Assignment 3
CSCI E 63 - Big Data Analytics

Problem 1:
Create your own Virtual Machine with a Linux operating system. The lecture notes speak about CentOS. You are welcome to work with another Linux OS. When creating the VM, create an administrative user. Call that user whatever you feel like. Please record the password of the new user. Once the VM is created transfer the attached text file Ulysses10.txt to the home of new user. You can do it using scp (secure copy command) or email. Examine the version of Java, Python and Scala on your VM. If any of those versions is below requirements for Spark 2.2 install proper version. Set JAVA_HOME environmental variable. Set your PATH environmental variable properly, so that you can invoke: java, sbt and python commands from any directory on your system.

Answer:

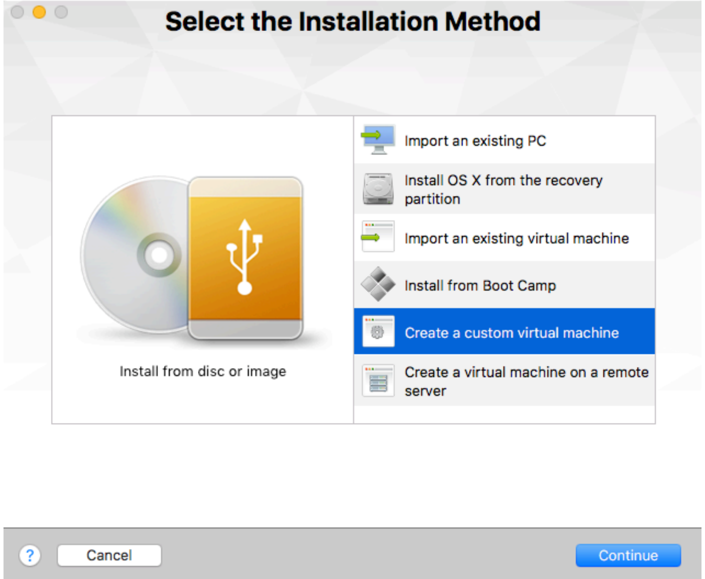
Steps followed fro create a new Virtual Machine:

Downloads:

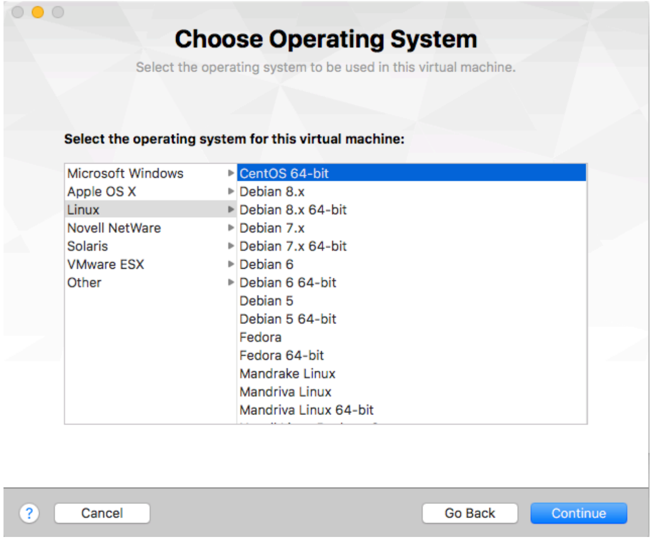
- Go to https://www.centos.org, select the button: GetCentOS Now. Choose ‘Everything ISO’ on the next page and select a link near you. Normally, VMWare Workstation installs a Linux OS from an iso image.
- Download VMware Fusion 8

Steps:

- Select Add → New → Create a custom virtual machine


Select the Installation Method
Install from disc or image
Import an existing PC
Install OS X from the recovery partition
Import an existing virtual machine
Install from Boot Camp
Create a custom virtual machine
Create a virtual machine on a remote server
?
Cancel
Continue

Assignment 3
CSCI E 63 - Big Data Analytics



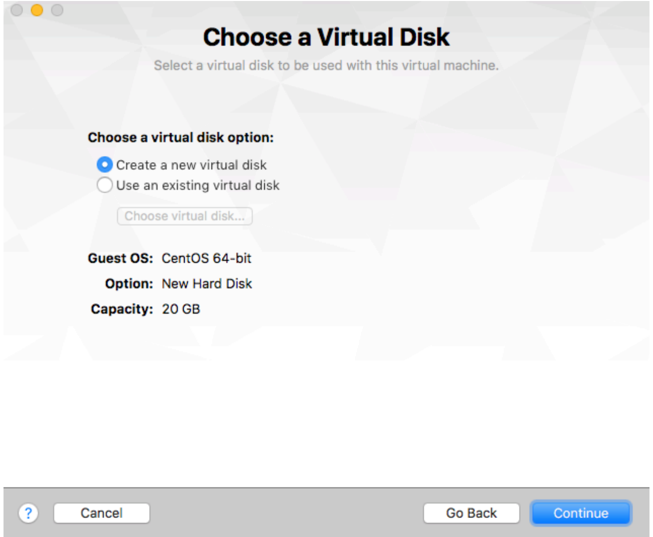
Choose Operating System

Select the operating system to be used in this virtual machine.

Select the operating system for this virtual machine:

Microsoft Windows
Apple OS X
Linux
Novell NetWare
Solaris
VMware ESX
Other

CentOS 64-bit
Debian 8.x
Debian 8.x 64-bit
Debian 7.x
Debian 7.x 64-bit
Debian 6
Debian 6 64-bit
Debian 5
Debian 5 64-bit
Fedora
Fedora 64-bit
Mandrake Linux
Mandriva Linux
Mandriva Linux 64-bit



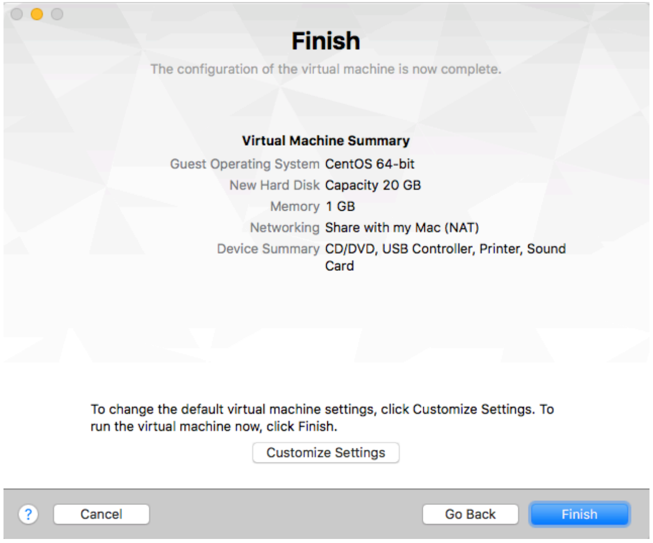
Choose a Virtual Disk

Select a virtual disk to be used with this virtual machine.

Choose a virtual disk option:

● Create a new virtual disk
○ Use an existing virtual disk

Guest OS: CentOS 64-bit
Option: New Hard Disk
Capacity: 20 GB



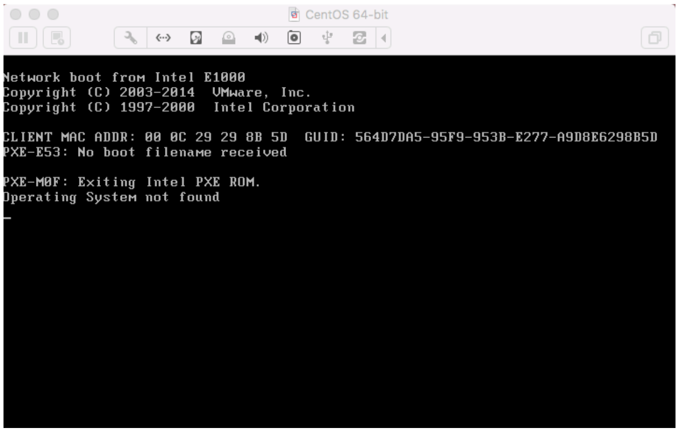
Finish

The configuration of the virtual machine is now complete.

