A propos du Cours Big Data

Objectifs du cours Big Data

- Découvrir la méthodologie map/reduce.
- Apprendre à utiliser Hadoop.
- Apprendre à rédiger et exécuter des programmes pour Hadoop.
- Apprendre à utiliser MongoDB (BD Non-structuré).
- Apprendre à utiliser Apache Spark.

Pré-requis

Pour pouvoir tirer le maximum de ce cours, il faut connaître :

- **Systèmes Répartis**
- **Programmation orientée objets (Java)**
- **Bases des Données**
- **Algorithmes parallèles et distribués**

Mode d’évaluation

L’évaluation finale se fait à travers:

- Un **examen final**, qui compte pour **60%** de la note finale.
- **Évaluation continue** à raison de **40%** restant (Contrôle TP, Exposé, Interro).

Plan de Cours

1-introduction au Big Data

2-Systeme Hadoop

3-Apache Spark

4- BD NoSql (Mongo DB)

Plan de TDs

1- Métriques de performance

2- Partie 1: Algorithmique avec MapReduce
2- Partie 2: HDFS

3- NoSQL

Plan de TPs

1- Préparation de l'environnement pour les TPs

2- Analyse de Ventes en utilisant Hadoop

3- Problème du réseau social et des « amis en commun » en utilisant Apache Spark

4- MongoDB – Utilisation et API Java

Exposés sur les technologies de Big Data

1- Apache Cassandra
2- Apache Kafka
3- QlikView
4- Qlik Sense
5- Tableau
7- Apache Storm
8- Apache Hive
9-Apache Pig
10 – Presto
11- Apache Flink
12-Apache Sqoop
13- Rapidminer
14- KNIME
15- Elasticsearch

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire de Mila
Institut des Mathématiques et Informatique
Département de l'Informatique

Big Data

– Chapitre 1 –

Introduction au Big Data

s.meghzili@centre-univ-mila.dz

Plan du chapitre 1

- **Motivations**
- **Définitions du Big Data**
- **Données**
- **Défis du Big Data**
- **Les 5 Vs du Big Data**
- **Hiérarchie de la mémoire**

Motivations: Exemple de recherche sur Google

- Le **nbre** moyenne de requêtes de recherche dans **Google** = **63000** requête par Seconde
  - **Donc, Combien de requête par année** ?

  - 63000* 3600**h***24**j***365**année** = **1 986 768 000 000** requête par Seconde
  - Donc, **Analyser ces données** vous donnera un **vertige**

Définitions du **Big Data**

Le terme **Big Data** se réfère:

• A une **accumulation** de données très **larges** et très **complexes** pour être traitées par les outils **classiques** de gestion des bases de données.

•Aux solutions **hardware** et **software** développées pour la gestion de données **volumineuses**.

•A la branche de l'informatique qui étudie les solutions de gestion de données **volumineuses**.

Données: Source de données

Plusieurs sources de données:

- **Messages texte** sur les médias sociaux (Twiter, Facebook)
- **Images** numériques et **vidéos** publiées en ligne (Youtube)
- Enregistrements transactionnels ***d'achat en ligne (Amazon)***
- Signaux **GPS** de **téléphones mobiles**
- les informations **climatiques**
- …

Donc, ces données appelées **Big Data** ou **Données Massives**

Données : Nature des données (0)

Niveau de structuration	Modèle de données	Exemples	Facilité de traitement
Structuré	Système de données relationnel objet/colonne	Base de données d'entreprise...	Facile (indexé)
Semi-structuré	XML, JSON, CSV, logs	API Google, API Twitter, web, logs...	Facile (non indexé)
Non structuré	Texte, image, vidéo	web, e-mails, documents...	Complexe

