Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique
Filière Génie Informatique et Digitalisation

Fondements du Big Data

Pr. N. EL FADDOULI
nfaddouli@gmail.com

2023-2024

Déroulement du Cours

- Ressources et activités pédagogiques sur **Moodle** ***(http://rime.emi.ac.ma)***
- Séance de 4h: cours et TP
- Auto-évalution après chaque chapitre (QCM)
- Evaluation formative: QCM + TP noté
- Evaluation sommative

Plan – Introduction générale

- Définitions
- Exemples d'utilisations
- Données au repos – Données en mouvement.
- Caractéristiques du Big Data
- L'écosystème Big Data - Hadoop

Traces numériques

- Les traces générées lors de l'utilisations de la technologie numérique sont composées de tout type de données saisies ou générées automatiquement.

- L'idée de base du Big Data est que les traces, de plus en plus volumineuses, de notre utilisation des outils numériques peuvent être **collectées**, **stockées** et **analysées** afin d'en tirer des connaissances pouvant aider à la **prise de décision**, la **prévision des risques** ou **l'améliorations des processus**, …

Définition

- Le cabinet d'études Gartner définit le Big Data comme suit:

"le Big Data est une forte volumétrie, haute Vélocité et grande Variété de données qui exigent des techniques innovantes et rentables de traitement d'information pour une meilleure compréhension, une prise de décision et l'automatisation et amélioration des processus."

- Ernst and Young propose la définition suivante:

"*Big Data refers to the **dynamic**, **large** and **disparate volumes** of data being created by people, tools and machines. It requires new, innovative, and scalable technology to **collect**, **host** and **analytically process** the vast amount of data gathered in order to derive **real-time business insights** that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value."*

Définition

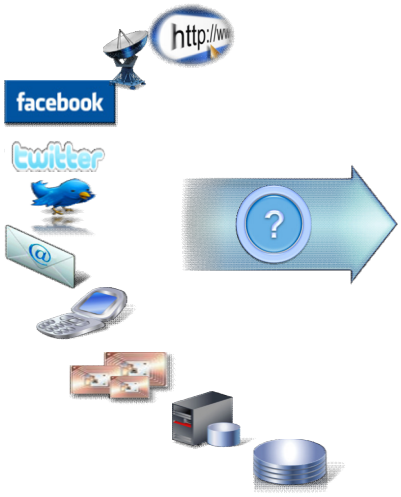
- **Lisa Arthur (Forbes contributor)** définit le **Big Data:**

"*a collection of data from traditional and digital sources inside and outside a company that represent a source of ongoing discovery and analysis."*

- **Quelques exemples d'utilisation du Big Data**

- Science
- Astronomy
- Atmospheric science
- Genomics
- Biogeochemical
- Biological
- Social
- Social networks
- Social data
  - *** Twitter**
  - *** Facebook**
  - *** LinkedIn**
- Commercial
- Web / event / database logs
- Sensor networks
- Internet text and documents
- **Medical** records
- Photographic archives
- Video / audio archives
- Government
- Military and homeland security surveillance

Secteurs d'utilisation des Big Data



Analyse du sentiment et de l'expérience client multicanal, …

Détection à temps des causes potentiellement mortelles dans les hôpitaux pour intervenir, …

Prédire les conditions météorologiques pour planifier l'utilisation optimale des éoliennes et optimiser les dépenses , …

Prendre des décisions de risque basées sur des données transactionnelles en temps réel, …

Identifier les criminels et les menaces provenant de sources vidéo, audio et de données disparates

Système d'unités / Système binaire d'unités

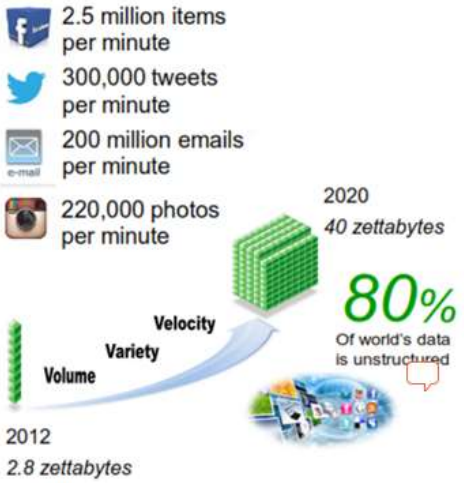
International System of Units (SI)			Binary Usage (deprecated)
kilobyte	KB	10^3	2^{10}
megabyte	MB	10^6	2^{20}
gigabyte	GB	10^9	2^{30}
terabyte	TB	10^{12}	2^{40}
petabyte	PB	10^{15}	2^{50}
exabyte	EB	10^{18}	2^{60}
zettabyte	ZB	10^{21}	2^{70}
yottabyte	YB	10^{24}	2^{80}

