

# TP 3 : Spark et le calcul parallèle

Introduction to Big Data & Data Engineering

ESME SUDRIA

Luc Marchand - [luc.marchand.pro@proton.me](mailto:luc.marchand.pro@proton.me)

Maxence Talon - [maxencetallon@gmail.com](mailto:maxencetallon@gmail.com)



# Global Syllabus

01

Introduction and main concepts

02

SQL, set up env and practical work

03

NoSQL world

04

Introduction to Big Data & Data Engineering

05

Kafka & event driven architectures

06

**Spark & Delta**

07

Warehouse, DBT & BI

08

IA - MLOps & RAG



# Pré-requis et installation

Pré - requis : docker

- 1 - Télécharger l'image docker : `docker pull quay.io/jupyter/pyspark-notebook:python-3.12`
- 2 - Lancer la commande suivante : `docker run --detach --name spark_TP -p 8888:8888 -p 4040:4040 -p 4041:4041 quay.io/jupyter/pyspark-notebook:python-3.12`
- 3 - regarder les logs du docker pour retrouver les urls de connexion avec la commande : `docker logs spark_TP`
- 4 - importer le notebook du tp

```
maxence.tallon@HEL-DF-XCFD7FJ27P:~$ docker logs spark_TP
Entered start.sh with args: start-notebook.py
Running hooks in: /usr/local/bin/start-notebook.d as uid: 1000 gid: 100
Done running hooks in: /usr/local/bin/start-notebook.d
Running hooks in: /usr/local/bin/before-notebook.d as uid: 1000 gid: 100
Sourcing shell script: /usr/local/bin/before-notebook.d/10activate-conda-env.sh
Sourcing shell script: /usr/local/bin/before-notebook.d/10spark-config.sh
Done running hooks in: /usr/local/bin/before-notebook.d
Executing the command: start-notebook.py
[I 2024-11-19 21:21:09.357 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2024-11-19 21:21:09.359 ServerApp] jupyter_server_mathjax | extension was successfully linked.
[I 2024-11-19 21:21:09.360 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2024-11-19 21:21:09.362 ServerApp] jupyterlab | extension was successfully linked.
[I 2024-11-19 21:21:09.362 ServerApp] jupyterlab_git | extension was successfully linked.
[I 2024-11-19 21:21:09.363 ServerApp] nbclassic | extension was successfully linked.
[I 2024-11-19 21:21:09.363 ServerApp] nbdtm | extension was successfully linked.
[I 2024-11-19 21:21:09.365 ServerApp] notebook | extension was successfully linked.
[I 2024-11-19 21:21:09.365 ServerApp] Writing Jupyter server cookie secret to /home/jovyan/.local/share/jupyter/runtime/jupyter_cookie_secret
[I 2024-11-19 21:21:09.498 ServerApp] notebook_shim | extension was successfully linked.
[I 2024-11-19 21:21:09.498 ServerApp] notebook_shim | extension was successfully loaded.
[I 2024-11-19 21:21:09.514 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2024-11-19 21:21:09.514 ServerApp] jupyter_server_mathjax | extension was successfully loaded.
[I 2024-11-19 21:21:09.514 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2024-11-19 21:21:09.516 LabApp] JupyterLab extension loaded from /opt/conda/lib/python3.12/site-packages/jupyterlab
[I 2024-11-19 21:21:09.516 LabApp] JupyterLab application directory is /opt/conda/share/jupyter/lab
[I 2024-11-19 21:21:09.516 LabApp] Extension Manager is 'pypi'.
[I 2024-11-19 21:21:09.522 ServerApp] jupyterlab | extension was successfully loaded.
[I 2024-11-19 21:21:09.523 ServerApp] jupyterlab_git | extension was successfully loaded.
[I 2024-11-19 21:21:09.525 ServerApp] nbclassic | extension was successfully loaded.
[I 2024-11-19 21:21:09.551 ServerApp] nbdtm | extension was successfully loaded.
[I 2024-11-19 21:21:09.553 ServerApp] notebook | extension was successfully loaded.
[I 2024-11-19 21:21:09.553 ServerApp] Serving notebooks from local directory: /home/jovyan
[I 2024-11-19 21:21:09.553 ServerApp] Jupyter Server 2.14.2 is running at:
[I 2024-11-19 21:21:09.553 ServerApp] http://c9f8329666bd:8888/lab?token=4ef2cd097455d85421e2f62093c79a976a580dac470dd7d7
[I 2024-11-19 21:21:09.553 ServerApp] http://127.0.0.1:8888/lab?token=4ef2cd097455d85421e2f62093c79a976a580dac470dd7d7
[I 2024-11-19 21:21:09.553 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 2024-11-19 21:21:09.554 ServerApp]
```

To access the server, open this file in a browser:

file:///home/jovyan/.local/share/jupyter/runtime/inserver-8-open.html

or copy and paste one of these URLs:

http://c9f8329666bd:8888/lab?token=4ef2cd097455d85421e2f62093c79a976a580dac470dd7d7

http://127.0.0.1:8888/lab?token=4ef2cd097455d85421e2f62093c79a976a580dac470dd7d7

```
[I 2024-11-19 21:21:44.774 LabApp] 302 GET /lab/ (@172.17.0.1) 2.79ms
hon-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server, unified-language-server, vscode-css-languageserver-bin, vscode-html-languageserver-bin, vscode-json-languageserver-bin, yaml-language-server
```