Large-Scale Data Management and Distributed Systems

Installing SPARK on Your Laptop

Vania Marangozova

Vania.Marangozova@imag.fr

2023-2024

Adapted from

- the notes of T.Ropars

Introduction

- These slides provide the basic information to configure your personal laptop for the labs on Apache Spark.

Warning

- The instructions assume that you run on Linux
- It will most probably work also on MacOS
- Not sure about Windows machines

Description

We are going to use a docker container that is already configured with Spark properly installed.

Main steps

1. Installing and configuring Docker
2. Downloading the Spark container
3. Testing that everything works well

Installing Docker

- We will use Docker CE (Community Edition)
- To install Docker CE on Ubuntu, please follow the instructions here:
  - I https://docs.docker.com/install/linux/docker-ce/ ubuntu/#install-docker-ce
  - On the same page you can find instructions for other Linux distributions
- By default, Docker containers can be launched only by the root user.
  - This can inconvenient. To allow a normal user to run a docker container, follow these instructions:
    https: //docs.docker.com/install/linux/linux-postinstall/

More on Docker

- Once you are done with installing Docker, you should test that it works by running the hello world of docker in a terminal:

```
docker run hello-world
```

- If everything works, congrats, you are almost done :-)
- If you want to know more about Docker, you may start from here:
  - https://docs.docker.com/engine/docker-overview/

Getting the Spark Docker Container

- Once Docker is installed, you should pull the Docker image containing Spark that we will use. To do so, run in a terminal:

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

WARNING

- The image is big. It will take time.

Last step: testing that everything works

- To test that you are able to run Spark on your machine, run the following command in a terminal:

```
docker run -it --rm -p 4040:4040 \
        jupyter/pyspark-notebook \
        /usr/local/spark-3.5.0-bin-hadoop3/bin/pyspark
```

pyspark Running?

- At this point, Spark is running on your machine and you have started the Python interactive console (pyspark).
- You can run your first Spark command by checking the default level of parallelism used by Spark on your machine, by typing in pyspark:

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

- You can even connect to the Spark Web UI here: `http://localhost:4040/`
- To exit pyspark and terminate Spark, simply type Ctrl-D

jupyter Notebook

- To use Spark in a jupyter notebook

```
docker run -it --rm -p 8888:8888
                 jupyter/pyspark-notebook
```

- Read what is written on the terminal and connect to the Jupyter server

```
http://127.0.0.1:8888/lab?token=...
```

- You can import the data, create the notebook, ...
  - to save your progress, create a new image container

```
docker ps -a
docker commit containerID newImageID
docker images -a
```

You are ready for the lab!