UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Technological Infrastructures for Data Science

2324-2-FDS01Q016

Aims

The course aims at providing a solid understanding of the technological platforms (Cloud and Containers) data collection and management, as well as of the computing platforms (architectures, algorithms, and infrastructures) that can be used to analyze those data.

The exercises will provide the student with the basic capabilities necessary to interact with such platforms.

Contents

Virtualized platforms for collecting and handling data characterized by volume and velocity. Data processing architectures, processing infrastructure, Big Data platforms for Data Science, examples of platforms.
Software development and ML models and practices: Waterfall, DevOps, DataOps, MLOps.

Detailed program

1. Course introduction

- Importance of the Data Engineer role in the professional environment
- Data Pipelines
- NIST Big Data Reference Architecture

2. Cloud Computing:

- Virtualization as the major enabler of Cloud Computing

- Cloud Computing introduction, including Service Models, Deployment Models, and Characteristics
- Deep dive into actual cloud offerings, comparing AMAZON EC3 and Azure web app
- Serverless computing

3. Containerization:

- Introduction to Linux containers and Docker
- Running a container
- Volumes: persistent data for containers
- Networking for containers
- Multi-container applications
- Multi-host resource management for containerized workloads using Kubernetes

4. Data Organization and Distribution:

- Data lake concept
- Big Data Data Warehouse (DW)
- HDFS (Hadoop Distributed File System), Avro, Parquet (data storage formats)

5. Big Data Processing Platforms:

- Resource Management using Apache YARN
- Batch processing basics and Apache Spark
- Stream processing basics, Apache Storm, and Spark Streaming

6. Software Development Process:

- Services and Service Computing
- Introduction to Software Engineering
- Waterfall and Agile development methodologies
- DevOps and DataOps concepts

Prerequisites

Basic knowledge of:

- a programming language (e.g., Python)
- the architecture of a computer (CPU, Memory, Disk...).

Teaching form

Classroom lectures, classroom exercises. The course will be held in English

Textbook and teaching resource

Lecture notes and slides provided by the lecturers.

Semester

Second year, first semester

Assessment method

The exam will consist of two parts. The two parts will have to be **done in the same session**.

The first part will consist of a set of closed and open questions to be taken in one hour (approximately there will be 9 closed and 4 open questions, however there may be minor variations in the structure of the exam). The first part of the exam will be computer based (esami online) and the weight of each question for grade formation will be explicitly stated containing the questions. The exam will be delivered in English, the student has the option to answer in English or Italian.

The second part will consist of an in-depth study of a topic agreed upon with the professor to be carried out in groups of 2 and its oral exposition.

Examples of project types:

- Analysis and testing of a particular technology platform (advantages, disadvantages, costs, learning curve, building an application using that technology)
- Design of a cloud application (choice of provider, type of virtual machines, services, cost estimation, quality of service estimation, supported data size estimation, corrective activities needed in case of unforeseen events)

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

1. For both parts, the student will have scored more than half of the 15 points available to them
2. The sum of the points of the two parts is greater than or equal to 18.

In this case, the student will be able to record a grade consisting of the sum of the points.

Office hours

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

Sustainable Development Goals

INDUSTRY, INNOVATION AND INFRASTRUCTURE