Modern Data Systems

Master EDT-IDS – Orsay University

Bachar Wehbi me@bachwehbi.net

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

Module Organization

- Session 1 – Apr 30: Data Systems & Architectures
- Session 2 – May 2: Data Storage + Exercises
- Session 3 – May 7: Data Processing with Spark - Exercises
- Session 4 – May 9: Data streaming with Kafka + Exercises
- Session 5 – May 14: Exposing Data with NoSQL + Exercises
- Session 6 – May 16: Practical work 1 (rated)
- Session 7 – Jun 18: Practical work 2 (rated)
- Session 8 – Jun 19: Project presentation (Rated)

Module Organization

- Score will be composed as follows
  - 20%: Presence and participation
  - 40%: Practical work
  - 40%: Project

Module Practical Work

- All exercises will be on Linux
- Use VirtualBox if you use MS Windows
  - Ubuntu 16.04 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 come back with an answer the net 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

Module Organization

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

Module Organization: Projects

- As part of this module, you are requested to make a study on a selected subject
- Objective: Deepen your understanding of a key Data subject, practice writing documents and sharing ideas
- Deliverables:
  - Short report of 10 pages maximum
  - Presentation and discussion on June 19 (20 min presentation + 10 min discussion/QA)

Module organization: Projects

Project subjects

- GDPR: The new European General Data Privacy Rules
  - What is it about, how it impacts companies and citizens, what needs to be done on be GDPR compliant.
- Apache Arrow:
  - Arrow project presentation, main features and how it will simplify data projects. Examples (code) are welcome.
- Data Lakes in the Cloud:
  - Presentation of the different services especially from AWS and Azure. What are the main features and the use cases.
- Hadoop v3.x (3.0 and 3.1):
  - Presentation of the major new features and how it will impact (Big) data projects in the future.
- Graph Databases (NoSQL):
  - What is a graph database, presentation of the major tools, and discussion about the use cases.
- Your own topic
  - Please talk to me first