Défis du Big Data

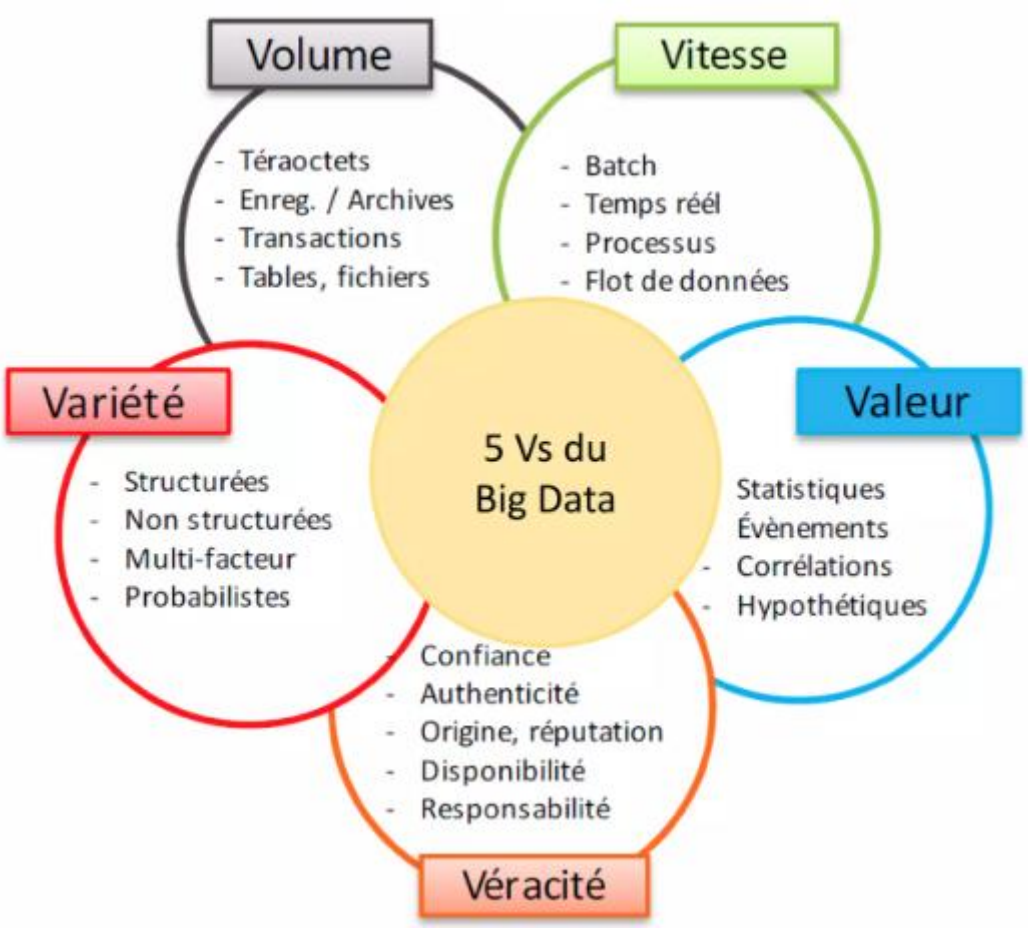
- ✓ **Réunir** un grand volume de données variées pour trouver de nouvelles **idées**

- ✓ **Difficulté** pour sauvegarder toutes ces données
  - Systèmes Distribués.
  - Bases de données (**relationnelles**/**NoSQL**) distribuées.

- ✓ **Difficulté** pour traiter ces données et les utiliser
  - **Parallélisation** du calcul sur les machines.
  - Bases de données **parallèles**.
  - Frameworks de traitement distribué (e.g., **Hadoop MapReduce**/**Spark**)

- ✓ Les données sont créées **rapidement**

Les 5 Vs du Big Data



Les 5 Vs du Big Data : Volume

- Le terme **volume** fait référence à la grande quantité d'éléments de données dans le big data.

- Exemples:
  - Les services de **streaming** comme **Netflix** ou **YouTube** accueillent des millions d'utilisateurs qui diffusent des vidéos en continu, ce qui génère une énorme quantité de données.

  - **Netflix** doit stocker ce volume colossal de données, les préférences des **utilisateurs**, leur **historique** de recherche et leurs **interactions**.

  - Le volume de données généré aide **Netflix** à utiliser des algorithmes sophistiqués pour **recommander** des émissions et des films

Les 5 Vs du Big Data : Varieté

- La nature **hétérogène** des données (structurées, semi-structurées, non structurées).
- Les données **structurées** comprennent des types de données bien définis, tels que des **bases de données** de noms et de chiffres.
- Les données **non structurées**, quant à elles, comprennent des types de données tels que **du texte**, des **sons**, des **images** et des **messages** sur les médias sociaux.
- Les données **semi-structurées** sont un mélange des deux.
  - **Par exemple**, dans le domaine des soins de santé, les données relatives aux patients peuvent inclure des **enregistrements structurés** tels que l'**âge**, le **diagnostic** et l'**historique** des traitements, ainsi que des données **non structurées** telles que des **notes médicales**, des **images** de santé et même des **informations génétiques**.