Virtual Machine Summary

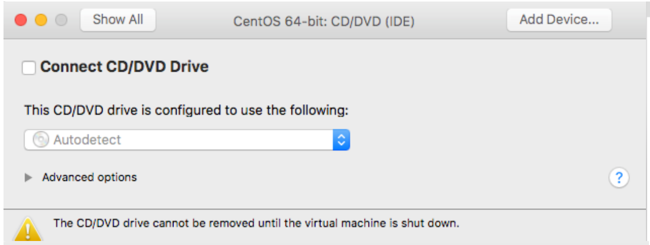
Guest Operating System CentOS 64-bit
New Hard Disk Capacity 20 GB
Memory 1 GB
Networking Share with my Mac (NAT)
Device Summary CD/DVD, USB Controller, Printer, Sound Card

To change the default virtual machine settings, click Customize Settings. To run the virtual machine now, click Finish.



```
Network boot from Intel E1000
Copyright (C) 2003-2014  VMware, Inc.
Copyright (C) 1997-2000  Intel Corporation

CLIENT MAC ADDR: 00 0C 29 29 8B 5D  GUID: 564D7DA5-95F9-953B-E277-A9D8E6298B5D
PXE-E53: No boot filename received

PXE-M0F: Exiting Intel PXE ROM.
Operating System not found
_
```

- "Operating system not found." This is where you use the iso image. Select Settings →CD/DVD

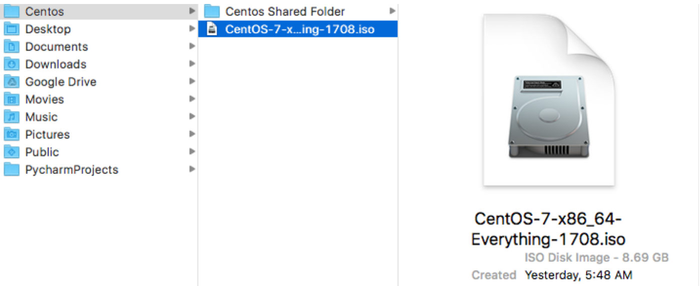


□ **Connect CD/DVD Drive**

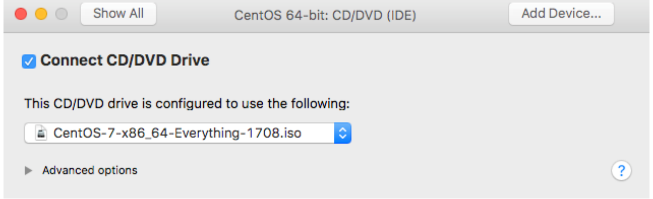
This CD/DVD drive is configured to use the following:

Autodetect

The CD/DVD drive cannot be removed until the virtual machine is shut down.



CentOS-7-x86_64-Everything-1708.iso
ISO Disk Image - 8.69 GB
Created Yesterday, 5:48 AM



☑ **Connect CD/DVD Drive**

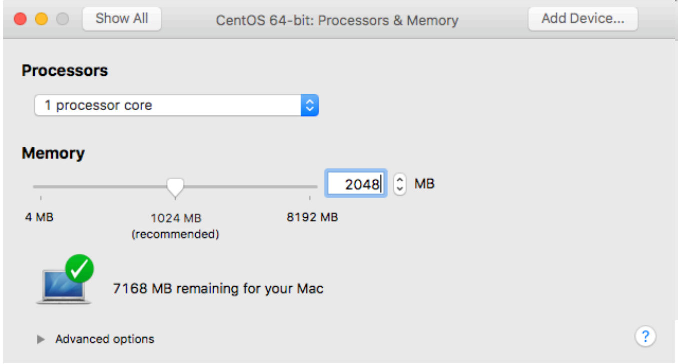
This CD/DVD drive is configured to use the following:

CentOS-7-x86_64-Everything-1708.iso

Karan A. Bhandarkar

Assignment 3
CSCI E 63 - Big Data Analytics

- The default Memory selection is 1024MB. Change it to at least 2048MB. Make sure the Virtual Disk has at least 30GB and is 'split into multiple files'.



CentOS 64-bit: Processors & Memory

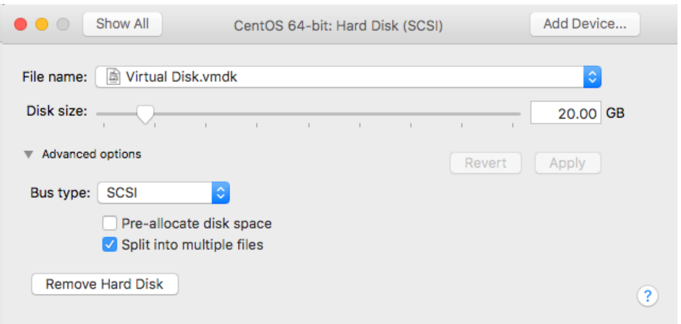
Processors

1 processor core

Memory

2048 MB

7168 MB remaining for your Mac



CentOS 64-bit: Hard Disk (SCSI)

File name: Virtual Disk.vmdk

Disk size: 20.00 GB

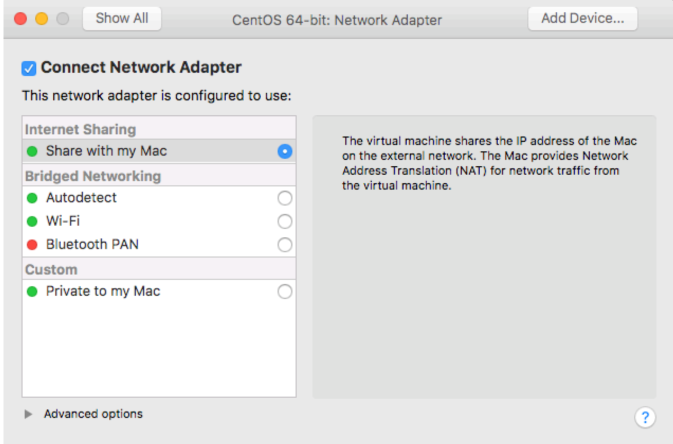
Advanced options

Bus type: SCSI

☐ Pre-allocate disk space

☑ Split into multiple files

- The default 'Share with my mac' Network Adapter is set up as below. Add another one using Add Device → Network Adapter as 'Private to my Mac'.



CentOS 64-bit: Network Adapter

☑ **Connect Network Adapter**

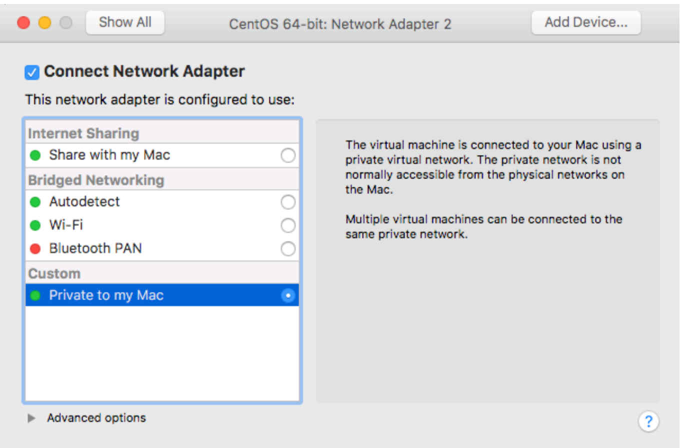
This network adapter is configured to use:

Internet Sharing
- Share with my Mac ●

Bridged Networking
- Autodetect ○
- Wi-Fi ○
- Bluetooth PAN ○

Custom
- Private to my Mac ○

The virtual machine shares the IP address of the Mac on the external network. The Mac provides Network Address Translation (NAT) for network traffic from the virtual machine.



