

CentraleSupélec  

**Big Data : Informatique pour les données et calculs massifs**

**1 - Définitions et objectifs du cours**

**Stéphane Vialle**

 université  Sciences et Technologies de l'Information et de la Communication (STIC)  

Stephane.Vialle@centralesupelec.fr  
<http://www.metz.supelec.fr/~vialle>

---

---

---

---

---

---

---

---


---

---

CentraleSupélec

**1 – Définition du « Big Data »**

Big Data  $\in$  Data Science  
 ou bien  
 Big Data  $\ni$  Data Science



« Data Science » : plutôt Math & Stats  
 « Big Data » : plutôt Informatique

---

---

---

---

---

---

---

---

---

---

CentraleSupélec

**1 – Définition du « Big Data »**

**Big Data :**

- « **Math-Info** » : Stats & Probas, Machine Learning, Fouille de données, Analyse de Graphes...
- **Informatique distribuée** (et parallèle) :
  - Paradigme de programmation **Map-Reduce**  
 « amener les codes de calcul sur les nœuds de données »  
 « traitements large échelle » ou même « web-scale »  
 Sur cluster Hadoop, sur matériel standard
  - Paradigmes du **Calcul à Haute Performance (HPC)**  
 Pour accélérer les algorithmes de « data analytics » ou de « machine learning »  
 Sur cluster de calcul intensif, sur GPU, sur SuperCalculateurs

---

---

---

---

---

---

---

---

---

---

**1 – Définition du « Big Data »**

**Big Data :**

- **BdD « NoSQL » : Not Only SQL**
  - BdD plus simples mais à très large échelle
  - Plusieurs types de BdD NoSQL

---

---

---

---

---

---

---

---

---

---

**1 – Définition du « Big Data »**

**Exemples d'application Big Data :**

- Analyse de fichiers de traces, de transactions, de données d'entreprise : prédiction de comportements, détection de nouveaux marchés...
- Analyse de signaux de capteurs : maintenance prédictive, prévention des risques...
- Analyse de graphes : fouille de réseaux sociaux, recherche de relations

---

---

---

---

---

---

---

---

---

---

**2 – Définition du « HPC »**

**HPC :**

- **Domaine applicatif** : types de simulations à réaliser, types de modèles à mettre au point, grands défis
- **Mathématiques numériques** : transformation du modèle en algorithme numérique
- **Informatique parallèle et optimisation du code** : pour tirer le maximum des architectures informatiques
- **Contrôle et visualisation de plus en plus interactives...**

---

---

---

---

---

---

---

---

---

---

**2 – Définition du « HPC »**

**Spécificité du HPC : l'Intrication ... à la place de la vision en couches !**

**Exemples d'applications HPC :**

- Prévision météo
- Simulations de neutronique
- Problèmes inverses : reconstruction de volumes en temps réel
- ...

The diagram illustrates the interdisciplinary nature of HPC. At the center is an oval labeled 'HPC'. Four boxes are connected to it: 'Mathématiques numériques' (containing 'Modèles numériques' and 'Etude de convergence et de précision'), 'Informatique parallèle et optimisée' (containing 'Optimisations sérielles', 'Vectorisation, parallélisation et distribution', and 'Analyse de performances'), 'Visualisation et contrôle interactifs des simulations', and 'Domaine applicatif' (containing 'Modélisation' and 'Validation des simulations'). A separate box 'Domaine connexe d'architecture des processeurs et des ordinateurs' is also connected to the central HPC oval.

---

---

---

---

---

---

---

---

**3 – Le HPC dans le Big Data**

**Grosses différences d'approches**

- Traitements amenés sur les nœuds de Bdd NoSQL
- Données en fichiers plats sur les machines de calcul
- Paradigme Map-Reduce
- Paradigmes multithreads, envois de messages, vectorisation...
- Objectifs fonctionnels à très large échelle
- Objectif/obsession de la performance

**Mais progressivement ...**

- Besoin de performances en Big Data (réimplantation d'outils)
- Concept de High Performance Data Analytics (HPDA)

---

---

---

---

---

---

---

---

**3 – Le HPC dans le Big Data**

**Architectures (hard et soft) hybrides : schéma générique**

The flowchart shows a data pipeline. On the left, 'Data srcs' and 'Data lake' feed into a blue box 'Large scale data management architecture' (with subtext 'Data extraction, cleaning, and filtering'). This feeds into a green box 'Intensive Computing Architecture' (with subtext 'Machine Learning, or Iterative Data Analytics'). Above the flow, a horizontal arrow indicates the process: 'Data srcs' → 'Data Ingeneering on commodity hardware' → 'Machine Learning on HPC hardware' → 'Data Visualisation'. On the right, there are icons for a data visualization and a globe.