Les 5 Vs du Big Data : Vélocité (vitesse)

- Elle comprend la **vitesse** de **génération** des données, ainsi que la vitesse à laquelle les professionnels les **collectent** et les **traitent**.

  - **Par exemple**, Système **d'analyse des Sentiments** qui traite les **tweets** afin de trouver l'ambiance générale du candidat politique.
  - **Vélocité**: flot constant de données (**7,500 tweets/second**).

Les 5 Vs du Big Data : Véracité

- Le terme **véracité** fait référence à la **fiabilité** et à la **qualité** des données.

- Avec un tel **volume** de données générées quotidiennement, il est toujours difficile de s'assurer que les données avec lesquelles vous travaillez sont **impartiales** et représentent **correctement** ce qu'elles sont censées représenter.

Les 5 Vs du Big Data : Valeur

- Elle provient des informations et des **modèles** que vous pouvez trouver dans les données.

- Comme les **big data** intègrent des données provenant de sources et de formats **divers**, vous pouvez obtenir des informations sur les paramètres qui vous intéressent, tels que le **comportement des clients**, l'**évolution du marché**, les **performances de l'entreprise**, etc.

  Par exemple,

  - alors que les **données structurées** peuvent révéler des **tendances** et des **modèles numériques**,
  - les données textuelles **non structurées** provenant de sources telles que les messages sur les **médias sociaux** ou les **commentaires** des clients peuvent révéler les **sentiments** et les **opinions** qui déterminent le comportement humain.

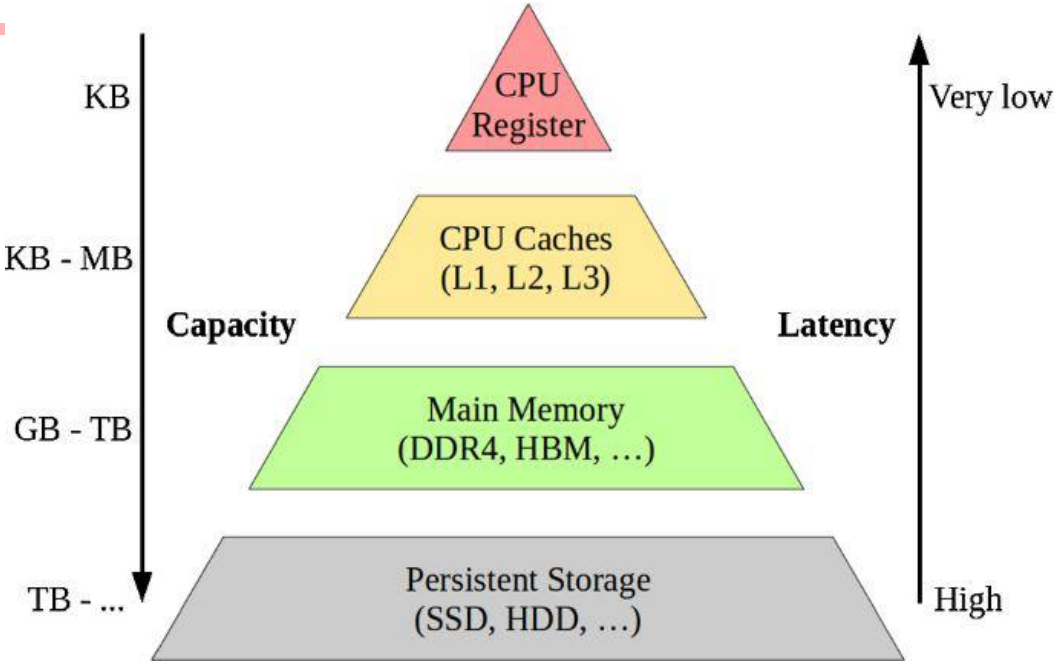
Hiérarchie de la mémoire (1)

