Tutorial for Running Hadoop/MapReduce (v3.1.2)

1) Install Oracle VirtualBox
https://www.virtualbox.org/wiki/Downloads
[take screenshot of VirtualBox running]

2) Install Ubuntu in a Virtual Machine (VM) on VirtualBox

- click on Add button
- click on Expert mode
- put in Name (e.g. Ubuntu)
- select Linux as Type
- select Ubuntu 64-bit as Version (or 32-bit if that's the system you have)
- select 4GB as the memory. It could be less if only have 4GB in your system, but make it at least 3GB.
- click Create and then create the hard drive image (VDI, dynamically allocated)

[take screenshot of your VM configuration]

Download Ubuntu iso image:
https://www.ubuntu.com/download/desktop

Start the Ubuntu VM by double-clicking on it

- select the iso image
- install ubuntu
- select minimal installation
- click install... install now...
- go through rest of dialog boxes (region selection, username info)
- make sure to write down your username and password in a safe location in case you forget it

once Linux starts you may need to install guest additions to get full features (go through the top menu); you may also need to increase the processor CPU cores and video RAM through settings. You will need to shut down the VM to do that first.

[take screenshots of Ubuntu VM running – must show your username]

3) Start Terminal in Ubuntu VM
Click on Apps (lower-left corner), then Terminal

4) Install Java JDK
type the following in Terminal:

```
sudo apt update
sudo apt install default-jre
```

```
sudo apt install default-jdk
```

then check versions:

```
javac -version
java -version
```

[take screenshot showing Terminal output at this point]

5) Install prerequisite software
type the following in Terminal:

```
sudo apt-get install ssh
sudo apt-get install pdsh
```

5) Install Hadoop
use the follow directions to setup a pseudo-distributed Hadoop installation:
https://hadoop.apache.org/docs/r3.1.2/hadoop-project-dist/hadoop-common/SingleCluster.html

Note that the JAVA_HOME must be set to current version in the configuration files. This should be the following:
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64

Note: the namenode web interface port was changed in latest version of Hadoop
http://localhost:9870

[take screenshot of the browser window showing the namenode web interface]

Note that Hadoop will not work well with Java 11. For example, you won't be able to browse the file system. You can fix that by typing this in:

```
export HADOOP_CONF_DIR=~/hadoop/hadoop-3.1.2/etc/hadoop
URL=https://jcenter.bintray.com/javax/activation/javax.activatio
  n-api/1.2.0/javax.activation-api-1.2.0.jar
sudo wget $URL -P $HADOOP_CONF_DIR/lib
sudo echo 'export HADOOP_CLASSPATH+="
  $HADOOP_CONF_DIR/lib/*.jar"' >> $HADOOP_CONF_DIR/hadoop-env.sh
```

6) Do the MapReduce tutorial up to running WordCount 1.0 - show output
https://hadoop.apache.org/docs/r3.1.2/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

if previously, you copied any directories into the input directory, those need to be removed, e.g.:

```
bin/hadoop fs -rm -r /user/username/input/shellprofile.d
```

If you are having trouble compiling and running WordCount based on the tutorial, try this:

```
export HADOOP_CLASSPATH=$(bin/hadoop classpath)
javac -classpath ${HADOOP_CLASSPATH} -d WordCount/
   WordCount.java
jar -cvf WordCount.jar -C WordCount/ .
bin/hadoop jar WordCount.jar WordCount /user/username/input
   /user/username/output/wordcount_output_dir
```

change username and wordcount_output_dir to appropriate names

[take screenshots showing the mapreduce output and the contents of the output file]