Appendix B. Install Spark-2.2.1 on Ubuntu-16.04 with JDK 8

1. Install JDK 8

Verify Java installation

```
$ java -version
```

If Java is not installed, we install it via the following commands.

```
$ sudo add-apt-repository ppa:webupd8team/java
$ sudo apt-get update && sudo apt-get install oracle-java8-installer
```

It may take some time to download the install. When it is done, set the path as follows.

```
$ sudo gedit /etc/environment
```

Append the following line at the end of the file and save it.

JAVA_HOME="/usr/lib/jvm/java-8-oracle"

2. Install Scala

Download Scala in http://www.scala-lang.org/download/

Other resources

You can find the installer download links for other operating systems, as well as documentation and source code archives for Scala 2.12.4 below.

Archive	System	Size
scala-2.12.4.tgz	Mac OS X, Unix, Cygwin	18.83M
scala-2.12.4.msi	Windows (msi installer)	126.38M
scala-2.12.4.zip	Windows	18.87M
scala-2.12.4.deb	Debian	145.23M
scala-2.12.4.rpm	RPM package	125.81M
scala-docs-2.12.4.txz	API docs	56.52M
scala-docs-2.12.4.zip	API docs	109.65M
scala-sources-2.12.4.tar.gz	Sources	

```
$ cd /home/Spark/Downloads
$ tar xvf scala-2.12.4.tgz
```

Use the following commands to move the Scala files to the directory /usr/local/scala

```
$ su -
Password:
# cd /home/Spark/Downloads/
# mv scala-2.12.4 /usr/local/scala
# exit
```

If you have not set the password for root account, use the following command to set it.

```
$ sudo passwd
```

Set the path for Scala.

```
$ sudo gedit /etc/environment
```

Append the following clause to the end of PATH = "/usr/local/sbin:….." in the file, and save it.

:/usr/local/scala/bin

3. Install Maven (to compile java files)

Download Maven from https://maven.apache.org/download.cgi

Files

Maven is distributed in several formats for your convenience. Simply pick a ready-made binar
the installation instructions. Use a source archive if you intend to build Maven yourself.

In order to guard against corrupted downloads/installations, it is highly recommended to verif
bundles against the public KEYS used by the Apache Maven developers.

	Link	Checksum
Binary tar.gz archive	apache-maven-3.5.2-bin.tar.gz	apache-maven-3.5.2-bin.tar.gz.md5
Binary zip archive	apache-maven-3.5.2-bin.zip	apache-maven-3.5.2-bin.zip.md5
Source tar.gz archive	apache-maven-3.5.2-src.tar.gz	apache-maven-3.5.2-src.tar.gz.md5
Source zip archive	apache-maven-3.5.2-src.zip	apache-maven-3.5.2-src.zip.md5

Extract Maven files.

```
$ cd /home/Spark/Downloads
$ tar xvf apache-maven-3.5.2-bin.tar.gz
```

Use the following commands to move the Maven files to the directory /usr/local/maven

```
$ su -
Password:
# cd /home/Spark/Downloads/
# mv apache-maven-3.5.2 /usr/local/maven
# exit
```

Set the path for Maven.

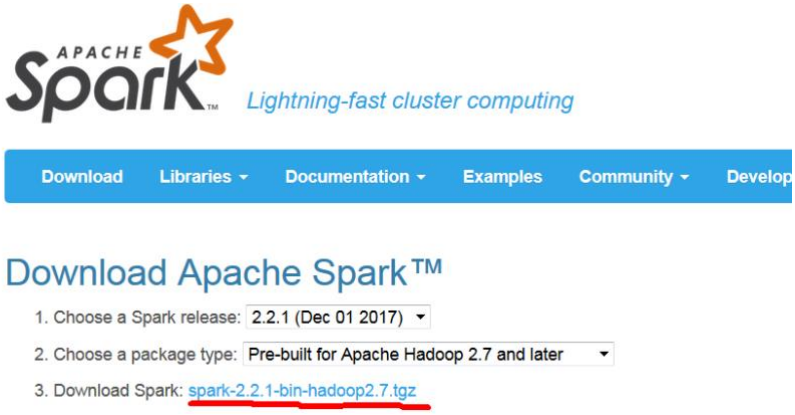
```
$ sudo gedit /etc/environment
```

Append the following clause at the end of PATH = "/usr/local/sbin:….." in the file, and save it.