CentOS 64-bit: Network Adapter 2

☑ **Connect Network Adapter**

This network adapter is configured to use:

Internet Sharing
- Share with my Mac ○

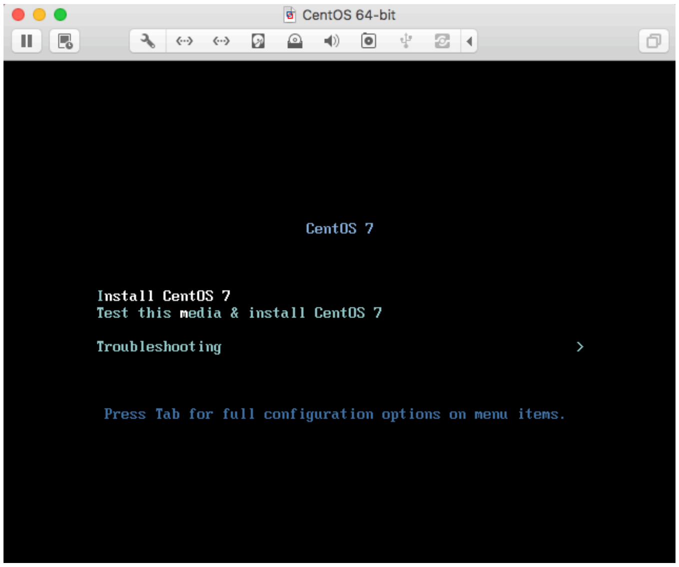
Bridged Networking
- Autodetect ○
- Wi-Fi ○
- Bluetooth PAN ○

Custom
- Private to my Mac ●

The virtual machine is connected to your Mac using a private virtual network. The private network is not normally accessible from the physical networks on the Mac.

Multiple virtual machines can be connected to the same private network.

- Start the VM. Use arrow keys and choose 'Select Install CentOS 7'.



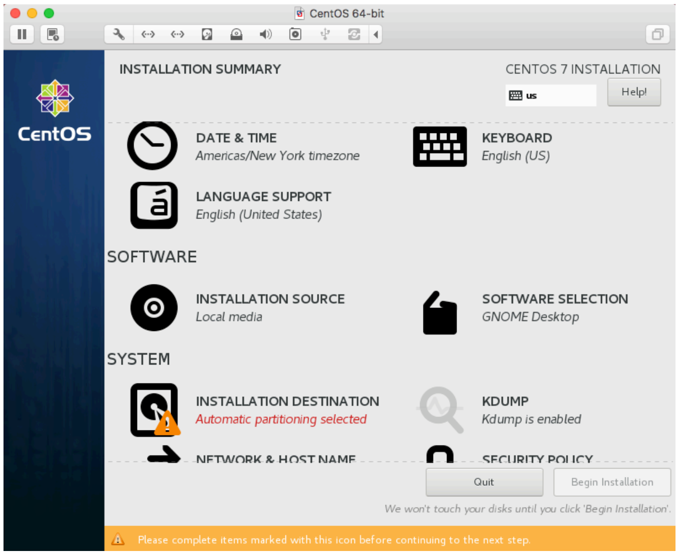
CentOS 64-bit

```
                CentOS 7


  Install CentOS 7
  Test this media & install CentOS 7

  Troubleshooting                                >


   Press Tab for full configuration options on menu items.
```

Assignment 3
CSCI E 63 - Big Data Analytics

- Steps for Installation Summary screen i.e. the screen that follows:

1. Click on Software Selection and choose GNOME Desktop



CentOS 64-bit

INSTALLATION SUMMARY

CENTOS 7 INSTALLATION

DATE & TIME: Americas/New York timezone

KEYBOARD: English (US)

LANGUAGE SUPPORT: English (United States)

SOFTWARE

INSTALLATION SOURCE: Local media

SOFTWARE SELECTION: GNOME Desktop

SYSTEM

INSTALLATION DESTINATION: Automatic partitioning selected

KDUMP: Kdump is enabled

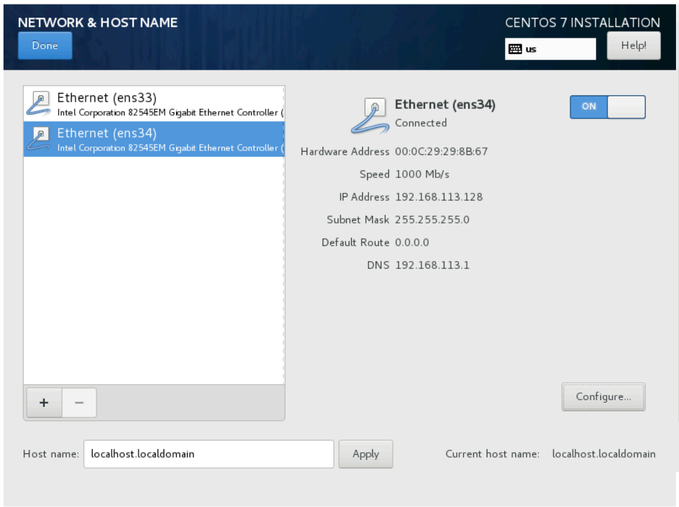
NETWORK & HOST NAME

SECURITY POLICY

We won't touch your disks until you click 'Begin Installation'.

Please complete items marked with this icon before continuing to the next step.

2. Select Network & Hostname. On the next screen toggle both network card to ON. Go to Configure → General. Check the option 'Automatically connect to this network when it is available.' This will make the network cards be on after every reboot.



NETWORK & HOST NAME

CENTOS 7 INSTALLATION

Ethernet (ens33)
Intel Corporation 82545EM Gigabit Ethernet Controller (

Ethernet (ens34)
Intel Corporation 82545EM Gigabit Ethernet Controller (

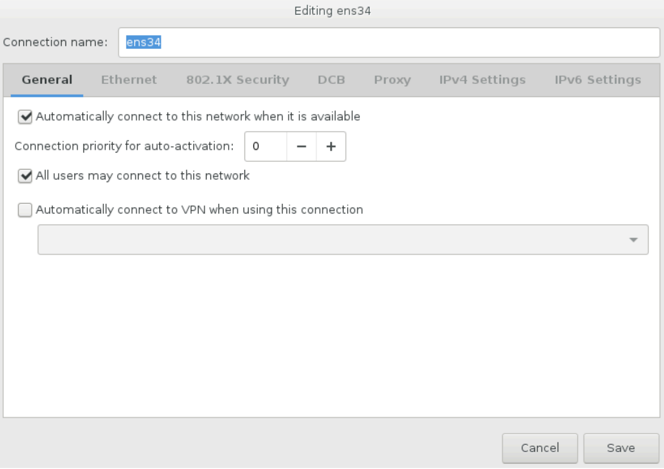
Ethernet (ens34)
Connected

ON

Hardware Address 00:0C:29:29:8B:67
Speed 1000 Mb/s
IP Address 192.168.113.128
Subnet Mask 255.255.255.0
Default Route 0.0.0.0
DNS 192.168.113.1

Host name: localhost.localdomain

Current host name: localhost.localdomain



Editing ens34

Connection name: ens34

☑ Automatically connect to this network when it is available

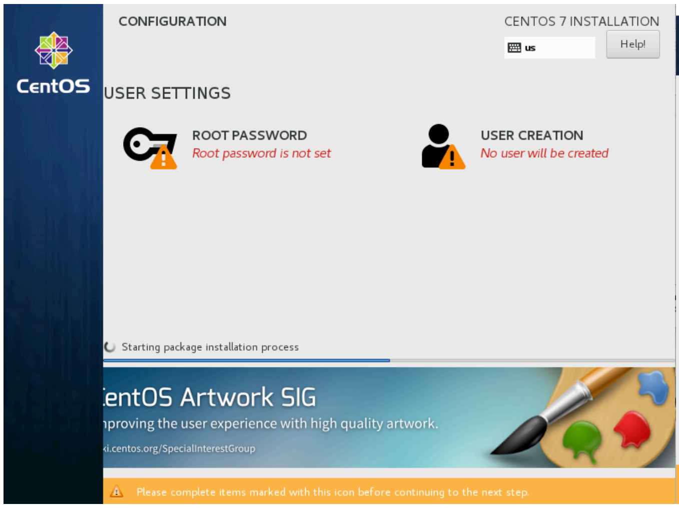
Connection priority for auto-activation: 0

☑ All users may connect to this network

☐ Automatically connect to VPN when using this connection

Assignment 3
CSCI E 63 - Big Data Analytics

- On the Configuration screen i.e. the screen that follows, select first Root Password and then create a user account through User Creation. Do not forget to check 'Make this user administrator'. We will use this user to run sudo commands.



