



UNIVERSITÀ  
DEGLI STUDI DI MILANO-BICOCCA

## SYLLABUS DEL CORSO

### Technological Infrastructures for Data Science

2223-2-F9101Q052

---

#### Aims

The main purpose of the course is to provide the student with a solid, foundational understanding of the main **technology solutions** and **software development methodologies** in support of data science. Hands-on sessions will provide the student with the basic skills needed to interact with such tools.

#### Contents

The course comprises the following modules:

**Module 1 - Infrastructure:** Introduction to Virtualization, Cloud Computing and Containerization.

**Module 2 - Platform:** Data organization and distribution, Data Lake, HDFS, YARN

**Module 3 - Processing:** Batch vs. Streaming vs. Messaging, the cases of Hadoop, Spark, Storm, Kafka

**Module 4 - Software Development:** Waterfall, Agile, DevOps, DataOps, MLOps

#### Detailed program

Course topics divided by modules:

**Module 1 - Infrastructure:**

- The figure of the data engineer

- The reference architecture
- Virtualization
- Cloud Computing (Introduction, Service and deployment models, essential features)
- Containerization with Docker
- Serverless

#### **Module 2 - Platform:**

- The Data Lake
- HDFS and YARN

#### **Module 3 - Processing:**

- Batch processing (Apache Hadoop and Apache Spark).
- Stream processing (Apache Storm, Apache Spark, and Apache Flink)
- Messaging (Apache Kafka)

#### **Module 4 - Software Development:**

- Service computing
- Software engineering
- Development methodologies (Waterfall, Agile, DevOps, DataOps, MLOps)

### **Prerequisites**

No prerequisites

### **Teaching form**

Classroom lectures, classroom exercises

The course will be delivered in English

### **Textbook and teaching resource**

Lecture notes and slide decks.

The following textbooks are referenced for further study:

- The basics of cloud computing ISBN-13: 978-0124059320 Authors: Derrick Rountree, Ileana Castrillo
- Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems ISBN-13: 978-1449373320 Author: Martin Kleppmann

## Semester

Second year, first semester

## Assessment method

The assessment consists of two parts: a **written test** and an **oral discussion** on an in-depth topic.

The written test consists of open and closed questions on the course topics. This has a duration of about one hour and a maximum score of 17 points.

The oral test consists of a discussion of a topic not covered in the course or an in-depth study of a topic covered in the course. The deepening work (research work and slide creation) can be done in groups of up to 3 people but the discussion and evaluation are personal.

The topic of the deepening must be agreed in advance with the lecturer. The oral test entitles the student to a maximum of 15 points.

Once the student has taken both tests, the exam will be considered passed if both of these conditions are met:

- for both parts the student has scored more than (or equal to) 7 points
- the sum of the points for the two parts is greater than or equal to 18

In that case a grade consisting of the sum of the points can be registered.

No partial examination will be issued during the course.

## Office hours

Tuesday 12:30-14:30 ask for email confirmation

## Sustainable Development Goals

INDUSTRY, INNOVATION AND INFRASTRUCTURE

---