Installing Spark

Master M2 – Université Grenoble Alpes & Grenoble INP

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.

As of today (2025), I have tried with the most recent version which is 4.0.1.

1. Download the latest already compiled version of Spark here: `https://spark.apache.org/downloads.html`

2. Extract the downloaded archive (replace X.Y.Z and AB with the corresponding version numbers of Spark and Hadoop):

   ```
   tar zxvf spark-X.Y.Z-bin-hadoopAB.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-X.Y.Z-bin-hadoopAB

   export PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
   ```

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

```
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 specific Java versions (for 4.0.1 I needed Java 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` (replace with the folder of your Java installation):

```
export JAVA_HOME=FOLDER_OF_JAVA_INSTALL
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

You need Docker to be installed on your machine. This is very probably already done for your Hadoop installation.

More detailed information on the Docker image may be found gere `https://jupyter-docker-stacks.readthedocs.io/en/latest/using/running.html`

To download the image to work with Spark

```
docker pull jupyter/pyspark-notebook
```

Be patient, the image is big and the download will take time.

To run the container (there are other options):

```
docker run -it --rm -p 4040:4040 -p 8888:8888 -v .:/home/jovyan/work jupyter/pyspar
```

This command outputs a trace which finishes with the URL to use to connect to the jupyter server. The current directory is mapped to the /home/jovyan/work directory in the notebook. You need to put your data and code in there to be able to work with them.

In the container, you can either use the Python notebook, or use a terminal (open terminal) and commands on pyspark.

A first test that your install is working may be

```
print(sc.defaultParallelism)
```

You can play with teh WordCount example or test some basic transformation and action operations (see lecture).

You can access the Spark web UI at `http://localhost:4040/`

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`.