CentOS
CONFIGURATION
CENTOS 7 INSTALLATION
us
Help!
USER SETTINGS

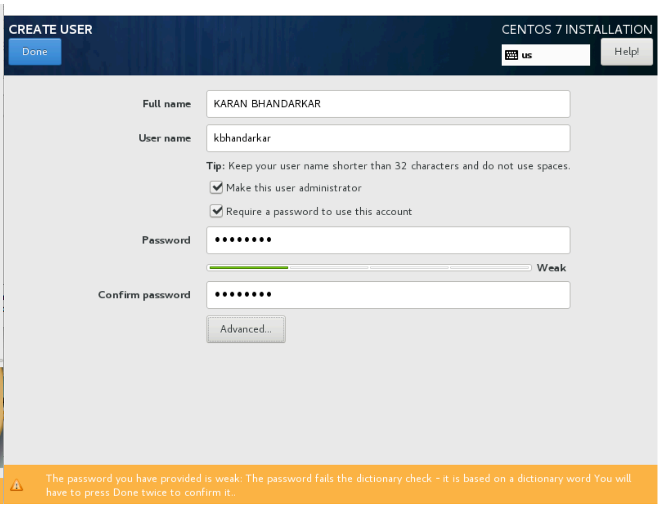
ROOT PASSWORD
Root password is not set

USER CREATION
No user will be created

Starting package installation process

entOS Artwork SIG
nproving the user experience with high quality artwork.
i.centos.org/SpecialInterestGroup

Please complete items marked with this icon before continuing to the next step.



Full name: KARAN BHANDARKAR

User name: kbhandarkar

Tip: Keep your user name shorter than 32 characters and do not use spaces.

☑ Make this user administrator

☑ Require a password to use this account

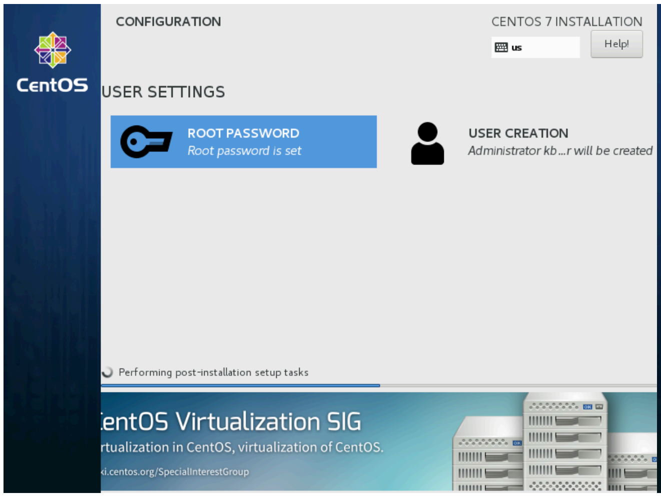
Password: ••••••••

Weak

Confirm password: ••••••••

The password you have provided is weak: The password fails the dictionary check - it is based on a dictionary word You will have to press Done twice to confirm it.

- The installation will proceed another 10-20 minutes. Wait. On the bottom of the configuration screen you can see names of various Linux modules that are being installed. After a while you will be asked to Reboot. Do it.

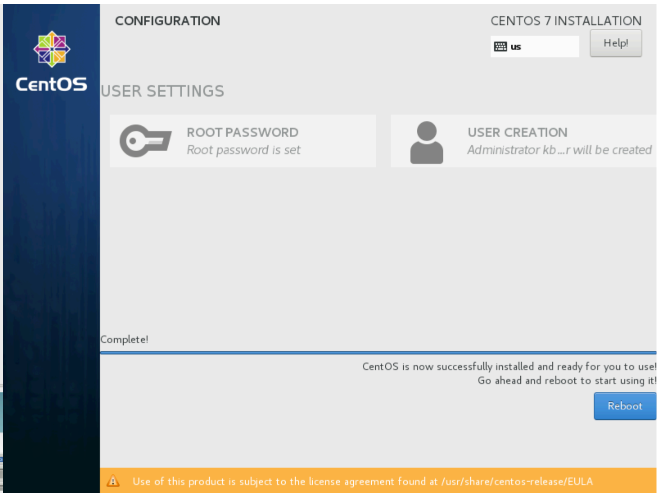


ROOT PASSWORD
Root password is set

USER CREATION
Administrator kb...r will be created

Performing post-installation setup tasks

entOS Virtualization SIG
rtualization in CentOS, virtualization of CentOS.
i.centos.org/SpecialInterestGroup



ROOT PASSWORD
Root password is set

USER CREATION
Administrator kb...r will be created

Complete!

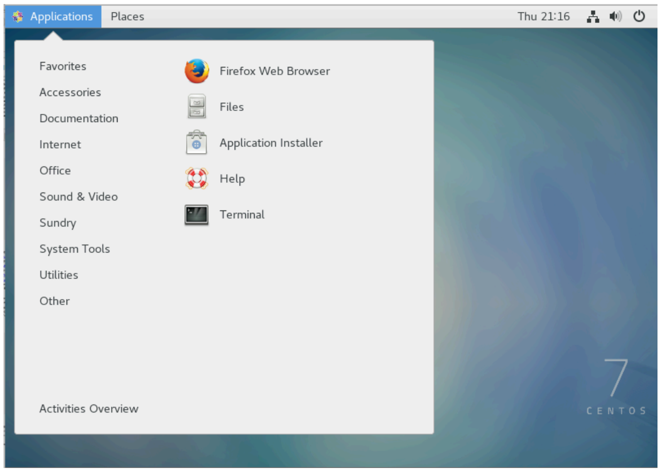
CentOS is now successfully installed and ready for you to use!
Go ahead and reboot to start using it!

Use of this product is subject to the license agreement found at /usr/share/centos-release/EULA

- After the reboot, accept the License Information and select 'Finish Configuration'.

Assignment 3
CSCI E 63 – Big Data Analytics

- Command terminal is at Applications > Favorites > Terminal

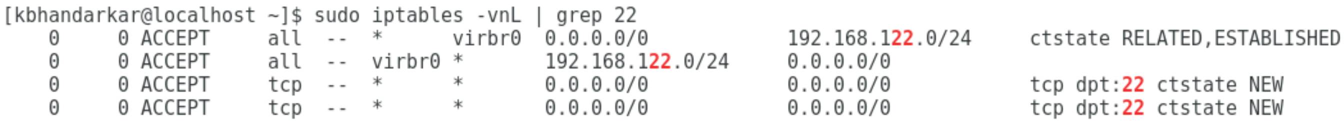


- Open port 22 from the terminal using commands:

  $ sudo firewall-cmd --zone=public --add-port=22/tcp --permanent

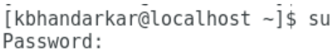
  $ sudo firewall-cmd --reload

- You can check whether port 22 has been added using iptables command:



```
[kbhandarkar@localhost ~]$ sudo iptables -vnL | grep 22
    0     0 ACCEPT     all  --  *      virbr0  0.0.0.0/0            192.168.122.0/24     ctstate RELATED,ESTABLISHED
    0     0 ACCEPT     all  --  virbr0 *       192.168.122.0/24     0.0.0.0/0
    0     0 ACCEPT     tcp  --  *      *       0.0.0.0/0            0.0.0.0/0            tcp dpt:22 ctstate NEW
    0     0 ACCEPT     tcp  --  *      *       0.0.0.0/0            0.0.0.0/0            tcp dpt:22 ctstate NEW
```

- Subsequently, you will have to reboot your system:



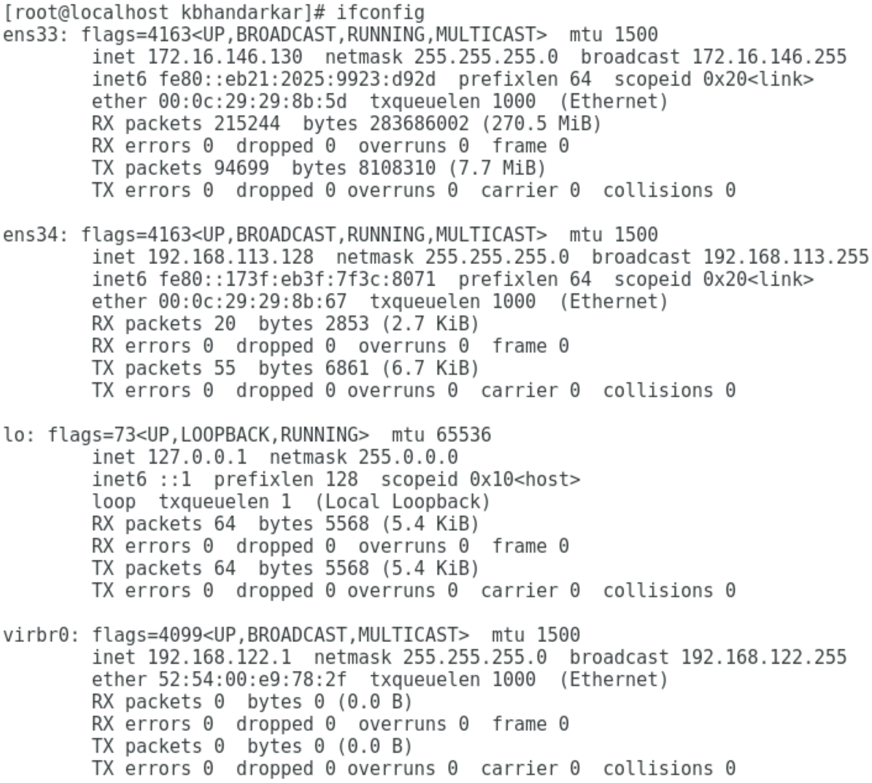
```
[kbhandarkar@localhost ~]$ su
Password:
```

- The Virtual Machine is now ready.

Karan A. Bhandarkar

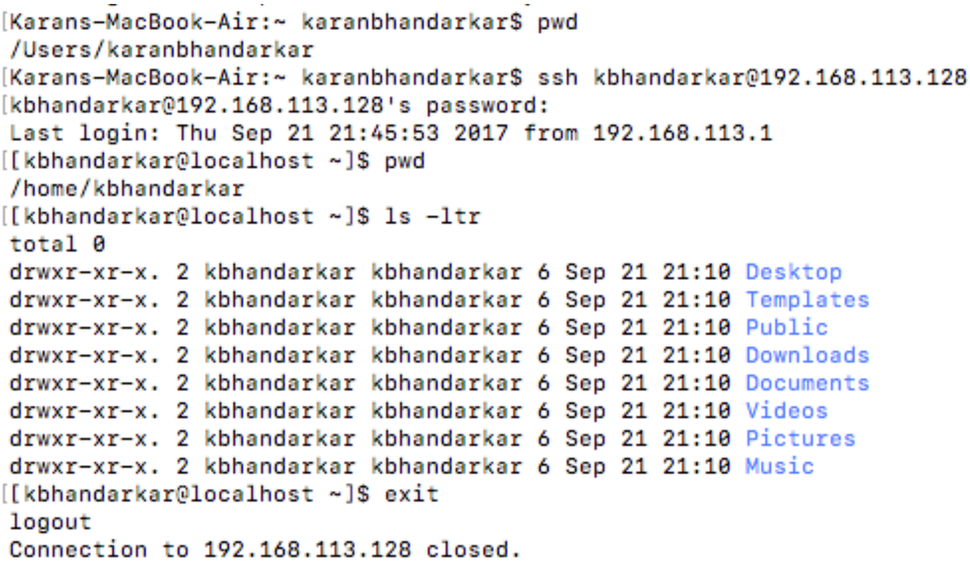
Steps followed to transfer a file:

- Open Terminal window and find the IP:



```
[root@localhost kbhandarkar]# ifconfig
ens33: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 172.16.146.130  netmask 255.255.255.0  broadcast 172.16.146.255
        inet6 fe80::eb21:2025:9923:d92d  prefixlen 64  scopeid 0x20<link>
        ether 00:0c:29:29:8b:5d  txqueuelen 1000  (Ethernet)
        RX packets 215244  bytes 283686002 (270.5 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 94699  bytes 8108310 (7.7 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

ens34: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.113.128  netmask 255.255.255.0  broadcast 192.168.113.255
        inet6 fe80::173f:eb3f:7f3c:8071  prefixlen 64  scopeid 0x20<link>
        ether 00:0c:29:29:8b:67  txqueuelen 1000  (Ethernet)
        RX packets 20  bytes 2853 (2.7 KiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 55  bytes 6861 (6.7 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1  (Local Loopback)
        RX packets 64  bytes 5568 (5.4 KiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 64  bytes 5568 (5.4 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

virbr0: flags=4099<UP,BROADCAST,MULTICAST>  mtu 1500
        inet 192.168.122.1  netmask 255.255.255.0  broadcast 192.168.122.255
        ether 52:54:00:e9:78:2f  txqueuelen 1000  (Ethernet)
        RX packets 0  bytes 0 (0.0 B)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 0  bytes 0 (0.0 B)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

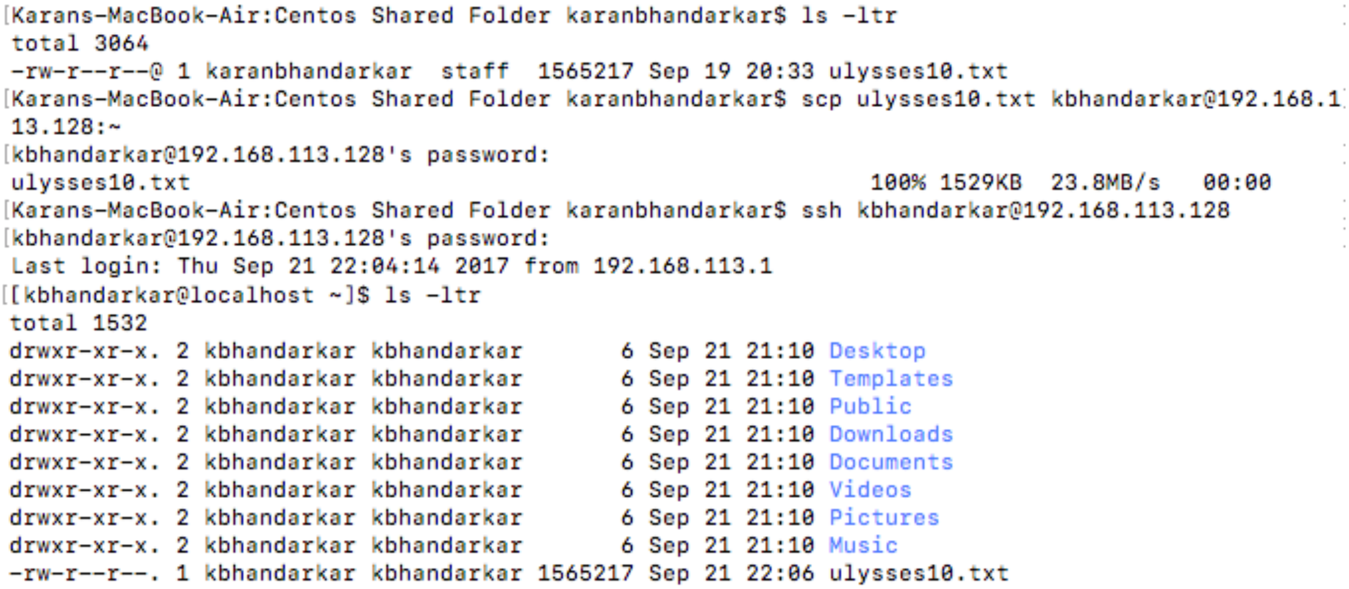
- We can see the IP address is 192.168.113.128. We can use the IP address to connect to the VM as if it is a server. Use ssh command to explore the folder structure.