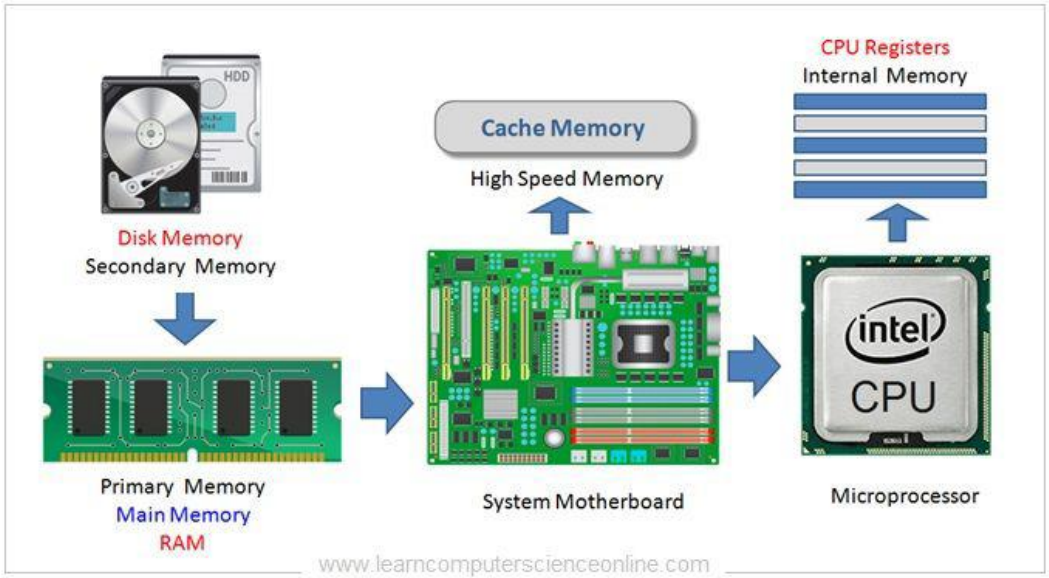
- Les systèmes réels ont plusieurs niveaux de types de **stockage**:
  - **Hiérarchie supérieure**: stockage **petit** et **rapide** à proximité du processeur
  - **Hiérarchie inférieure**: stockage **grand** et **lent** plus éloigné du Processeur

- Pour le programmeur / compilateur, pas besoin de savoir
  - Le matériel fournit une ***abstraction***: la mémoire ressemble à un seul grand tableau,

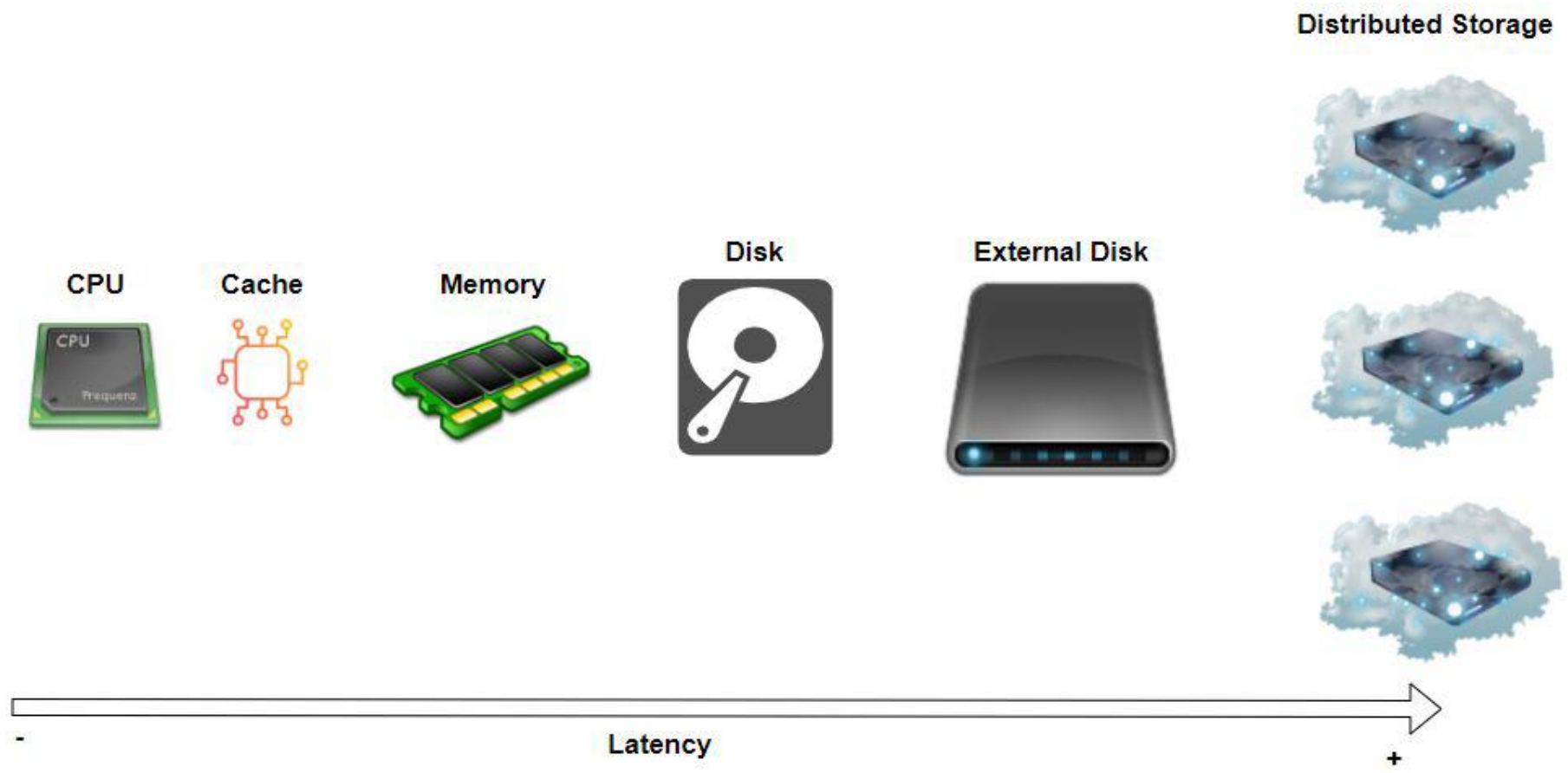
- Mais les performances **dépendent** du **modèle d'accès** du programme.

