Question 1

1.
The command **chmod 400 "dsci2024.pem"** is used to change the access permissions of the file ‘dsci2024.pem’ to read only.
The first argument is a Unix/Linux command that is used to change the access permissions of files and directories in file system. The name of it is short for **ch**ange **mod**e.

The second argument **400** is the numerical format permission that accepts up to 4 digits. Here we only use three. This is the argument where the users can set, to specify the modes for this file.
There are 3 kinds of permission types, r(read), w(write) and x(execute), each is represented by a different digit.

Read	(r)	– 4
Write	(w)	– 2
Execute	(x)	– 1

Together they can indicate the access permission of a file system object by using only 1 digit.
e.g. A value of 7 means a *class* is allowed to read, write and execute;
6 means read and execute;
0 means this group is disallowed to this object.
So that is why when professor shows us the file mode of **"dsci2024.pem"**, it is

```
-r--------
```

It can be sliced into three parts, each accords to the digit 4, 0 and 0.

The last argument is the name of the file that can be found in the current directory. This is why the professor keep on reminding us where you stored the pem file and how you should be able to locate it in the terminal.

2.
The command

```
ssh -i "dsci2024.pem" ubuntu@ec2-54-183-13-46.us-west-1.compute.amazonaws.com
```

is used to log into the remote site using the username ubuntu.

The first arugument **ssh** is short for secure shell. It is a connection protocal that provides a secure way for a remote computer to access a Unic/Linux shell.

The second argument ***-i*** is the option that means the user is using an identity_file.

The third argument “dsci2024.pem” is the identity file used to stores your SSH keypairs. (which is generated in the aws website.)

The last and longest argument is used to set up the **username** and **remotehost**, separated by ‘@’. In this example, the username is ubuntu, and the public DHS of the remotehost is ec2-54-183-13-46.us-west-1.compute.amazonaws.com.

3.

wget https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
This command is used to download python package spark using GNU Wget.

The first argument **wget** is an Internet protocol that is widely used for retrieving content from web servers.

The second argument is a https url where the python spark package is located. This is the address that wget needs to find and download it in terminal.

4.

tar xvf spark-3.5.0-bin-hadoop3.tgz
This command is used to uncompress the hadoop tgz file just downloaded from the web.

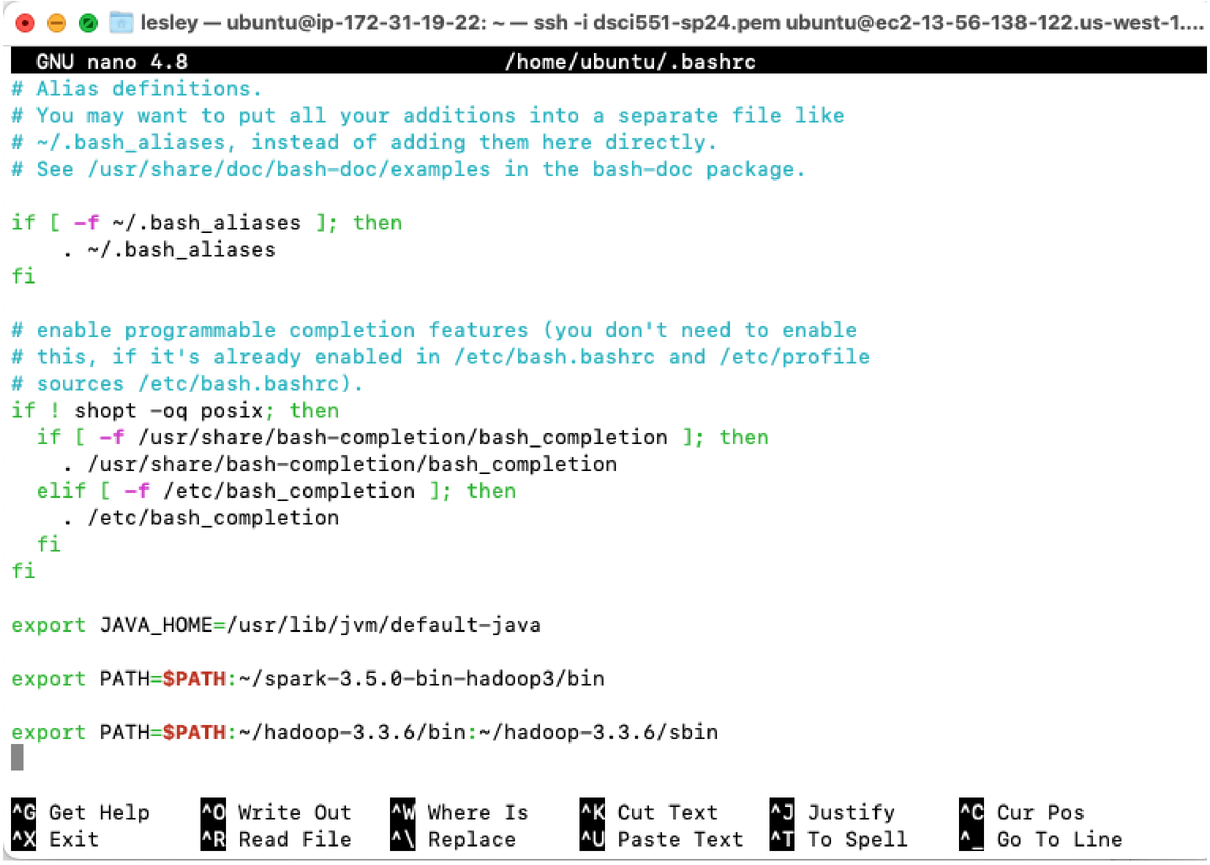
'Tar' is a tool in GNU core-utilities package that creates or extracts archived files. The name tar is short for **ta**pe **ar**chive.