International Electrotechnical Commission (IEC) - 1999		
kibibyte	KiB	2^{10}
mebibyte	MiB	2^{20}
gibibyte	GiB	2^{30}
tebibyte	TiB	2^{40}
pebibyte	PiB	2^{50}
exbibyte	EiB	2^{60}
zebibyte	ZiB	2^{70}
yobibyte	YiB	2^{80}

Source: Wikipedia, http://en.wikipedia.org/wiki/Kibibyte

La croissance des données

- **Chaque jour, 2,5 trillions ($2.5*10^{18}$) d'octets de données sont générées.**
- **90% des données ont été créées au cours des deux dernières années**



Information is at the center of a new wave of opportunity

Exemples de volumétrie (2017)

- 3,3 millions de publications sur Facebook sont créées chaque minute.
- 65 972 photos Instagram sont téléversées chaque minute.
- 448 800 tweets sont crées chaque minute.
- 500 heures de vidéos YouTube sont téléversées chaque minute.
- On analyse moins de 1% des données créées .
- Il est impossible de suivre le rythme de croissance des données et il n'est pas nécessaire de tout stocker.
- En 2020, on estime à 37% les données utiles susceptibles d'être analysées

Exemples de volumétrie

- 2.5 petabytes

  Estimation maximale de la capacité mémoire du cerveau humain (1TB – 2.5PB)

- 13 petabytes

  Quantité pouvant être téléchargée d'Internet en deux minutes, si environ 300 millions de personnes se trouvaient connectées simultanément.

- 4.75 exabytes

  Séquences totales du génome de tous les habitants de la terre.

- 124 exabytes

  Quantité des données des data center dans le monde en 2018

- 40 zettabyte

  Quantité de données en 2020

Données au repos et Données en mouvement

- Informations rapides en temps réel, souvent transitoires sur les réseaux:
  - Informations provenant de capteurs
  - Informations provenant des journaux en temps réel et des moniteurs d'activité
  - Contenu en streaming comme l'audio et la vidéo
  - Transactions à grande vitesse telles que les tickers (*rapports d'évolutions de prix*), systèmes de trafic, …et

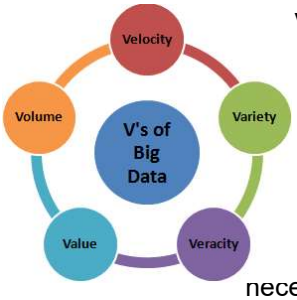
Information streams

- Informations stockées en dehors des systèmes conventionnels.
  Les données peuvent provenir du Web ou de différents systèmes internes.
  - Collection de données qui étaient en mouvement
  - Informations provenant des médias sociaux, journaux, courriels, …
  - Documents non structurés ou semi-structurées
  - Données structurées de divers systèmes

Information oceans

Caractéristiques du Big Data

- Il n'y a pas de définition unique du Big Data, mais certains éléments sont communs à toutes les définitions: la **vélocité**, le **volume**, la **variété**.
- Extraire, de manière rapide et rentable, des connaissances à partir de données volumineuses , variées et d'une vitesse élevée.



Volume: varie de terabytes à zettabytes. On doit traiter efficacement ce volume croissant.

Variété: Gérer divers types et structures de données

Vélocité: Analyser les flux de données en continu et les grands volumes de données persistantes.

