

Systèmes, Langages et Paradigmes pour le Big Data

ECTS : 3

Volume horaire : 24

Description du contenu de l'enseignement :

Le cours s'articule en trois parties.

Dans la première, l'accent est sur le paradigme MapReduce et le système Hadoop, avec un focus sur son système de fichiers HDFS. Le cours illustrera les mécanismes de base de Hadoop pour le support de l'exécution parallèle de 'dataflow' MapReduce sur des clusters de machines. Une attention particulière sera donnée aux aspects algorithmiques et d'optimisation de dataflow MapReduce.

La deuxième partie présentera des langages de requête et d'analyse de données caractérisés par des mécanismes de haut niveau et qui sont compilés sur MapReduce. Le focus sera sur les langages Pig Latin et Hive, des langages incluant des mécanismes à la SQL. Les techniques de compilation vers MapReduce seront présentées.

La troisième partie sera consacrée à des évolutions de Hadoop, et en particulier au système Spark et au langage de support Scala. Le focus sera sur l'architecture de Spark, la notion de RDD, l'évaluation lazy de transformations et actions sur des collections distribuées RDD.

Les notions apprises seront mises en pratique dans un projet où les étudiants devront concevoir un dataflow pour l'analyse de grands volumes de données. L'implémentation sera faite tant en MapReduce qu'en Spark, et une analyse expérimentale sera effectuée pour comparer les performances des deux implémentations.

Compétence à acquérir :

Apprentissage des aspects fondamentaux des technologies Big Data pour la gestion et analyse de données massives.