DATA SCIENCETECH INSTITUTE

***MSc in Applied Data Science & Big Data***
***MapReduce Ecosystem***

# The Hadoop Ecosystem

## *(Applied Data Science & Big Data – DSBD3-001)*

## Volume of classes hours: 50 hrs *(± same personal work expected)*

## Course summary:

Apache Hadoop has evolved to be the main platform on top of which most Big Data components are developed. By the end of the course, students will be familiar with this large and growing ecosystem. We firstly cover the 3 bricks of Apache Hadoop Framework: the filesystem (HDFS), the processing pattern and its API (MR), and the resource manager YARN.
We then cover the complementary tasks of the data analyst and learn how to leverage some of the most popular software that are used on top of the Hadoop platform.
We will also experiment functional programming through Spark.
Through these lessons, we will cover the basics of DevOps paradigm and expose students to multiple usual hardware configurations and their impacts on Big Data systems.

**Topics to be covered are:**

1. The Hadoop ecosystem
2. File System, scheduling and resource management
3. Data analytics
4. NoSQL databases
5. Indexing engines
6. Workflow Management & ETL
7. Dataflow Management
8. Scalable Enterprise Serial Bus
9. Realtime
10. Multi-tenancy: security, resource allocation, data governance
11. Agile methodology
12. Machine learning
13. Data exploration & visualization

**Course objectives:**

- Distributed Systems: pros and cons, associated paradigms
- Massively Parallel Processing
- Design a Big Data solution on top of Hadoop framework
- Functional Programming to be able to put multiple models and tools in competition and select the most mathematically valid one.

## Course mini-projects description:

Using a case proposal, a commonly used case will be isolated to illustrate the role of each component and how they interact together. It will cover the following steps:

- Data acquisition
- Buffering
- Transformation
- Storage
- Analysis

## - Visualization

## Theoretical background used:

No book is required. A course material will be provided (presentation slides), and students will learn how-to retrieve up-to-date information from internet (wikis, articles, blogs, source code).

## Technologies used:

This course is best suited for students familiar with programming and database systems. Basic knowledge in Java and SQL is required. Being comfortable with Linux and networking is a plus.
Prior knowledge of Distributed Systems, Apache Hadoop, NoSQL, or functional programming is not necessary.