Spell
```



Use cat command to display the contents of the file "cat <filename>"



```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ nano test.txt  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$ cat test.txt  
HI This is my Frst Assignment  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$
```

- 3) Opening HDFS Homepage using the url "http://localhost:50070"

The screenshot shows a Mozilla Firefox browser window titled "Namenode information - Mozilla Firefox". The address bar displays "localhost:50070/dfshealth.html#tab-overview". The page has a green navigation bar with tabs: "Hadoop", "Overview", "Datanodes", "Snapshot", "Startup Progress", and "Utilities". The "Overview" tab is selected, showing the title "Overview 'localhost:8020' (active)". Below the title is a table with the following data:

<b>Started:</b>	Sun Feb 25 14:43:19 IST 2018
<b>Version:</b>	2.6.5, re8c9fe0b4c252caf2ebf1464220599650f119997
<b>Compiled:</b>	2016-10-02T23:43Z by sjlee from branch-2.6.5
<b>Cluster ID:</b>	CID-7b3f9bd8-f34c-4fb8-87aa-f76b6dfbd809
<b>Block Pool ID:</b>	BP-437583619-127.0.0.1-1517555661954

Below the table is a section titled "Summary" with the following text:

Security is off.  
Safemode is off.  
23 files and directories, 1 blocks = 24 total filesystem object(s).  
Heap Memory used 31.06 MB of 59.89 MB Heap Memory. Max Heap Memory is 966.69 MB.

The browser's taskbar at the bottom shows the terminal window "[acdgild@localhost:~]" and the "Namenode informatio..." window.

Namenode information

localhost:50070/dfshealth.html#tab-overview

Search

Heap Memory used 31.06 MB of 59.89 MB Heap Memory. Max Heap Memory is 966.69 MB.  
 Non Heap Memory used 51.01 MB of 52.19 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	17.11 GB
DFS Used:	48 KB
Non DFS Used:	12.02 GB
DFS Remaining:	5.08 GB
DFS Used%:	0%
DFS Remaining%:	29.72%
Block Pool Used:	48 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 Started	2/25/2018, 2:43:19 PM

acadmild@localhost:~]

Namenode informatio...

Namenode information - Mozilla Firefox

Namenode information x +

localhost:50070/dfshealth.html#tab-overview | Search

Number of Blocks Pending Deletion

0

Block Deletion Start Time

2/25/2018, 2:43:19 PM

## NameNode Journal Status

Current transaction ID: 354

Journal Manager	State
FileJournalManager(root=/home/acadgild/install/data/dfs/name)	EditLogFileOutputStream(/home/acadgild/install/data/dfs/name/current/edits_inprogress_000000000000000354)

## NameNode Storage

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

Hadoop, 2016.

Legacy UI

[acadgild@localhost:~]

Namenode informatio...

Legacy UI