Hiérarchie de la mémoire (2)





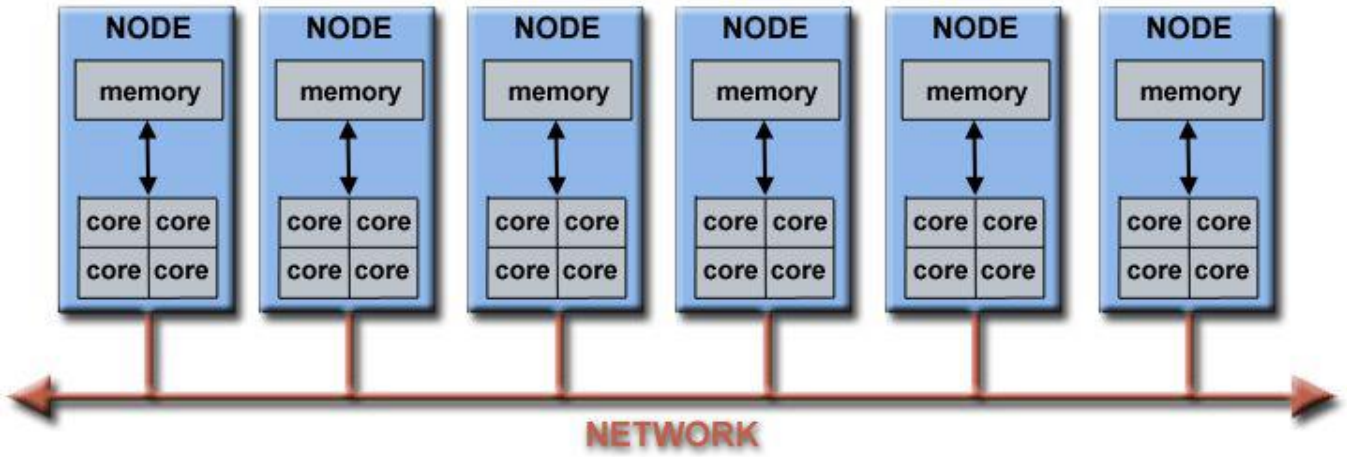
Hiérarchie de la mémoire (3)



Plus les ensembles de données sont **volumineux**, plus la latence est **grande**

Hiérarchie de la mémoire (4)

- Un **cluster** de traitement de données est simplement de nombreux ordinateurs reliés via une connexion **Ethernet**.
- Le stockage est **partagé**
- Localité: les données seront stockées sur l'ordinateur qui les **traitera**



Disciplines du Big Data

- **Informatique distribuée:**
  - Paradigme de programmation ***Map-Reduce****: «amener les codes de calcul sur les noeuds de données»* «***traitements large échelle*** *» sur cluster* ***Hadoop****, …*

- **•Informatique parralèle:**
  - 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...

- **•Bases de Données « NoSQL » : Not Only SQL**
  - BdD plus simples mais à très large échelle
  - Plusieurs types de BdD NoSQL: stockage distribué, interrogation sur les données.