2EL1590 - Cloud computing et informatique distribuée

Responsables : **Gianluca QUERCINI , Francesca BUGIOTTI**

Département de rattachement : **DÉPARTEMENT INFORMATIQUE**

Langues d'enseignement : **FRANCAIS**

Type de cours : **Electif 2A**

Campus où le cours est proposé : **CAMPUS DE PARIS - SACLAY**

Nombre d'heures d'études élèves (HEE) : **60**

Nombre d'heures présentielles d'enseignement (HPE) : **30**

Année académique : **2024-2025**

Catégorie d'électif : **Sciences fondamentales**

Niveau avancé : **oui**

Présentation, objectifs généraux du cours :

De nos jours, la stratégie marketing des entreprises repose de plus en plus sur l'analyse de données massives et hétérogènes qui nécessite d'une grande puissance de calcul.
Au lieu d'investir sur l'acquisition de matériel et de logiciel, les entreprises souvent font recours à la puissance de calcul et de stockage mise à disposition par des plateformes de cloud computing via internet.

L'objectif du cours est de présenter les concepts fondamentaux des systèmes distribués et du calcul distribué qui sont à la base du cloud computing.
Le cours abordera les principes de la virtualisation et de la conteneurisation, ainsi que les méthodes et les outils pour effectuer des calculs distribués (par exemple, MapReduce, HDFS, Spark).
Le cours introduira aussi des techniques et des algorithmes avancés pour l'analyse de données massives et hétérogènes (PageRank, apprentissage supervisé, clustering) et une introduction à un ensemble techniques de stockage optimisées Spark-compliant (Parquet).

Période(s) du cours (n° de séquence ou hors séquence) :

SG8

Prérequis :

Programmation en Python, bases de données, des notions en réseaux seront aussi appréciées.

Plan détaillé du cours (contenu) :

Introduction

- Cloud computing : motivation et terminologie.
- Présentation des cloud publiques (Amazon AWS, Microsoft Azure)

- Démarrage d'une machine virtuelle sur le cloud Microsoft Azure

Virtualisation

- Principes de base de la virtualisation.
- Principes de la conteneurisation.
- Architecture de Docker.
- Images, conteneurs, volumes et réseaux en Docker.
- Déploiement d'applications avec Docker.

Applications multi-services et orchestration

- Architecture microservices.
- Principes de l'orchestration.
- Présentation de Kubernetes.
- Déploiement d'applications avec Kubernetes.
- Déploiement d'applications dans le cloud.

Programmation cloud et environnements logiciel.

- Calcul parallèle, paradigmes de programmation.
- Hadoop MapReduce.
- Apache Spark.
- Apache Parquet.

Analyse de données.

- Environnements Cloud et stockage de données.
- Données distribués.
- Dataframes.

Déroulement, organisation du cours :

Introduction.

- **Cours magistral** : 3h

- **Virtualisation et conteneurisation.**
  - **Cours magistral :** 3h
  - **TD** : 3h
- **Applications Multi-service**.
  - **Cours magistral** : 3h
  - **TD** : 3h
  - **TP (noté)** : 1,5h
- **Programmation cloud et environnements logiciel**.
  - **Cours magistral** : 7,5h
  - **TD** : 3h
  - **TP (noté)** : 1,5h
- **Exam**: 2h

16,5h cours magistral, 9h TD, 3h TP, 2h exam.

Des ponts réguliers de suivi seront programmés pendant le cours, notamment pour la finalisation des TP notés. L'équipe enseignante sera aussi disponible à des créneaux horaires fixés pour suivre le travail individuel requis pour le cours et répondre aux questions des élèves.

Organisation de l'évaluation :

Examen écrit à la fin du cours (QCM + exercises) sur la plateforme Evalmee (examen dématerialisé).

- 2 TP notés.
- Chaque TP compte pour 30% de la note finale et l'examen écrit pour 40%

Moyens :

Equipe enseignante : Francesca Bugiotti, Gianluca Quercini, Idir Ait Sadoune, Marc-Antoine Weisser, Arpad Rimmel
Taille des TP : 25 élèves
Outils logiciels et nombre de licence nécessaire : Utilisation de logiciels licence libre

Acquis d'apprentissage visés dans le cours :

A l’issue de ce cours, l’élève sera capable de :
- Comprendre les concepts à la base du cloud computing.
- Maîtriser la notion de virtualisation et conteneurisation dans le cloud.
- Connaître les différentes plateformes cloud.
- Utiliser les paradigmes de calcul distribué, tels que MapReduce et Spark.
- Concevoir des algorithmes de calcul distribué sur les données.

Description des compétences acquises à l'issue du cours :

C.2 Develop in-depth skills in an engineering field and a family of professions

- Assimilate new knowledge into operational and efficient tools or methods for the given problem

C.6 Be operational, responsible, and innovative in the digital world

- Process data

Bibliographie :

- 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 19.x. Packt Publishing,. Print.
- Lisdorf, Anders. Cloud Computing Basics: A Non-technical Introduction. Apress, 2021.
- Linthicum, David. An Insider's Guide to Cloud Computing. Addison-Wesley, 2023