

# Spark Workshop

---

## Recommended Background

Hands-on experience with at least one programming language is required. Experience with Java, C#, C/C++, Python, JavaScript, or any other modern language is sufficient.

## Suggested Reading

The workshop is self-contained. However, we will cover a lot of material over a period of two days; it may feel like drinking water from a firehose. Therefore, it is recommended to read the book "[Big Data Analytics with Spark](#)" by Mohammed Guller before the boot camp. The book is not mandatory, but reading it before the boot camp will help you get more out of the hands-on sessions. You will be better prepared for the programming exercises.

## Operating System

The ideal operating system for developing Spark applications is Linux or Mac OS X. If your laptop is running Windows, it is recommended to install Linux as a guest OS in a virtual machine.

## Install Linux as a guest OS on a Windows machine

1. Download and install VirtualBox.  
<https://www.virtualbox.org>
2. Download and install KUbuntu as a guest OS in a Virtual Machine (VM) inside VirtualBox.  
<http://www.kubuntu.org>
3. Assign at least 2 cores and 2GB memory to the guest OS.
4. Start the Linux VM you installed in step 2.
5. Run the following commands in a terminal (inside KUbuntu) to verify that KUbuntu is installed with the right hardware resources:

```
$ lscpu
```

The output of this command should show that the VM has 2 or more CPU(s).

```
$ free -m
```

The output of this command should show that the VM has 2 GB or more memory.

6. Install all the remaining software listed in this document (JDK, Spark, etc.) on the VM / KUbuntu.

## JDK

JDK 7 or JDK 8 is required.

### Install JDK (skip this step if you already have JDK installed)

7. Download and install JDK

<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

8. Run the following command in a terminal to make sure that JDK is correctly installed

```
$ javac -version
```

## Spark

9. Download Spark binaries (not source code).

<https://spark.apache.org/downloads.html>

Select the following options from the download page:

Spark release: 2.0.2

Package type: Pre-built for Hadoop 2.7 and later

10. Extract the Spark binaries in your home directory.

```
$ tar xzf spark-2.0.2-bin-hadoop2.7.tgz
```

11. Verify that you are able to launch the spark-shell.

```
$ cd spark-2.0.2-bin-hadoop2.7/
```

```
$ ./bin/spark-shell
```

You should see the following:

```
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).
16/11/28 18:58:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/11/28 18:58:35 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 192.168.127.128 instead (on interface eth0)
16/11/28 18:58:35 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
16/11/28 18:58:36 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.
Spark context Web UI available at http://192.168.127.128:4040
Spark context available as 'sc' (master = local[*], app id = local-1480388315638).
Spark session available as 'spark'.
Welcome to

  ____      _
 / ___|    / \
 \___ \  / _ \
  ___) / / ___\
 /____/_/_/___ \
               \_/_

version 2.0.2

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_77)
Type in expressions to have them evaluated.
Type :help for more information.

scala> █
```

If you do not get the Scala prompt as shown above, something is wrong.

## SBT

12. Download and install SBT.

<http://www.scala-sbt.org>

Installation instructions at: <http://www.scala-sbt.org/0.13/tutorial/Manual-Installation.html>

Please go through all the above steps before your arrive to the boot camp. **You will not be able to participate in the hands-on sessions without the above listed software.** If you have any questions, email me at: mohammedguller@yahoo.com

**\*\* Important: Some of the files are large and it takes time to install them. You will not be able to keep up with the rest of the class if you try to install them during the boot camp. \*\***