

Lab 1: EC2

DSCI 551 – Spring 2024

1. [4 points] Explain what each of the following commands does, and what each of its arguments means.

1. `chmod 400 "dsci2024.pem"`
2. `ssh -i "dsci2024.pem" ubuntu@ec2-54-183-13-46.us-west-1.compute.amazonaws.com`
3. `wget https://d1cdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz`
4. `tar xvf spark-3.5.0-bin-hadoop3.tgz`

Solution:

The detailed explanation of the commands is given below:

1. `chmod 400 "dsci2024.pem"` - This command is used to set read-only permissions for the owner on the file "dsci2024.pem"
 - a. `chmod`: `chmod` means change mode, mainly used to change the permission of a file or directory.
 - b. `400`: Represents the permission assigned to the file "dsci2024.pem", which means it grants read-only access to the owner file without permission for the group and others.
 - c. `dsci2024.pem`: This is a file to which permissions are granted.
2. `ssh -i "dsci2024.pem" ubuntu@ec2-54-183-13-46.us-west-1.compute.amazonaws.com` - Establish a secure ssh connection to a remote server using a specified private key and username
 - a. `ssh`: `ssh` means secure shell and is used to connect to a remote server securely
 - b. `-i`: This is an identity file
 - c. `dsci2024.pem`: Specifies the private key file which can be used for authentication purposes
 - d. `Ubuntu`: This is the username used to login to the remote server
 - e. `@ec2-54-183-13-46.us-west-1.compute.amazonaws.com`: The address of the remote server
3. `wget https://d1cdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz` - Download the Apache spark distribution archive from a specified URL
 - a. `wget`: `wget` means stands for web get and is used to download files from the internet
 - b. `https://d1cdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz`: URL of Apache spark distribution archive to be downloaded.

4. tar xvf spark-3.5.0-bin-hadoop3.tgz - Used to extract the contents of the Spark archive
 - a. tar: tar means tape archive and is used for manipulating archives in Unix-like operating systems
 - b. xvf: Options for extracting (x), being verbose (v), and specifying the archive file (f)
 - c. spark-3.5.0-bin-hadoop3.tgz: This is the name of the archive file to be extracted

2. [4 points] Please add the following lines to the end of ~/.bashrc file (i.e., .bashrc file under the home directory). Show a screenshot of the file with modified content [last few lines of the file].

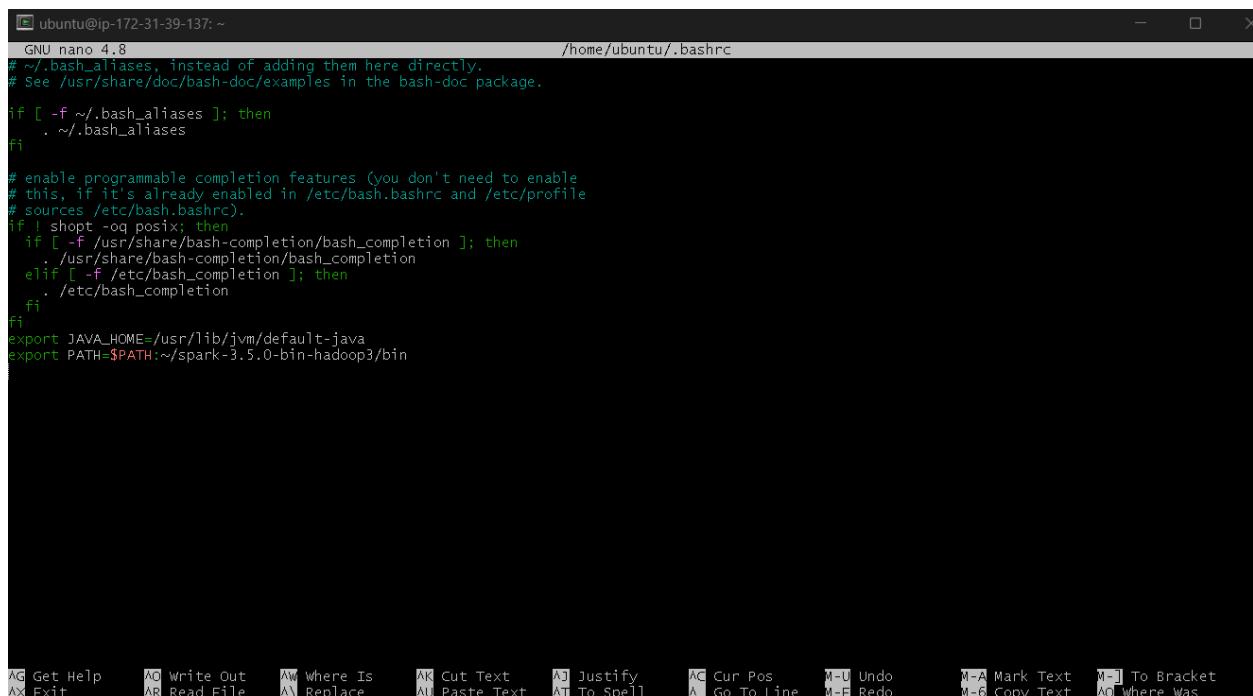
Explain what each of the following commands does.

