Big Data

In brief

Course langage: English

Presentation

Prerequisites

Basic understanding of how computer systems work: processor, memory, disk operations and operating system functions.

Good knowledge of relational database management systems.

Learning objectives

This course aims to present the main technologies for tackling the many challenges posed by Big Data.

Big Data is a term used to describe a collection of data that is enormous in volume and growing exponentially over time. In short, this data is so voluminous and complex that none of the traditional data management tools are capable of storing or processing it efficiently.

In the first part, this course presents the existing technologies that enable large volumes of data to be processed efficiently, namely Hadoop MapReduce and Apache Spark.

In the second part, we will look at solutions for storing and querying these volumes of data; we will focus on a variety of NoSQL databases (using MongoDB as a case study).

Description of the programme

Introduction and MapReduce programming.

Basic concepts and reasons for Big Data.

Overview of Hadoop.

Introduction to MapReduce.

Hadoop and its ecosystem: HDFS.

In-depth description of the Hadoop Distributed File System (HDFS).

Introduction to Apache Spark.

Apache Spark, its architecture and features.

Resilient distributed datasets: transformations and actions.

Spark Structured APIs and Structured Streaming

SparkSQL, Spark streaming.

Distributed databases and NoSQL.

Data distribution (replication, sharding, CAP theorem).

Overview of NoSQL databases.

Document-oriented databases: MongoDB.

Presentation of MongoDB.

How knowledge is tested

Evaluation on machine

Bibliography

- 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

Teaching team

- Gianluca QUERCINI
- Stéphane VIAL

Total des heures **0h**

Useful info

Name responsible for EU

Lead Instructor

Stéphane Vialle

svialle@intervenants.centrale-marseille.fr

Lead Instructor

Gianluca Quercini

gquercini@intervenants.centrale-marseille.fr