Gestion de données massives en utilisant HADOOP et Spark

ECTS : 2

Enseignant responsable : Khalid BELHAJJAME (https://dauphine.psl.eu/recherche/cvtheque/belhajjame-khalid)

Langue du cours : Français

Volume horaire : 18

Description du contenu de l'enseignement :

L'objectif de ce cours est de vous former aux principaux modèles, architectures et outils utilisés dans l'ingénierie de Data Science pour gérer de grand volumes de données.
La richesse et l'importance de l'information véhiculée par les données a conduit à une augmentation rapide de l'influence des données sur les individus et la société. Les données de toutes sortes, tels que les énormes collections de données sur l'internet, sont devenues omniprésentes dans pratiquement tous les aspects de notre société. Les données numériques sont à présent des ingrédients clés des innovations dans divers domaines, tels que l'énergie, l'économie, la santé ou le climat, ainsi que dans la science, du web, les entreprises doivent relever le défi de capturer, stocker, rechercher, partager et visualiser les données pour faciliter la prise de décision à tous les niveaux (politique, économique, scientifique, social, etc.). L'objectif de ce cours est de vous former aux principaux modèles, architectures et outils utilisés dans l'ingénierie de Data Science pour gérer de grand volumes de données.
L'objectif de ce cours est de vous former aux principaux modèles, architectures et outils utilisés dans l'ingénierie de Data Science pour gérer de grand volumes de données. En particulier :
- Les clusters
- Le modèle de programmation MapReduce
- L'écosystème Hadoop
- SPARK
- Les base de données NoSQL,

Pré-requis : Connaissances des bases de données relationnelles.