Dipartimento di Ingegneria Informatica
Automatica e Gestionale Antonio Ruberti

Sapienza
Università di Roma

Laboratory of Advanced Programming

How to install Spark/Scala

Marco Calamo, Matteo Marinacci, Jacopo Rossi
{calamo, marinacci, j.rossi}@diag.uniroma1.it

How to install Spark/Scala

How to install Spark/Scala:

- On Ubuntu/Debian
- using Docker
- On Mac OS
- On Windows

How to install it on Ubuntu/Debian (1/4)

1. **Install Java**

   a. `sudo apt update`

   b. `sudo apt install default-jre`

   c. `java -version`

2. **Install Scala**

   a. `sudo apt search scala` ⇒ Search for the package

   b. `sudo apt install scala` ⇒ Install the package

   c. `scala -version`

How to install it on Ubuntu/Debian (2/4)

3. **Install Apache Spark in Ubuntu**

   a. Now go to the official [Apache Spark download]page and grab a version (i.e. 3.1.1) at the time of writing this article. Alternatively, you can use the [wget command]to download the file directly in the terminal.

```
wget https://apachemirror.wuchna.com/spark/spark-3.1.1/spark-3.1.1-bin-hadoop2.7.tgz
```

   b. Now open your terminal and switch to where your downloaded file is placed and run the following command to extract the Apache **Spark** tar file. Finally, move the extracted Spark directory to **/opt** directory.

```
tar -xvzf spark-3.1.1-bin-hadoop2.7.tgz
sudo mv spark-3.1.1-bin-hadoop2.7 /opt/spark
```

How to install it on Ubuntu/Debian (3/4)

4. **Configure Environmental Variables for Spark**
   Now you have to set a few environmental variables in your **.profile** file before starting up the spark.

```
a.    echo "export SPARK_HOME=/opt/spark" >> ~/.profile
b.    echo "export PATH=$PATH:/opt/spark/bin:/opt/spark/sbin" >> ~/.profile
c.    echo "export PYSPARK_PYTHON=/usr/bin/python3" >> ~/.profile
```

To make sure that these new environment variables are reachable within the shell and available to Apache Spark, it is also mandatory to run the following command to take recent changes into effect.

```
source ~/.profile
```

All the spark-related binaries to start and stop the services are under the **sbin** folder.

```
ls -l /opt/spark
```

How to install it on Ubuntu/Debian (4/4)

5. **Start Apache Spark in Ubuntu**
Run the following command to start the **Spark** master service and slave service.

```
a.  start-master.sh
b.  start-workers.sh spark://localhost:7077
```

Once the service is started go to the browser and type the following URL access spark page. From the page, you can see my master and slave service is started.

- `http://localhost:8080/` OR `http://127.0.0.1:8080`

You can also check if spark-shell works fine by launching the spark-shell command.

- `spark-shell`

Apache Spark and Docker

Follow this guide in order to use Docker:
[Apache Spark and Docker]

How to install it on Mac OS

1. **Get Homebrew**
   You can get Homebrew by following the instructions on it's website. Which basically just tells you to open your terminal and type:

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

2. **Installing xcode-select**
   Go on your terminal and run: `xcode-select --install`
3. **Use Homebrew to install Java**
   Go on your terminal and run: `brew cask install java`
4. **Use Homebrew to install Scala**
   Go on your terminal and run: `brew install scala`
5. **Use Homebrew to install Apache Spark**
   Go on your terminal and run: `brew install apache-spark`
6. **Start the Spark Shell**
   Try this command to start the interactive shell: `spark-shell`

How to install it on Windows (1/4)

Prerequisite — Java 8

1. Before you start make sure you have Java 8 installed and the environment variables correctly defined:

   Download **Java JDK 8** from Java's official website

2. Set the following environment variables:
   a. `JAVA_HOME = C:\Program Files\Java\jdk1.8.0_161`
   b. `PATH += C:\Program Files\Java\jdk1.8.0_161\bin`

Note: In this guide I'll be using my C drive but obviously you can use another drive also

Optional: **`_JAVA_OPTIONS`** `= -Xmx512M -Xms512M` (To avoid common Java Heap Memory problems whith Spark)

How to install it on Windows (2/4)

Spark: Download and Install

1. Download Spark from Spark's official website Choose your release

   Choose the newest package type (e.g. Pre-built for Hadoop 2.7 or later)
2. Download the **.tgz** file
3. Extract the **.tgz** file into `C:\Spark`
4. Set the environment variables. In my case:

```
SPARK_HOME = D:\Spark\
PATH += D:\Spark\bin
```

Note: In this guide I'll be using my C drive but obviously you can use another drive also

How to install it on Windows (3/4)

Spark: Some more stuff (winutils)

1. Download winutils.exe from here: https://github.com/steveloughran/winutils
2. Choose the same version as the package type you choose for the Spark .tgz file you chose in section 2 *"Spark: Download and Install" (e.g. hadoop-2.7.1)*
3. You need to navigate inside the **hadoop-X.X.X folder**, and inside the bin folder you will find **winutils.exe**
4. Move **ONLY the winutils.exe** file to the bin folder inside **SPARK_HOME**, In my case: `C:\Spark\bin`
5. Set the following environment variable to be the same as **SPARK_HOME**:
   a. `HADOOP_HOME = C:\Spark\`

How to install it on Windows (4/4)

Install Scala

If you are planning on using Scala instead of Python for programming in Spark, follow this steps:

1. Download Scala from their official website
   a. Download the Scala binaries for Windows *(.msi file)*
2. Install Scala from the **.msi** file
3. Set the environment variables:

```
a.   SCALA_HOME = C:\Program Files (x86)\scala
b.   PATH += C:\Program Files (x86)\scala\bin
```

4. Check if scala is working by running the following command in the **cmd (run as administrator):**

```
a.   cmd> scala -version
```