Data Lake:
Modern Data Systems
Master EDT-IDS – Orsay University

Bachar Wehbi

bachwehbi@gmail.com

What we will learn in this Module

- Understand Modern Data Systems & Data Architectures for Big Data
- Data Lake: Store massively and efficiently
- Process: Revisiting Spark
- Ingest in real-time: Streaming data systems
- Expose: NoSQL Data Systems
- Data Lake in the Cloud

Module Organization

- Session 1 – Data Systems & Architectures
- Session 2 – Data Storage + Exercises
- Session 3 – Data Processing with Spark - Exercises
- Session 4 – Data streaming with Kafka + Exercises
- Session 5 – Exercises - Continued
- Session 6 – Exposing Data with NoSQL + Exercises
- Session 7 – Data Lakes in the Cloud + Exercises

Module Organization

- Score will be composed as follows
  - 20%: Presence and participation
  - 80%: Practical work

Module Practical Work

- All exercises will be on Linux
- Use VirtualBox if you use MS Windows
  - Ubuntu 18.04 LTS
  - Please have it ready as before next session
- All exercises include the installation procedures for dependencies
- Module contents will be available at:

https://github.com/bachwehbi/data-systems

Module Organization

- Interactions and discussions
  - Stop me when things are not clear
  - Ask questions: there is no bad questions, only bad answers!
    - I might not have answers for everything, but I'll always try to come back with an answer the next session.
  - Provide feedback on the content of the module. This helps make it better
    - Open issues at the module repository on Github
    - If you have suggestions or corrections, open Pull Requests
    - Or just send me an email

Module Organization

- I need your emails to share content
- Please send an email now to bachwehbi@gmail.com
  - Subject: Data Systems
  - Your full name