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

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

Screenshot:

By following the instructions given in the manual above 2 lines were added to the bash file.



```
ubuntu@ip-172-31-39-137: ~
GNU nano 4.8 /home/ubuntu/.bashrc
# ~/.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
```

Solution:

1. export JAVA_HOME=/usr/lib/jvm/default-java
 - a. export: The "export" command in the shell is utilized to define an environment variable, and in this instance, it is employed to establish the JAVA_HOME variable.
 - b. JAVA_HOME=/usr/lib/jvm/default-java: In the command, "JAVA_HOME=/usr/lib/jvm/default-java," this segment involves assigning the

JAVA_HOME variable to the path /usr/lib/jvm/default-java. Widely utilized in various applications and scripts, the JAVA_HOME variable plays a crucial role in determining the directory of the Java installation.

Upon executing this command, the JAVA_HOME variable becomes configured for the ongoing shell session. It's noteworthy that this modification exclusively influences the environment of the current shell and any offspring processes initiated from it. To ensure the persistence of this change across multiple sessions, it might be necessary to include this line in a shell configuration file such as .bashrc or .bash_profile, depending on the specific shell in use.

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

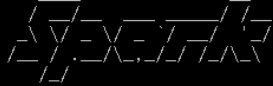
- a. export: The "export" command in the shell is utilized to define an environment variable, and in this instance, it is employed to establish the PATH variable.
- b. PATH=\$PATH:~/spark-3.5.0-bin-hadoop3/bin: The command "PATH=\$PATH:~/spark-3.5.0-bin-hadoop3/bin" alters the PATH environment variable. This portion involves assigning the current value of the path variable, represented by \$PATH. The addition of " ~/spark-3.5.0-bin-hadoop3/bin" appends the bin directory of Apache Spark to the PATH. Notably, the tilde (~) serves as a shorthand for the user's home directory.

The command "export PATH=\$PATH:~/spark-3.5.0-bin-hadoop3/bin" is employed to augment the system's PATH environment variable by appending the bin directory of Apache Spark to it.

3. [2 points] Install Spark by following the instructions in the handout (note the handout has been updated and make sure you use the latest version on d2l) and submit a screenshot showing the successful starting up of pyspark. [screenshot should include the prompt from "pyspark" command]

Screenshot:

By following the instructions given in the manual pyspark was installed

```
ubuntu@ip-172-31-39-137: ~  
spark-3.5.0-bin-hadoop3/bin/sparkR.cmd  
spark-3.5.0-bin-hadoop3/bin/spark-shell2.cmd  
spark-3.5.0-bin-hadoop3/bin/load-spark-env.cmd  
spark-3.5.0-bin-hadoop3/bin/run-example  
spark-3.5.0-bin-hadoop3/bin/sparkR2.cmd  
spark-3.5.0-bin-hadoop3/bin/beeline.cmd  
spark-3.5.0-bin-hadoop3/bin/find-spark-home  
ubuntu@ip-172-31-39-137:~$ 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).  
24/01/17 00:51:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... 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-172-31-39-137.us-east-2.compute.internal:4040  
Spark context available as 'sc' (master = local[*], app id = local-1705452720880).  
SparkSession available as 'spark'.  
>>> |
```