

2EL1590 – Cloud computing and distributed computing

Instructors: Francesca Bugiotti, Gianluca Quercini **Department:** DÉPARTEMENT INFORMATIQUE

Language of instruction: FRANCAIS
Campus: CAMPUS DE PARIS - SACLAY

Workload (HEE): 60

On-site hours (HPE): 35,00

Elective Category: Fundamental Sciences

Advanced level: Yes

Description

Nowadays, the marketing strategies of most companies is based on the analysis of massive and heterogeneous data that needs a considerable amount of computational power. Instead of purchasing new hardware and software infrastructures, companies often resort to the computational and storage power offered by *cloud computing* platforms over the Internet.

The objective of this course is to present the fundamental principles of distributed systems and distributed computing that are at the heart of cloud computing.

The course will cover the principles of virtualization and containerization and the methods and tools used for distributed processing (for instance, *MapReduce*, *HDFS*, and *Spark*).

The course will also introduce advanced techniques and algorithms for the analysis of massive and heterogeneous data (PageRank, supervised learning, and *clustering*) and a brief introduction to some optimized Spark-compliant data formats (i.e., Parquet).

Quarter number

SG8

Prerequisites (in terms of CS courses)

Python programming, databases, basics of networking will be appreciated.

Syllabus

Introduction

- Cloud computing: motivation and terminology.
- Introduction to the public cloud providers (Amazon AWS, Microsoft Azure).
- Setup of a virtual machine on Microsoft Azure.

Virtualisation



- Virtualisation basics.
- Containerisation basics.
- Docker architecture.
- Images, containers, volumes and networks in Docker.
- Application deployment with Docker.

Multi-service applications and orchestration.

- Microservices architecture.
- Orchestration principles.
- Presentation of Kubernetes.
- Application deployment with Kubernetes.
- Application deployment in the cloud.

Cloud programming and software environments.

- Parallel computing, programming paradigms.
- Hadoop MapReduce.
- Apache Spark.
- Apache Parquet.

Data analysis.

- Cloud environments and data storage.
- Data distribution.
- Dataframes.

Class components (lecture, labs, etc.)

Introduction.

o Lecture: 3h

Virtualisation and containerisation.

Lecture: 3hTutorial: 3h

Multi-service applications.

Lecture : 3hTutorial : 3h



• Lab assignment (graded): 3h

Cloud programming and software environments.

Lecture : 9hTutorial : 3h

• Lab assignment (graded): 3h

• Exam: 2h

18h lecture, 9h tutorials, 6h lab assignments, 2h exam.

Grading

Written examination at the end of the course (MCQ + exercises) on the Evalmee platform (paperless exam).

• 2 lab assignments are graded.

Course support, bibliography

- Hwang, Kai, Jack Dongarra, and Geoffrey C. Fox. *Distributed and cloud computing: from parallel processing to the internet of things*. Morgan Kaufmann, 2013.
- Erl, T., Puttini, R., & Mahmood, Z. (2013). *Cloud computing:* concepts, technology & architecture. Pearson Education.
- Tel, G. (2000). *Introduction to distributed algorithms*. Cambridge university press.
- Miner, D., & Shook, A. (2012). MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems.
 O'Reilly Media, Inc.
- Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning spark: lightning-fast big data analysis*. O'Reilly Media, Inc.
- Schenker, Gabriel. Learn Docker Fundamentals of Docker 18.x. Packt Publishing, Print.

Resources

Teaching staff: Francesca Bugiotti, Gianluca Quercini, Idir Ait Sadoune,

Marc-Antoine Weisser, Arpad Rimmel Maximum lab enrollment: 25 students

Software, number of licenses required: Use of free software



Learning outcomes covered on the course

At the end of this course, the students must be able to:

- Understand the fundamental concepts of cloud computing.
- Master the notion of virtualization and containerisation in the cloud.
- Be acquainted with the different cloud platforms.
- Use the distributed computing paradigms, such as MapReduce and Spark.
- Design distributed algorithms on data.

Description of the skills acquired at the end of the course

Operate all types of data, structured or unstructured, including big data.

• Conceive, design, implement and authenticate complex software.