Big Data

En bref

Langue de cours: Anglais

Présentation

Prérequis

Compréhension de base du fonctionnement des systèmes informatiques : processeur, mémoire, opérations sur disque et fonctions du système d'exploitation.

Bonne connaissance des systèmes de gestion de bases de données relationnelles.

Objectifs d'apprentissage

Ce cours vise à présenter les principales technologies permettant de relever les nombreux défis posés par le "Big Data".

Le "Big Data" est un terme utilisé pour décrire une collection de données dont le volume est énorme et qui croît de manière exponentielle au fil du temps. En bref, ces données sont si volumineuses et complexes qu'aucun des outils traditionnels de gestion des données n'est capable de les stocker ou de les traiter efficacement.

Dans une première partie, ce cours présente les technologies existantes qui permettent de traiter efficacement de grands volumes de données, à savoir Hadoop MapReduce et Apache Spark.

Dans une seconde partie, nous étudierons les solutions qui permettent de stocker et d'interroger ces volumes de données ; nous nous concentrerons sur une variété de bases de données NoSQL (en utilisant MongoDB comme étude de cas).

Description du programme

- Introduction et programmation MapReduce.

Notions de base et motivations du Big Data.

Vue d'ensemble de Hadoop.

Introduction à MapReduce.

- Hadoop et son écosystème : HDFS.

Description approfondie du système de fichiers distribués Hadoop (HDFS).

- Introduction à Apache Spark.

Apache Spark, son architecture et ses fonctionnalités.

Les ensembles de données distribuées résilientes : transformations et actions.

- Spark Structured APIs et Structured Streaming

SparkSQL, Spark streaming.

- Bases de données distribuées et NoSQL.

Distribution des données (réplication, fragmentation, théorème CAP).

Aperçu des bases de données NoSQL.

- Bases de données orientées documents : MongoDB.

Présentation de MongoDB.

Modalité de contrôle des connaissances

Evaluation sur machine

Bibliographie

- Singh, Chanchal, and Manish Kumar. Mastering Hadoop 3: Big data processing at scale to unlock unique business insights. Packt Publishing Ltd, 2019.
- Mehrotra, Shrey, and Akash Grade. Apache Spark Quick Start Guide: Quickly learn the art of writing efficient big data applications with Apache Spark. Packt Publishing Ltd, 2019.
- Karau, Holden, et al. Learning spark: lightning-fast big data analysis.O'Reilly Media, Inc., 2015
- Giamas, Alex. Mastering MongoDB 4.x: Expert techniques to run high-volume and fault-tolerant database solutions using MongoDB 4.x. Packt Publishing Ltd, 2019.
- Bradshaw, Shannon, Eoin Brazil, and Kristina Chodorow. MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. O'Reilly Media, 2019.
- Scifo, Estelle, Hands-on Graph Analytics with Neo4j. Packt Publishing Ltd, 2020

Equipe pédagogique

- Gianluca QUERCINI
- Stéphane VIAL

Total des heures		**21h**
CM	Cours Magistral	9h
TD	Travaux Dirigés	12h

Infos pratiques

Nom responsable UE

Responsable pédagogique

Stéphane Vialle

svialle@intervenants.centrale-marseille.fr

Responsable pédagogique

Gianluca Quercini

gquercini@intervenants.centrale-marseille.fr