```
[Karans-MacBook-Air:~ karanbhandarkar$ pwd
 /Users/karanbhandarkar
[Karans-MacBook-Air:~ karanbhandarkar$ ssh kbhandarkar@192.168.113.128
[kbhandarkar@192.168.113.128's password:
 Last login: Thu Sep 21 21:45:53 2017 from 192.168.113.1
[[kbhandarkar@localhost ~]$ pwd
 /home/kbhandarkar
[[kbhandarkar@localhost ~]$ ls -ltr
 total 0
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Desktop
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Templates
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Public
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Downloads
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Documents
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Videos
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Pictures
 drwxr-xr-x. 2 kbhandarkar kbhandarkar 6 Sep 21 21:10 Music
[[kbhandarkar@localhost ~]$ exit
 logout
 Connection to 192.168.113.128 closed.
```

Assignment 3
CSCI E 63 – Big Data Analytics

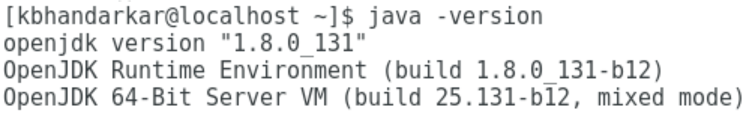
- Use scp command to transfer files to the VM. Use ssh command to connect to VM and validate.



```
[Karans-MacBook-Air:Centos Shared Folder karanbhandarkar$ ls -ltr
 total 3064
 -rw-r--r--@ 1 karanbhandarkar  staff  1565217 Sep 19 20:33 ulysses10.txt
[Karans-MacBook-Air:Centos Shared Folder karanbhandarkar$ scp ulysses10.txt kbhandarkar@192.168.1
 13.128:~
[kbhandarkar@192.168.113.128's password:
 ulysses10.txt                                          100% 1529KB  23.8MB/s   00:00
[Karans-MacBook-Air:Centos Shared Folder karanbhandarkar$ ssh kbhandarkar@192.168.113.128
[kbhandarkar@192.168.113.128's password:
 Last login: Thu Sep 21 22:04:14 2017 from 192.168.113.1
[[kbhandarkar@localhost ~]$ ls -ltr
 total 1532
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Desktop
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Templates
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Public
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Downloads
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Documents
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Videos
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Pictures
 drwxr-xr-x. 2 kbhandarkar kbhandarkar       6 Sep 21 21:10 Music
 -rw-r--r--. 1 kbhandarkar kbhandarkar 1565217 Sep 21 22:06 ulysses10.txt
```

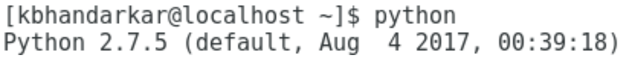
Steps followed to verify Spark 2.2 requirements:

- For Spark 2.2 to run you must have Java 1.8+ installed. Verify java version:



```
[kbhandarkar@localhost ~]$ java -version
openjdk version "1.8.0_131"
OpenJDK Runtime Environment (build 1.8.0_131-b12)
OpenJDK 64-Bit Server VM (build 25.131-b12, mixed mode)
```

- Spark 2.2 needs Python 2.7+ or 3.4+



```
[kbhandarkar@localhost ~]$ python
Python 2.7.5 (default, Aug  4 2017, 00:39:18)
```

You need pip utility. The python-pip package for CentOS is in EPEL. Run commands:

```
yum --enablerepo=extras install epel-release
sudo pip install --upgrade pip
sudo yum install python-wheel
```

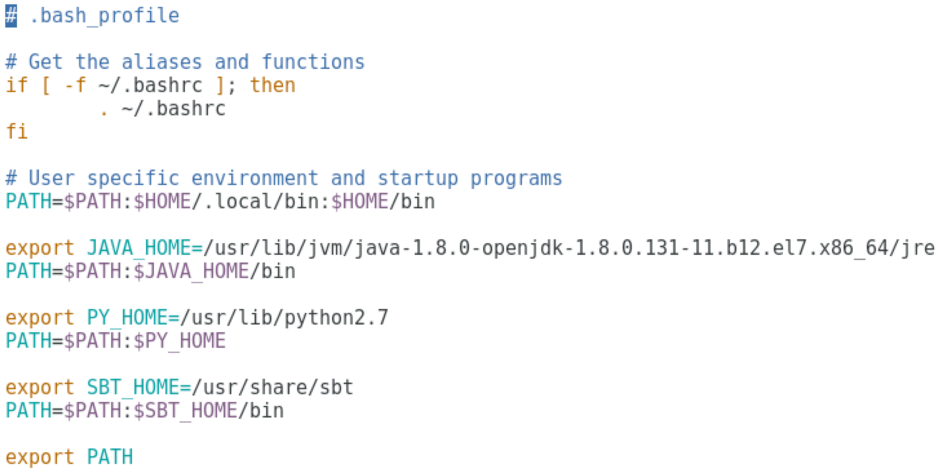
- Spark 2.2 needs Scala 2.11+. Run commands:

```
sudo curl https://bintray.com/sbt/rpm/rpm > /etc/yum.repos.d/bintray-sbt-rpm.repo
sudo yum install sbt
sbt
```

Assignment 3
CSCI E 63 – Big Data Analytics

Steps followed to set up PATH environment variable

Update file .bash_profile in your home directory as below:



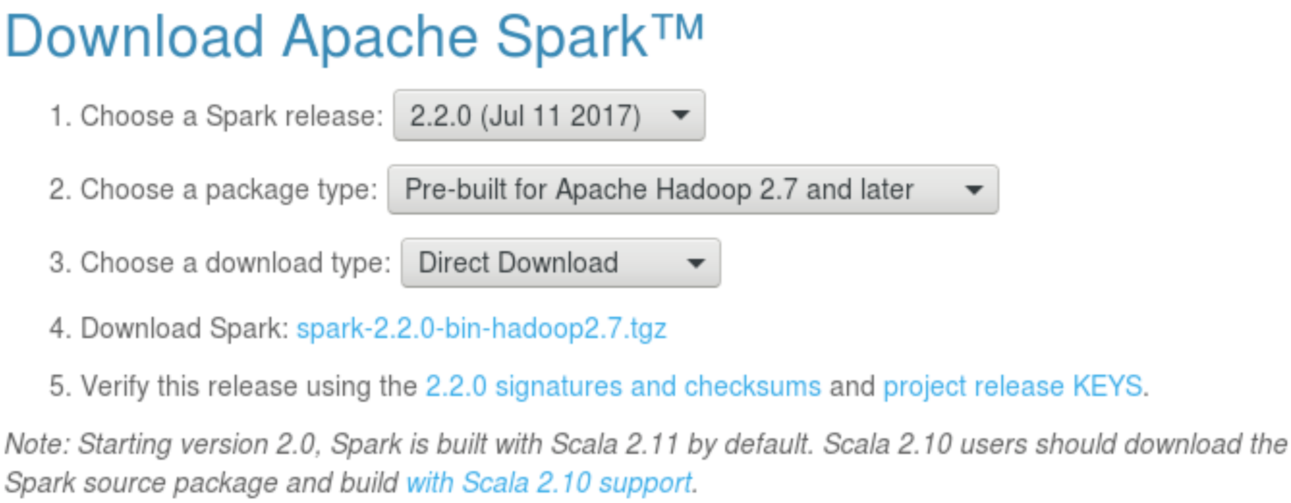
```
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
        . ~/.bashrc
fi

# User specific environment and startup programs
PATH=$PATH:$HOME/.local/bin:$HOME/bin

export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.131-11.b12.el7.x86_64/jre
PATH=$PATH:$JAVA_HOME/bin

export PY_HOME=/usr/lib/python2.7
PATH=$PATH:$PY_HOME

export SBT_HOME=/usr/share/sbt
PATH=$PATH:$SBT_HOME/bin

export PATH
```

Assignment 3
CSCI E 63 – Big Data Analytics

Problem 2:
Install Spark 2.2 on your VM. Make sure that pyspark is also installed. Demonstrate that you can successfully open spark-shell and that you can eliminate most of WARNing messages.

Answer

- Download Spark 2.2.2 from https://spark.apache.org/downloads.html as below:

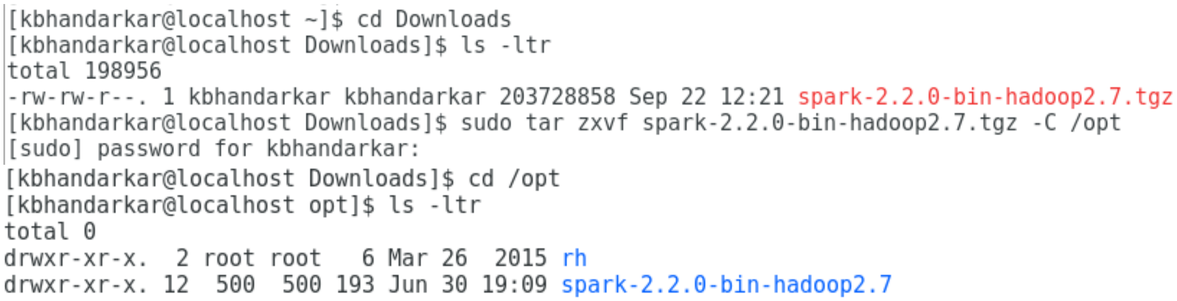


Download Apache Spark™

1. Choose a Spark release: 2.2.0 (Jul 11 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Choose a download type: Direct Download
4. Download Spark: spark-2.2.0-bin-hadoop2.7.tgz
5. Verify this release using the 2.2.0 signatures and checksums and project release KEYS.

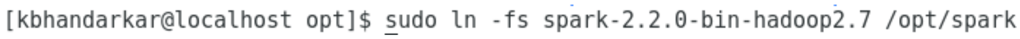
Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

- Unpack Spark in the /opt directory and validate