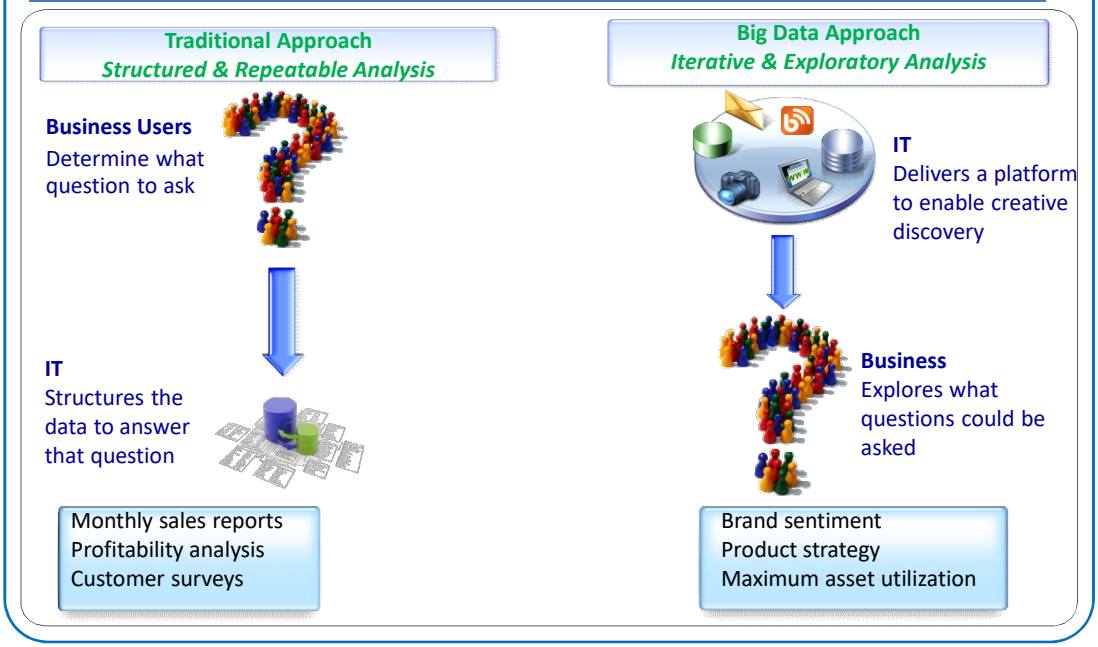
Véracité: authenticité et fiabilité des données necessitant une *grande rigueur* dans la collecte, le recoupement, le croisement et l'enrichissement des données pour lever l'incertitude pour garantir l'intégrité des données.

Valeur: le point le plus important, une solution Big Data doit apporter une valeur ajoutée pour l'entreprise en répondant à des objectifs commerciaux ou Marketing qui orientent l'utilisation des Big Data.

Apport du Big Data

- Aide intelligente pour la prise de décision,
- Réductions des coûts,
- Réductions de temps,
- Optimisation des processus de l'entreprise et développement de nouveaux produits ou services

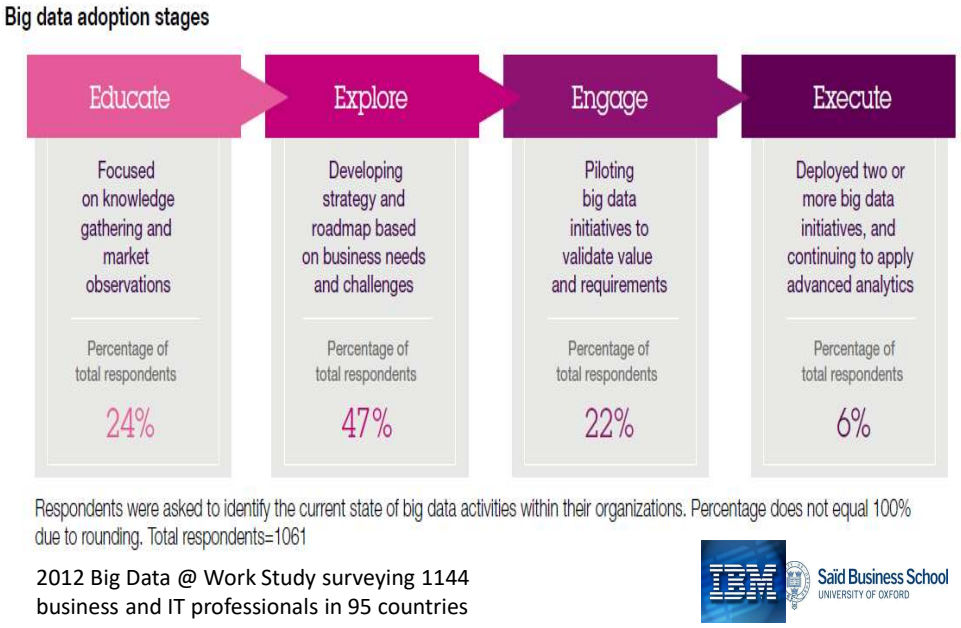
Big Data dans un système décisionnel



Projets Big Data

- **Déforestation** : Projet Planetary SKIN (Par la NASA, CISCO, …)
- **Suivi astronomique en direct** : Projet LSST (Large Synoptic Survey Telescope) (20 Tb chaque nuit)
- **Traitement du Cancer**: projet ICGC (International Cancer Genome Consortium), analyse de plusieurs BD sur des tumeurs de 5O types de cancers
- **Détection d'épidémies en temps réel** : Projet Healthmap puis ResistanceOpen pour la collecte, le stockage et l'analyse de divers types de données pour le suivi et la surveillance d'épidémie au niveau mondial.

