Installing Spark On Linux

Step 1 - Update the System

Before starting, you need to update and upgrade all system packages to the latest version. You update all of them by running the following command:

```
apt update -y apt upgrade -y
```

Once all the packages are updated, you can proceed to the next step.

Step 2 - Install Java JDK

Following step with Apache Spark is a Java based application. So you need to install Java JDK on your server. You can install it by running the following command:

```
apt-get install default-jdk -y
```

Once the Java JDK is installed, you can verify the Java version using the following command:

```
java --version
```

Now you should see the Java version in the following output:

```
openjdk 11.0.17 2022-10-18 OpenJDK Runtime Environment (build 11.0.17+8-post-Ubuntu-1ubuntu220.04) OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Ubuntu-1ubuntu220.04, mixed mode, sharing)
```

Step 3 - Install Scala

Then, you also need to install Scala on your server. Install it by running the following command:

```
apt-get install scala -y
```

After the installation, you can verify the Scala version with the following command:

```
scala -version
```

You will get the following output:

```
Scala code runner version 2.11.12 -- right 2002-2017, LAMP/EPFL
```

Now connect to the Scala with the following command:

```
scala
```

At this stage you should see the Scala shell in the following output:

```
Welcome to Scala 2.11.12 (OpenJDK 64-Bit Server VM, Java 11.0.17). Type in expressions for evaluation. Or try :help.
scala>
```

To test the Scala, run the following command:

```
scala> println("Testing Scala")
```

You should get the following output:

```
Testing Scala
```

Press **CTRL+D** to exit from the Scala shell.

Step 4 - Install Apache Spark

First, visit the Apache Spark official download page and download the latest version using the following command:

```
wget https://archive.apache.org/dist/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
```

Once the Apache Spark is downloaded, you can extract the downloaded file with the following command:

```
tar -xvzf spark-3.3.1-bin-hadoop3.tgz
```

Next, move the extracted directory to the **/mnt** directory with the following command:

```
mv spark-3.3.1-bin-hadoop3 /mnt/spark
```

Step 5 - Start Apache Spark

In this step, you will need to edit the **.bashrc** file and define the Apache Spark path. Edit it with the following command:

```
nano ~/.bashrc
```

Add the following lines:

```
export SPARK_HOME=/mnt/spark export
PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin
```

Save and close the file then reload the environment variable with the following command.

```
source ~/.bashrc
```

Next, start the Apache Spark with the following command:

```
start-master.sh
```

You should see the following output:

```
starting org.apache.spark.deploy.master.Master, logging to
/mnt/spark/logs/spark-root-org.apache.spark.deploy.master.Master-
1-spark.out
```

If you want to stop the Apache Spark, run the following command:

```
stop-master.sh
```