```
[kbhandarkar@localhost ~]$ cd Downloads
[kbhandarkar@localhost Downloads]$ ls -ltr
total 198956
-rw-rw-r--. 1 kbhandarkar kbhandarkar 203728858 Sep 22 12:21 spark-2.2.0-bin-hadoop2.7.tgz
[kbhandarkar@localhost Downloads]$ sudo tar zxvf spark-2.2.0-bin-hadoop2.7.tgz -C /opt
[sudo] password for kbhandarkar:
[kbhandarkar@localhost Downloads]$ cd /opt
[kbhandarkar@localhost opt]$ ls -ltr
total 0
drwxr-xr-x.  2 root root    6 Mar 26  2015 rh
drwxr-xr-x. 12  500  500 193 Jun 30 19:09 spark-2.2.0-bin-hadoop2.7
```

- Create a symbolic link to make it easier to access



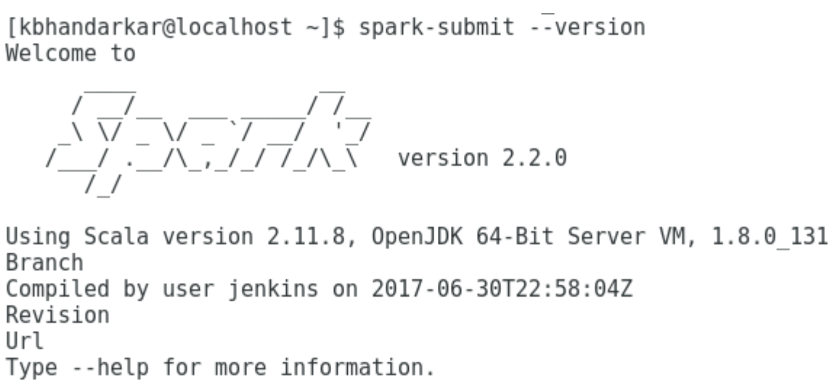
```
[kbhandarkar@localhost opt]$ sudo ln -fs spark-2.2.0-bin-hadoop2.7 /opt/spark
```

- Set the SPARK_HOME environment variable so it takes effect when you login to your VM.
  Add the lines to .bash_profile:

  export SPARK_HOME=/opt/spark
  PATH=$PATH:$SPARK_HOME/bin
  Export PATH

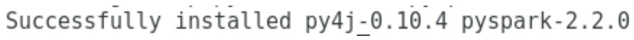
- Reload the environment variable using $ source ~/.bash_profile

- Confirm that spark-submit is now in the PATH



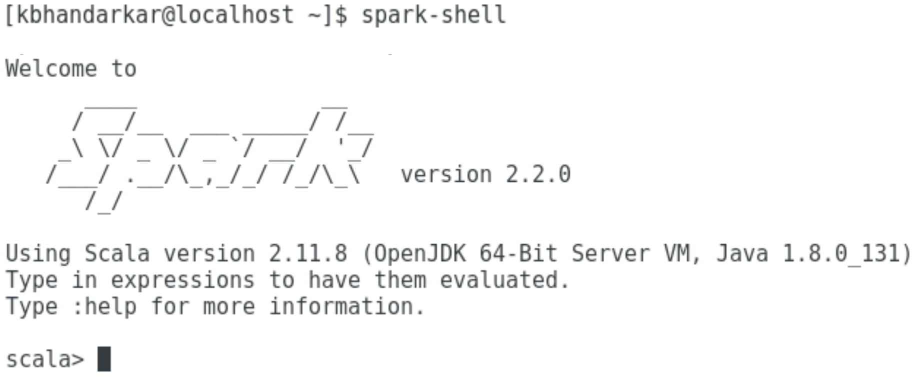
```
[kbhandarkar@localhost ~]$ spark-submit --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Scala version 2.11.8, OpenJDK 64-Bit Server VM, 1.8.0_131
Branch
Compiled by user jenkins on 2017-06-30T22:58:04Z
Revision
Url
Type --help for more information.
```

- Install pyspark. Run the commands:
  sudo pip install - -upgrade setuptools
  sudo pip install ez_setup
  sudo pip install pyspark - -no-cache-dir



```
Successfully installed py4j-0.10.4 pyspark-2.2.0
```

- Open spark shell using command spark-shell

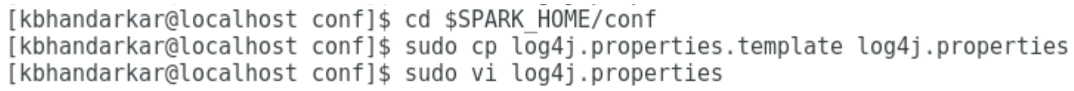


```
[kbhandarkar@localhost ~]$ spark-shell

Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_131)
Type in expressions to have them evaluated.
Type :help for more information.

scala> █
```

•There are a lot of WARNing messages before the welcome message. These can be removed by adjusting the log4j.properties file. You create that file by copying provided file log4j.properties.template and by changing line:

log4j.rootCategory=INFO, console to read: log4j.rootCategory=ERROR, console

That lowered the logging level so that we see only the ERROR messages, and above.

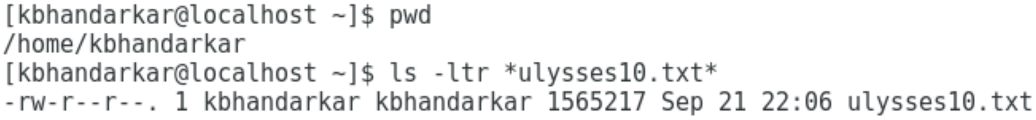


```
[kbhandarkar@localhost conf]$ cd $SPARK_HOME/conf
[kbhandarkar@localhost conf]$ sudo cp log4j.properties.template log4j.properties
[kbhandarkar@localhost conf]$ sudo vi log4j.properties
```

Problem 3:
Find the number of lines in the text file ulysses10.txt that contain word “afternoon” or “night” or “morning”. In this problem use RDD API. Do this in two ways, first create a lambda function which will test whether a line contains any one of those 3 words. Second, create a named function in the language of choice that returns TRUE if a line passed to it contains any one of those three words. Demonstrate that the count is the same. Use pyspark and Spark Python API. If convenient you are welcome to implement this problem in any other language: Scala, Java or R.

Answer:

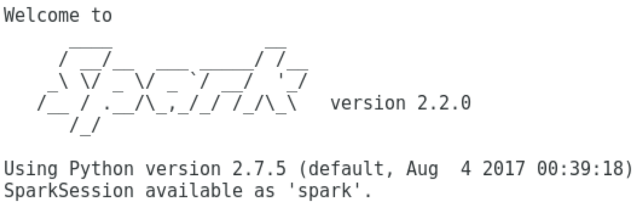
As part of problem1, the text file ulysses10.txt was transferred to the VM.



