Systèmes, paradigmes et langages pour les Big Data

ECTS : 3

Volume horaire : 24

Description du contenu de l'enseignement :

The main aim of this course is to give students a deep and solid understanding of the state of the art of Big Data systems and programming paradigms, and to enable them to devise and implement efficient algorithms for analysing massive data sets.

The focus will be on paradigms based on distribution and shared-nothing parallelism, which are crucial to enable the implementation of algorithms that can be run on clusters of computers, scale as the size of input data increases, and can be safely executed even in the presence of system failures.

Lectures will give articular emphasis to the MapReduce paradigm and the internal aspects of its related runtime support Hadoop, as well as to MapReduce-based systems, including Spark and Hive, that provide users with powerful programming tools and efficient execution support for performing operations related to complex data flows. The attention will be then given to mechanisms and algorithms for both iterative and interactive data processing in Spark. A particular attention will be given to SQL-like data querying, graph analysis, and the development of machine learning algorithms.

A large part of the course consists of lab-sessions where students develop parallel algorithms for data querying and analysis, including algorithms for relational database operators, matrix operations, graph analysis, and clustering. Lab-sessions rely on the use of both desktop computers and Hadoop clusters on the Amazon WS cloud.

Program:

1. Introduction to massive data management and processing.

2. A data operating system for distributed data management, Hadoop.

3. MapReduce paradigm, algorithm design, implementation and optimisation.

4. iterative and interactive massive data processing, algorithm design, implementation and optimisation in Spark

5. large scale data-warehouse in Hive

References:

Mining of Massive Datasets.
Jure Leskovec, Anand Rajaraman, Jeff Ullman
http://www.mmds.org/#top

Data-Intensive Text Processing with MapReduce.
Jimmy Lin and Chris Dyer.
Mogan & Claypool Publishers

Hadoop: The Definitive Guide - Tom White.
O'Reilly.

Apache Hadoop Yarn - Arun C.Murty, Vinod Kumar Vavilapalli, et al.
Addison Wesley

Programming Hive.
Edward Capriolo, Dean Wampler, Jason Rutherglen.
O'Reilly.

Big Data Analytics with Spark.

Université Paris Dauphine - PSL - Place du Maréchal de Lattre de Tassigny - 75775 PARIS Cedex 16 **- 01/07/2025**