From the documentation we can find the syntax of 'tar' command in Linux:

```
    tar [options] [archive-file] [file or directory to be archived]
```

The second argument **xvf** is a short-hand for three separate **options** that the user can choose to set.
The character **-x** means to extract files and diretories from an existing archive.
The character **-v** means to display verbose info of details during the extraction process, which is convenient for debugging for anything went wrong.
The last character **-f** specifies the filename to be extracted (or created). In this case, the file is **spark-3.5.0-bin-hadoop3.tgz**. This also explains what the third argument does.

Question 2


lesley — ubuntu@ip-172-31-19-22: ~ — ssh -i dsci551-sp24.pem ubuntu@ec2-13-56-138-122.us-west-1....
GNU nano 4.8
/home/ubuntu/.bashrc

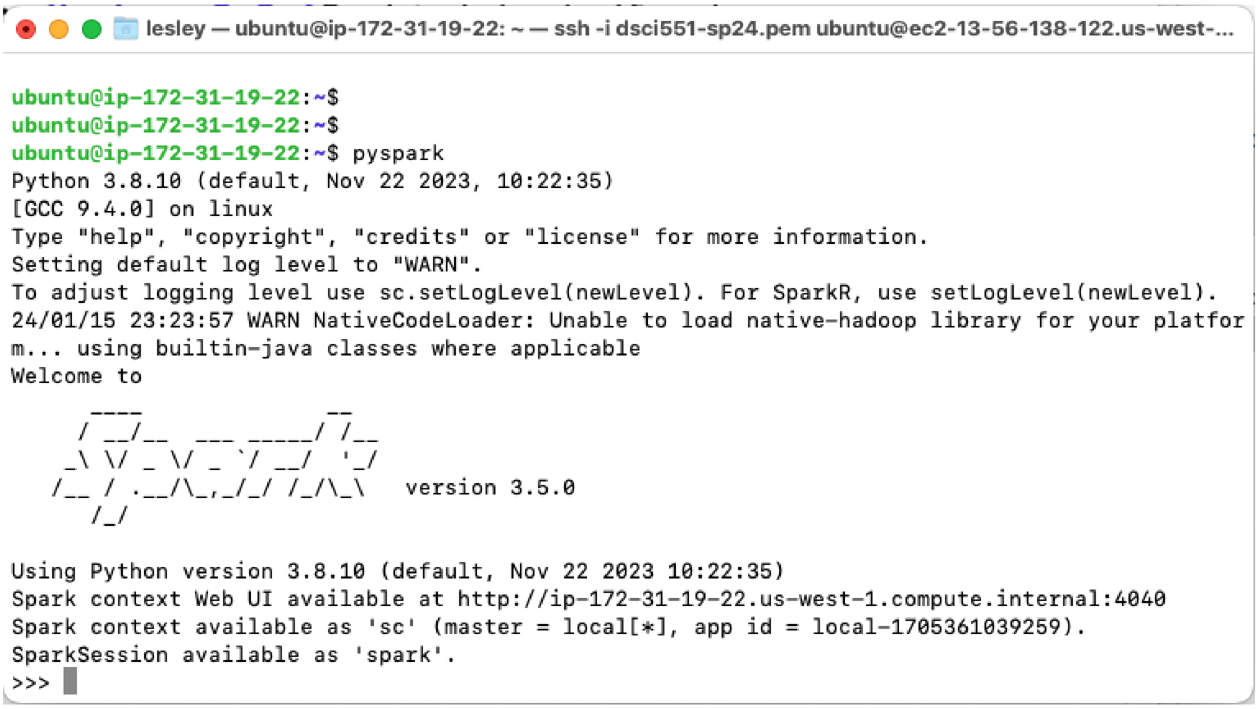
```
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export JAVA_HOME=/usr/lib/jvm/default-java

export PATH=$PATH:~/spark-3.5.0-bin-hadoop3/bin

export PATH=$PATH:~/hadoop-3.3.6/bin:~/hadoop-3.3.6/sbin
```

The above two lines of code are all added into .bashrc file. This script file contains a series of configurations for the terminal session when the user logs in. The configurations includes coloring, shell history, default paths etc. That is why we are writing them into this file.

The command `export JAVA_HOME=/usr/lib/jvm/default-java` sets up the default path of java, so that everytime the user is trying to run java, this version of java will be executed.
Here **JAVA_HOME** is the environment variable. And **/usr/lib/jvm/default-java** is the path.

The command `export PATH=$PATH:~/spark-3.5.0-bin-hadoop3/bin` added a version of python spark package into the system's default path, so that whenever the user tries to run spark in the command line, the system can find these executables from this specific path quickly.
Here $PATH is an essential variable in LINUX systems. It tell the shell where to find the executable files or a list of directories where those files are at. In this case, the directory is **~/spark-3.5.0-bin-hadoop3/bin**, where the python spark files are located.

Question 3


lesley — ubuntu@ip-172-31-19-22: ~ — ssh -i dsci551-sp24.pem ubuntu@ec2-13-56-138-122.us-west-...

```
ubuntu@ip-172-31-19-22:~$
ubuntu@ip-172-31-19-22:~$
ubuntu@ip-172-31-19-22:~$ pyspark
Python 3.8.10 (default, Nov 22 2023, 10:22:35)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/01/15 23:23:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfor
m... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Python version 3.8.10 (default, Nov 22 2023 10:22:35)
Spark context Web UI available at http://ip-172-31-19-22.us-west-1.compute.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1705361039259).
SparkSession available as 'spark'.
>>>
```