```
[kbhandarkar@localhost ~]$ pwd
/home/kbhandarkar
[kbhandarkar@localhost ~]$ ls -ltr *ulysses10.txt*
-rw-r--r--. 1 kbhandarkar kbhandarkar 1565217 Sep 21 22:06 ulysses10.txt
```

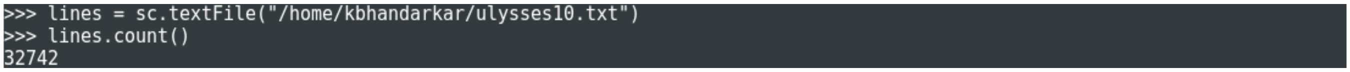
In Spark, we express our computation through operations on distributed collections that are automatically parallelized across the cluster. These collections are called resilient distributed datasets, or RDDs. When we load some data, i.e. a file into a shell variable, we are creating an RDD.

The Spark Python API (PySpark) exposes the Spark programming model to Python. Start pyspark using the command: pyspark



```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 2.7.5 (default, Aug  4 2017 00:39:18)
SparkSession available as 'spark'.
```

Load the ulysses.txt data using the spark context.



```
>>> lines = sc.textFile("/home/kbhandarkar/ulysses10.txt")
>>> lines.count()
32742
```

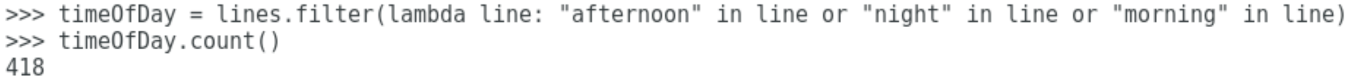
RDD method filter() takes a function returning True or False and applies it to a sequence (list) and returns only those members of the sequence for which the function returned True.

So, in the Python code :
timeOfDay = lines.filter(lambda line: "afternoon" in line or "night" in line or "morning" in line)

Assignment 3
CSCI E 63 – Big Data Analytics

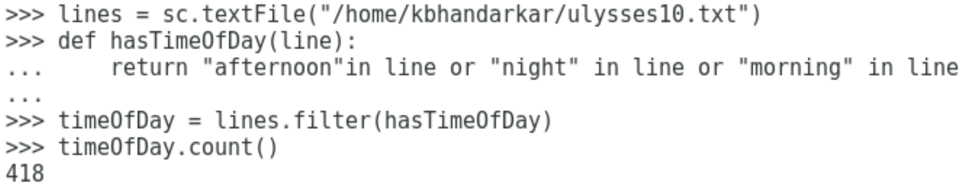
Method filter() acts on the collection lines, and passes every element of that collection as the variable line as the argument to the anonymous function created using lambda construct. That anonymous function uses a simple regular expression to test whether string “afternoon/night/morning“ exists in variable line.

If the regular expression returns True for a particular line, an element of collection lines, the anonymous function will return True and for that particular line, filter() will return/add variable line building up a new collection called heavens.



```
>>> timeOfDay = lines.filter(lambda line: "afternoon" in line or "night" in line or "morning" in line)
>>> timeOfDay.count()
418
```

Rather than using lambda constructs we could define named functions and then pass their names to Spark.



```
>>> lines = sc.textFile("/home/kbhandarkar/ulysses10.txt")
>>> def hasTimeOfDay(line):
...     return "afternoon"in line or "night" in line or "morning" in line
...
>>> timeOfDay = lines.filter(hasTimeOfDay)
>>> timeOfDay.count()
418
```

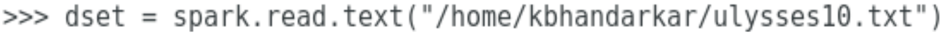
As you can see, the count from both the methods is the same i.e. 418.

Assignment 3
CSCI E 63 – Big Data Analytics

Problem 4:
Implement the above task, finding the number of lines with one of those three words in file ulysses10.txt using Dataset/DataFrame API. Again, use the language of your choice.

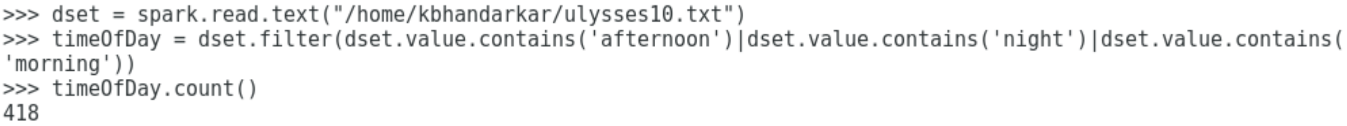
Answer

We can load data into a dataset as:



```
>>> dset = spark.read.text("/home/kbhandarkar/ulysses10.txt")
```

We want to extract all lines in the dataset with the words ‘afternoon’, ‘night’, ‘morning’. This can be done as:

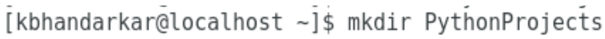


```
>>> dset = spark.read.text("/home/kbhandarkar/ulysses10.txt")
>>> timeOfDay = dset.filter(dset.value.contains('afternoon')|dset.value.contains('night')|dset.value.contains(
'morning'))
>>> timeOfDay.count()
418
```

Problem 5:
Create a standalone Python script that will count all words in file ulysses10.txt. You are expected to produce a single number. Do it using RDD API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R.

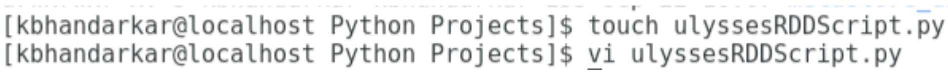
Answer:

Create a ‘PythonProjects’ folder for future python projects.



```
[kbhandarkar@localhost ~]$ mkdir PythonProjects
```

Create a python script as ulyssesRDDScript.py

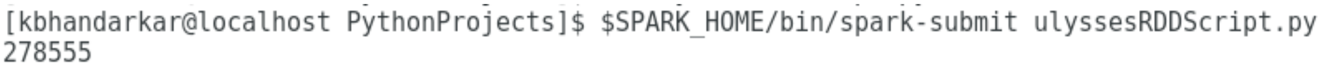


```
[kbhandarkar@localhost Python Projects]$ touch ulyssesRDDScript.py
[kbhandarkar@localhost Python Projects]$ vi ulyssesRDDScript.py
```

Script code:

```
from pyspark import SparkConf, SparkContext
conf = (SparkConf()
        .setMaster("local")
        .setAppName("Word count app")
        .set("spark.executor.memory", "1g"))
sc = SparkContext(conf = conf)
lines = sc.textFile("ulysses10.txt")
words = lines.flatMap(lambda x: x.split(' '))
print(words.count())
```

Running python applications through 'pyspark' is not supported as of Spark 2.0. We use spark-submit.

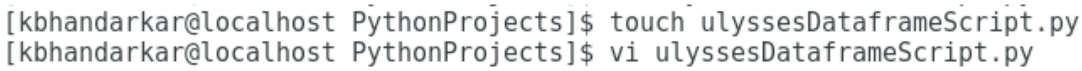


```
[kbhandarkar@localhost PythonProjects]$ $SPARK_HOME/bin/spark-submit ulyssesRDDScript.py
278555
```

Problem 6:
Create a standalone Python script that will count all words in file `ulysses10.txt`. You are expected to produce a single number. Do it using Dataset/DataFrame API. If convenient, you are welcome to implement this problem in other languages: Scala, Java or R.

Answer

Create a python script as ulyssesDataframeScript.py

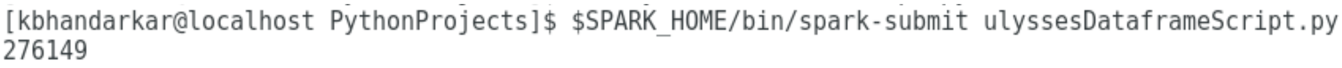


```
[kbhandarkar@localhost PythonProjects]$ touch ulyssesDataframeScript.py
[kbhandarkar@localhost PythonProjects]$ vi ulyssesDataframeScript.py
```

Script code:

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as fn
from pyspark.sql.functions import split
from pyspark.sql.functions import explode
from pyspark.sql.functions import col

spark = SparkSession.builder.master("local").appName("Word Count").getOrCreate()

df = spark.read.text("ulysses10.txt")

wordDF=df.select(explode(split(col("value"), "\s+")).alias("value"))
print(wordDF.count())
```

Running python applications through 'pyspark' is not supported as of Spark 2.0. We use spark-submit.



```
[kbhandarkar@localhost PythonProjects]$ $SPARK_HOME/bin/spark-submit ulyssesDataframeScript.py
276149
```