Using Hadoop/Spark with Python for Big Data Analytics

Tristan Glatard

Canada Research Chair (Tier II) on
Big Data Infrastructures for Neuroinformatics

18-20, 25 August, 2020

Learning objectives

Concepts of distributed storage and processing for Big Data

Practical experience with Apache Spark APIs

Applications of Apache Spark

Workshop design

20% theory, 80% practice

Mini-projects and use cases (1 per session)

Session 1: distributed data pre-processing in Spark

Session 2: unsupervised clustering in Spark

Session 3: supervised classification in Spark

Session 4: data stream analysis in Spark

Getting started

Material available as Jupyter notebooks

`session1.ipynb`, `session2.ipynb`, `session3.ipynb`, `session4.ipynb`

Start notebook with

```
jupyter notebook session1.ipynb
```

Full solutions in `session{1,2,3,4}_solution.ipynb`

Software dependencies

Python >= 3.6

and

pip install -r requirements.txt

./check_install.py

(not needed on VM or JupyterHub)