:/usr/local/maven/bin

4. Install Spark

Download Spark from https://spark.apache.org/downloads.html



Extract the file

```
$ cd /home/Spark/Downloads
$ tar xvf spark-2.2.1-bin-hadoop2.7.tgz
```

Move the Spark files to the directory /usr/local/spark

```
$ su -
Password:
# cd /home/Spark/Downloads/
# mv spark-2.2.1-bin-hadoop2.7 /usr/local/spark
# exit
```

Set the path for Spark.

```
$ sudo gedit /etc/environment
```

Append the following clause at the end of PATH = "/usr/local/sbin:…..", then save it.

:/usr/local/spark/bin

Now, restart the system to make those changes work!

5. Verify the Software Installations

```
$ java -version
```

If Java is installed successfully then you will find the following output.

```
spark@spark-VirtualBox: ~
spark@spark-VirtualBox:~$ java -version
java version "1.8.0_151"
Java(TM) SE Runtime Environment (build 1.8.0_151-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.151-b12, mixed mode)
spark@spark-VirtualBox:~$
```

```
$ scala -version
```

If Scala is installed successfully then you will find the following output.

```
spark@spark-VirtualBox: ~
spark@spark-VirtualBox:~$ scala -version
Scala code runner version 2.12.4 -- Copyright 2002-2017, LAMP/EPFL and Lightbend
, Inc.
spark@spark-VirtualBox:~$
```

```
$ mvn -version
```

If Maven is installed successfully then you will find the following output.

```
spark@spark-VirtualBox:~$ mvn -version
Apache Maven 3.5.2 (138edd61fd100ec658bfa2d307c43b76940a5d7d; 2017-10-18T15:58:1
3+08:00)
Maven home: /usr/local/maven
Java version: 1.8.0_151, vendor: Oracle Corporation
Java home: /usr/lib/jvm/java-8-oracle/jre
Default locale: en_SG, platform encoding: UTF-8
OS name: "linux", version: "4.10.0-42-generic", arch: "amd64", family: "unix"
```

```
$ spark-shell
```

If spark is installed successfully then you will find the following output.

```
spark@spark-VirtualBox:~$ spark-shell
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
18/01/08 18:59:33 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
18/01/08 18:59:33 WARN Utils: Your hostname, spark-VirtualBox resolves to a loop
back address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
18/01/08 18:59:33 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1515409175769
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_151)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Appendix C. My First Spark Program (with Python)

1. Download the example files from Lab1 folder (in.txt, wordcount.py).

2. Create a new folder named ”spark-application” with the files you downloaded. (in.txt, wordcount.py).

3. To execute the Spark program, using the following command under the folder “…/spark-application/”.

```
$ spark-submit wordcount.py in.txt outfile
```

```
spark@spark-VirtualBox:~/Downloads/spark-python/spark-application$ spark-submit
wordcount.py in.txt outfile
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/01/14 12:13:02 WARN Utils: Your hostname, spark-VirtualBox resolves to a loop
back address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
19/01/14 12:13:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
19/01/14 12:13:04 INFO SparkContext: Running Spark version 2.2.1
19/01/14 12:13:05 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
19/01/14 12:13:05 INFO SparkContext: Submitted application: wordcount.py
19/01/14 12:13:05 INFO SecurityManager: Changing view acls to: spark
19/01/14 12:13:05 INFO SecurityManager: Changing modify acls to: spark
```

...

```
19/01/14 12:13:12 INFO SparkContext: Successfully stopped SparkContext
19/01/14 12:13:12 INFO ShutdownHookManager: Shutdown hook called
19/01/14 12:13:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-cd9294
59-3369-4305-af19-9427fde05ea3
19/01/14 12:13:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-cd9294
59-3369-4305-af19-9427fde05ea3/pyspark-4dfe6e40-097f-484d-8900-72346cfb353f
spark@spark-VirtualBox:~/Downloads/spark-python/spark-application$
```

You can see a folder named *outfile* generated under ”.” directory. The result is in the inside file named *part-00000*.

```
part-00000 (~/Downloads/spark-python/spark-application/outfile) - gedit
Open                                                                  Save
(u'beautiful', 2)
(u'', 1)
(u'talk.', 1)
(u'people', 1)
(u'love,', 1)
(u'look,', 1)
(u'as', 8)
(u'walk', 1)
(u'only', 1)
(u'share.', 1)
(u'are', 2)
(u'they', 7)
(u'not', 1)
(u'or', 1)
(u'care', 1)
```