CDO: Installing Spark

Master M2 – Université Grenoble Alpes & Grenoble INP

2018

This document provides instructions about how to install Apache Spark on your personal laptop or on the computers from the lab room.

Installing Apache Spark using Docker on your laptop: On your laptop, the recommended solution is to rely on Docker containers. This solution is described in Section 1. With this solution, the only software to be installed on your machine is Docker.

Installing Apache Spark directly on your laptop: You may want to chose to install Apache Spark natively on your machine. In this case, you can follow the instructions provided in Section 2. Note that in this case, in addition to Apache Spark, you will have to install `Jupyter Notebook`. Note also that if you decide to go with this option, we will not provide support regarding missing dependencies or problems related to versions of software during the lab sessions.

Using the machines from the lab rooms: Installing Apache Spark in your account on the machines from the lab room is very easy. Please follow the instructions provided in Section 2.

1 Installing Spark using Docker

To install Spark using Docker on your laptop, please follow the instructions provided in this document: `https://tropars.github.io/downloads/lectures/LSDM/LSDM-Spark-on-your-Laptop.pdf`.

Note that this solution works very well on Linux and MacOS machines. It might also work on Windows but we did not test it and we will not provide support for problems related to Windows machines during the lab sessions.

Using Spark during the lab sessions Once you are done with the described installation process, we can start the docker container by running in a terminal:

```
docker run -v absolute_path_to_folder:/home/jovyan/work -it \
       --rm -p 8888:8888 -p 4040:4040 jupyter/pyspark-notebook
```

- In the previous command, `absolute_path_to_folder` should be replaced by the actual path to the directory where you are going to store your notebooks.
- When you launch this command, logs are written into the terminal. The last displayed line is an `url` that you should copy into a web browser. This `url` allows you to connect to a Jupyter Notebook that gives access to Spark.

2 Installing Spark in the lab rooms

These instructions should actually work on any Linux machine.

Here are the instructions to install and configure Spark in the lab rooms:

1. Download the latest already compiled version of Spark here: `https://www.apache.org/dyn/closer.lua/spark/spark-2.3.2/spark-2.3.2-bin-hadoop2.7.tgz`
2. Extract the downloaded archive:

   ```
   tar zxvf spark-2.3.2-bin-hadoop2.7.tgz
   ```

3. Configure the require environment file by updating the file `$HOME/.bashrc` by adding the following lines at the beginning of the file[1]:

   ```
   export SPARK_HOME=PATH_TO_DIR/spark-2.3.2-bin-hadoop2.7

   export PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
   export PYTHONPATH="${SPARK_HOME}/python/lib/py4j-0.10.7-src.zip:$PYTHONPATH"

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

   export PYSPARK_PYTHON=python3
   ```

   - where `PATH_TO_DIR` corresponds to the directory where your stored Spark.
4. Start a new terminal to make your changes active
5. In the new terminal, launch `pyspark` to check that everything works correctly

Using Spark during the lab session To use Spark, simply start `Jupyter Notebook`.

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