Introduction au cours de Big Data

M. Ben & R. Tavenard

Informations générales

- Installation des paquets Python nécessaires (une fois dans le dossier cloné) : télécharger le fichier `requirements.txt` sur CURSUS puis :
  `pip install -r requirements.txt`
- Contact des enseignants
  - Mathieu Ben, Romain Tavenard
  - prenom.nom@univ-rennes2.fr
  - U208-U212
- Évaluations (dates susceptibles de modifications)
  - 22 octobre + 16 décembre : 1/3 de la note pour chaque CC
  - TP noté 6 novembre : 1/3 de la note

Contenu du cours

- Comment se passe un calcul sur un ordinateur ?
- Comment accélérer un calcul ?
- Comment faire si les données ne tiennent pas en mémoire ?
- Présentation de plusieurs *frameworks*
  - Dask
  - Hadoop
  - Spark

Codage de l'information

Codage de l'information

- Stockage des données
  - En mémoire (RAM ou disque) : binaire (représentation en base 2)
  - Types de base (int, float, str, …)
    - Correspondance binaire $\Leftrightarrow$ valeur (1 par type)
  - Cas des flottants
    - Plusieurs types co-existent en Python (`float`, `np.float32`, `np.float64`, …)
    - Types `numpy` : 32 bits $\rightarrow$ 4 octets par flottant, 64 bits $\rightarrow$ 8 octets par flottant

Stockage des données

Stockage des données

Deux principaux types de stockage

	Sur disque (fichiers)	En mémoire (variables d'un programme)
Vitesse d'accès (lecture, écriture)	Lent	Rapide
Permanence (en cas d'extinction de l'ordinateur)	Oui	Non
Capacité	Grande (~1To)	Limitée (~10Go)

Stockage des données

- Principe de fonctionnement
    1. Chargement des données en mémoire
    2. Calculs à partir des données (processeur)
    3. Libération de la mémoire
        - Quand le programme termine
        - Quand le programme n'a plus besoin des données (libération explicite avec `del` en Python, ou action du ramasse-miette)
- En cas de dépassement de la mémoire
    - Utilisation du swap
    - Crash de l'application

Threads, process et calculs multi-coeurs

Calcul parallèle / distribué

- Calcul parallèle
  - Plusieurs tâches effectuées en parallèle
  - Possibilité de partager de la mémoire
  - Typiquement sur 1 machine
- Calcul distribué
  - Plusieurs tâches effectuées en parallèle
  - Sur des machines distinctes
  - Données possiblement réparties sur plusieurs serveurs

Notions de threads et de process

- Process
  - ~ 1 programme
  - Les process ne partagent pas de mémoire
- Threads
  - ~ 1 ensemble de calculs
  - Exemple : navigateur web (threads de récupération de données, de mise en page du site, etc.)
  - Chaque thread est généré par un process
  - Threads d'un même process partagent la mémoire

Processeur

- Processeur
  - Composant électronique qui effectue les calculs dans un ordinateur
  - Peut avoir plusieurs coeurs
    - Métaphore du restaurant
  - CPU vs GPU
    - CPU : moins de coeurs, chaque coeur plus puissant
    - GPU (généralement) : moins de mémoire spécifique, idéal pour multiples petits calculs parallèles