Cours Big Data, Master Industrie Numérique, Inseet

Acquis d'apprentissage (objectifs) :

- S'orienter parmi les technologies étiquetées « big data » les plus courantes
- Identifier les goulots d'étranglements dans l'exécution d'un traitement de données et adapter le traitement pour y remédier
- Choisir et mettre en œuvre une architecture adaptée à un traitement donné, et en particulier choix CPU vs. GPU, local vs. cloud, batch vs. flux, haut niveau vs. bas niveau etc.
- Produire des analyses statistiques simples avec Spark
- Provisionner une infrastructure simple sur AWS

Principales notions abordées :

Le terme de *big data* est de plus en plus utilisé aussi bien en entreprise que dans les médias généralistes. Malheureusement, souvent, il est utilisé en tant que terme fourre-tout. Ce cours débute avec une déconstruction de la notion de *big* data en présentant les V du *big data* et l'introduction de la notion de traitement de données à haute performance.

Il présente ensuite un panorama des technologie étiquetées *big data* et les architectures informatiques associés en les mettant en parallèle à des solutions classiques :

- Architecture générale du calcul en local (processeur, mémoire vive, stockage) et en distribué (centralisé vs. pee-to-peer ; avantages et inconvénients des systèmes répartis)
- Architectures de stockages (systèmes de fichier vs. base de données, local vs. Distribué)
- Zoom sur le stockage distribué avec HDFS
- Zoom sur le calcul distribué avec Spark et MapReduce
- Présentation du *cloud computing* avec manipulation de *Amazon Web Service*

Organisation du cours

- Jour 1 - matin : **Introduction aux Big Data**
    - Introduction aux enjeux du Big Data : historique, 3V du Big Data
    - Outils non big data pour traiter de grosses volumétrie de données
    - Problèmes Big Data courants et leurs solutions
    - Les enjeux sociétaux du big data
- Jour 1 - après midi : **Introduction au Cloud Computing**
    - Introduction au Cloud Computing : historique, concepts, services, avantages, inconvénients
    - TP1 : Cloud Computing
        - Création d'un bucket S3 et d'une instance EC2 sur AWS.

            - Déploiement d'un webservice basique hébergé dans un Auto Scaling Group dernière un Load Balancer
- Jour 2 - matin : Les outils pour stocker des données Big Data
    - Comment stocker des données dans un contexte de big data, File System vs Database, CAP theorem, avantages et inconvénients de la distributions
    - Présentation de HDFS et de Hadoop MapReduce
    - TP2 : Découverte de HDFS
        - Création d'un cluster EMR sur AWS
        - Upload de données sur HDFS
        - Manipulation des données via Hadoop MapReduce
- Jour 2 - après midi : **Spark, un outil open source de traitement pour le Big Data**
    - Présentation de Spark
    - TP3 : Découverte de Spark
        - Manipulation de tweets via Spark
- Jour 3 - matin : **Spark 2**
    - TP4 : Spark Machine Learning et Spark Streaming
        - Utilisation de Spark pour le Machine Learning (cas d'une régression linéaire)
        - Utilisation de Spark pour traiter des données en quasi temps réel
- Jour 3 - après midi : Utilisation de la suite AWS pour l'IoT
    - Atelier : Création d'un pipeline d'analyse de données IoT
        - Utilisation des services AWS : IoT Core, Kinesis Firehose, S3, Glue, Athena et Quicksight

Références bibliographiques :

- Carpenter, Jeff, and Eben Hewitt. 2020. *Cassandra: The Definitive Guide; Distributed Data at Web Scale*. 3rd ed. O'Reilly Media, Inc, USA.
- Chambers, Bill, and Matei Zaharia. 2018. *Spark: The Definitive Guide: Big Data Processing Made Simple*. O'Reilly Media, Inc, USA.
- Steen, Maarten van, and Andrew S. Tanenbaum. 2017. *Distributed Systems*. 3.01 ed. CreateSpace Independent Publishing Platform.
- White, Tom. 2015. *Hadoop - The Definitive Guide 4e-*. 4th ed. O'Reilly.
- Stephenson, D. (2018). Big Data Demystified : How to use big data, data science and AI to make better business decisions and gain competitive advantage (1re éd.). FT Press.
- Erl, T., Puttini, R., & Mahmood, Z. (2013). Cloud Computing : Concepts, Technology & Architecture (The Pearson Service Technology Series from Thomas Erl) (1re éd.). Pearson.