How to Install Spark on Ubuntu

Elaborated by Yudith Cardinale

Step 1: Install Java

The default Ubuntu repositories contain both Java 8 and Java 11. Use the following command to install it.

```
sudo apt update && sudo apt install openjdk-8-jdk (or  openjdk-11-jdk)
```

Once you have successfully installed it, check the current Java version:

```
java -version
```

```
yudithcardinale$ java -version

openjdk version "1.8.0_292"
OpenJDK Runtime Environment (AdoptOpenJDK)(build 1.8.0_292-b10)
OpenJDK 64-Bit Server VM (AdoptOpenJDK)(build 25.292-b10, mixed mode)
```

You can find the location of the JAVA_HOME directory by running the following command. That will be required for the configuration.

```
yudithcardinale$ dirname $(dirname $(readlink -f $(which java)))

/usr/lib/jvm/java-8-openjdk-amd64
```

Step 2: Install Python and Scala

Python and Scala installation are needed if you wanted to run PySpark examples (Spark with Python) or Spark with Scala on the Ubuntu server.

```
yudithcardinale$ sudo apt install python3
```

```
yudithcardinale$ sudo apt install curl mlocate git scala -y
```

Step 3: Install Spark

Use the following command to download Spark 3.3.2:

```
wget https://downloads.apache.org/spark/spark-3.3.0/spark-3.3.0-bin-
hadoop3.tgz
```

Once you've downloaded the file, you can unzip it to a folder on your hard drive.

```
tar -xzf spark-3.3.0-bin-hadoop3.tgz
```

```
mv spark-3.3.0-bin-hadoop3 spark
```

Next, you will need to configure Spark and Java Environment Variables on your system. Open the ~/.bashrc file in your favorite text editor:

```
vim ~/.bashrc
```

Append the below lines to the file. You can find the JAVA_HOME location by running dirname $ (dirname $(readlink -f $(which java))) command on the terminal.

```
export SPARK_HOME=/home/ycardinale/spark
export PATH=$PATH:$SPARK_HOME/bin

export PYSPARK_PYTHON=/usr/bin/python3
```

After the installation is complete, enter "**spark-submit --version**" to view the version. If there is an information receipt, the installation is successful.

```
yudithcardinale@MacBookYudith ~ % spark-submit --version

2023-05-21 12:50:39,396 WARN util.Utils: Your hostname, MacBook-Pro-de-Yudith.local
resolves to a loopback address: 127.0.0.1; using 10.88.2.113 instead (on interface
en0)
2023-05-21 12:50:39,400 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.0
      /_/

Using Scala version 2.12.15, OpenJDK 64-Bit Server VM, 1.8.0_292
Branch HEAD
Compiled by user hgao on 2022-01-20T19:26:14Z
Revision 4f25b3f71238a00508a356591553f2dfa89f8290
Url https://github.com/apache/spark
Type --help for more information.
```

Step 5: Spark testing with command line interfaces

Scala:
```
yudithcardinale@MacBookYudith ~ % spark-shell
```

Python:

```
yudithcardinale@MacBookYudith ~ % pyspark
```