---

---

---

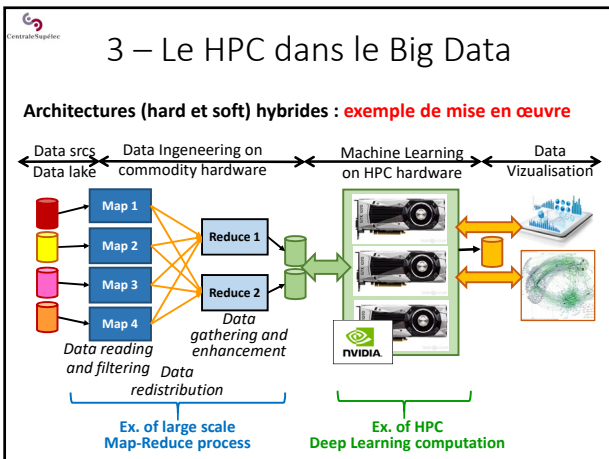
---

---

---

---

---




---

---

---

---

---

---

---

---

---

---

- ### 4 – Objectifs du cours
- I - Notions de bases d'informatique distribuée :**
- Rappels sur les composants *Hardware & Software*, problématique du déploiement distribué
  - Métriques d'évaluation de performance et de passage à l'échelle
- II - Distribution et parallélisation de traitements de données :**
- Schémas de parallélisation : barrières de synchronisation, modèle SPMD, Map-Reduce
  - Technologie d'Hadoop : localité calculs-données, vue d'ensemble et architecture(s) d'Hadoop
  - Algorithmique Map-Reduce : étapes de l'approche algorithmique, présentations de patrons de conception Map-Reduce
  - *Technologie de Spark*
  - *Solution de stockage S3 d'AWS*

---

---

---

---

---

---

---

---

---

---

- ### 4 – Objectifs du cours
- I - Notions de bases d'informatique distribuée**
- II - Distribution et parallélisation de traitements de données**
- III - Base de données NoSQL**
- Contexte d'émergence du NoSQL, classification des solutions
  - Caractéristiques communes, et technologies de HBASE et MongoDB
  - Environnements NoSQL de plus haut niveau
  - *Architecture spécifique de Neo4j pour l'analyse de graphes*
  - Aspects pratiques d'utilisation de MongoDB
- IV - Introduction au Machine Learning**
- Exploration et préparation des données
  - Objectifs et principes du Machine Learning
  - Algorithmes de Machine Learning
  - Exemple de parallélisation d'un algorithme de clustering
- + objectifs de performance pour l'analyse de données**

---

---

---

---

---

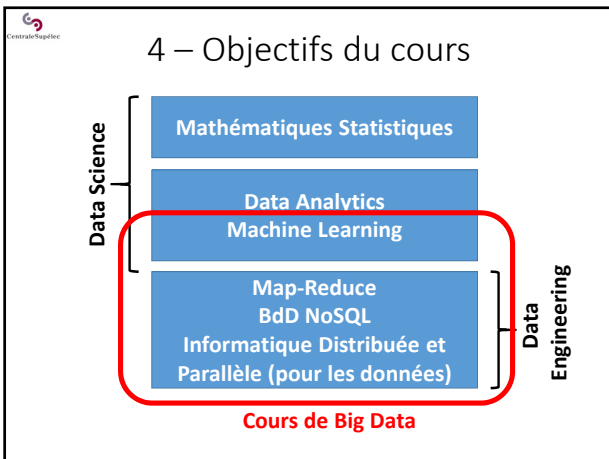
---

---

---

---

---



---

---

---

---

---

---

---

---



---

---

---

---

---

---

---

---