Installing Spark

Master M2 – Université Grenoble Alpes & Grenoble INP

2023

The document provides the basic instructions for installing and configuring Spark before the first lab session of the course "Data Management in Large Scale Distributed Systems".

The recommended method is to install Spark directly on your machine if you have a Linux system (or a Linux virtual machine). This method will also work on the computers of the lab rooms at Ensimag.

An alternative method that should work on any recent OS is to use a docker image.

The last method is using Google Colab, an online solution to execute python notebooks. This solution can be a temporary alternative if you are experiencing issues with all the other methods. However, it can only be a temporary solution in our opinion.

The three methods are described below.

1 Native installation of Spark on a Linux machine

Please find below the instructions to install Spark on a recent Linux machine. These instructions are valid both for your laptop and for the machines of the lab rooms.

The instructions are for Spark 3.3.5, the latest version of Spark.

1. Download the latest already compiled version of Spark here: `https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz`

2. Extract the downloaded archive:

   ```
   tar zxvf spark-3.3.5-bin-hadoop3.tgz
   ```

3. Configure the required environment variables in the file `$HOME/.bashrc` by adding the following lines at the beginning of the file[1], where `PATH_TO_DIR` should correspond to the directory where your stored Spark.

   ```
   export SPARK_HOME=PATH_TO_DIR/spark-3.3.5-bin-hadoop3
   ```

[1] To open this file, simply run `nano ~/.bashrc` in a terminal

```
export PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
export PYTHONPATH="${SPARK_HOME}/python/lib/py4j-0.10.9.7-src.zip:$PYTHONPATH"

export PATH=${SPARK_HOME}/bin:$PATH
```

4. Start a new terminal to make your changes active
5. In the new terminal, launch `pyspark` to check that everything works correctly

Troubleshooting

This version of Spark only works with Java 8/11/17. If after executing the previous commands, you experience some problems, it might be that the default Java version in your system is newer than this.

In the lab rooms, you can select the correct Java version to be used by adding the following line to the file `$HOME/.bashrc`:

```
export JAVA_HOME="/usr/lib/jvm/java-17-openjdk-amd64/"
export PATH=${JAVA_HOME}/bin:$PATH
```

About Scala

To use Scala on your laptop, in addition to installing Spark, you will need to install `sbt`, which is a project builder for Scala projects.

To this end, simply follow the instructions here: `https://www.scala-sbt.org/1.x/docs/Installing-sbt-on-Linux.html`. We recommend using the DEB or RPM package if possible.

2 Installing Spark using a Docker container

Please refer to the slides on the LSTD teacing section.

3 Using Spark with Google Colab

A last solution to work with Spark is to use Google Colab. Please refer to this short introduction to start using Spark with Google Colab: `https://colab.research.google.com/drive/1-co8gEHx_EJLURFWfw0WZq1uik0uRfqC?usp=sharing`.