Les étapes d'adoption du Big Data

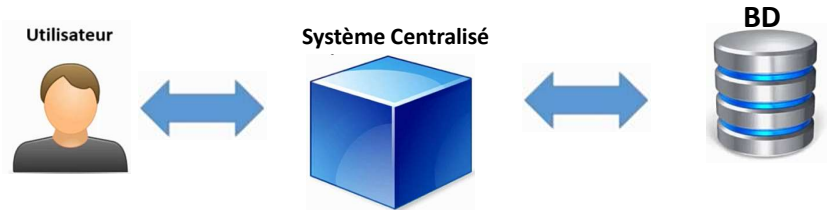


2012 Big Data @ Work Study surveying 1144 business and IT professionals in 95 countries

IBM Saïd Business School UNIVERSITY OF OXFORD

Approche traditionnelle Vs Approche Big Data

Approche traditionnelles

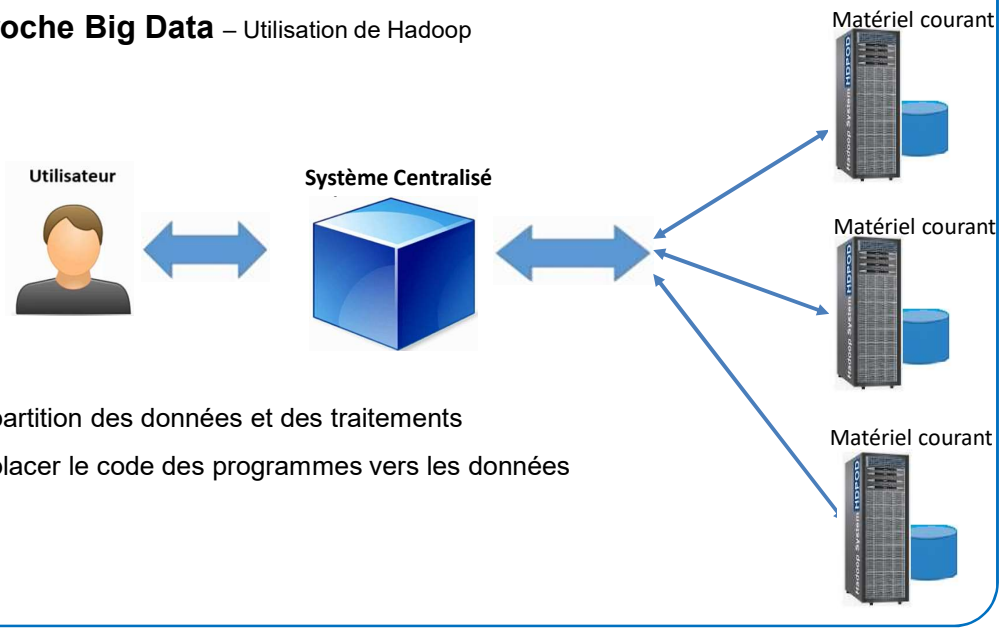


Insuffisance lorsqu'il y a:

- Grande croissance de données
- Des données sans schémas

Approche traditionnelle Vs Approche Big Data

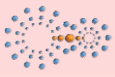
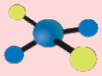
Approche Big Data – Utilisation de Hadoop



- Répartition des données et des traitements
- Déplacer le code des programmes vers les données

Fonctionnalités d'une plateforme Big Data

- Pour tirer profit du Big Data, la plate-forme utilisée doit permettre:

Understand and Navigate Federated Big Data Sources		Federated Discovery and Navigation
Manage and Store Huge Volume of any Data		Hadoop File System MapReduce
Structure and Control Data		Data Warehousing
Manage Streaming Data		Stream Computing
Analyze Unstructured Data		Text Analytics Engine
Integrate and Govern all Data Sources		Integration, Data Quality, Security, ILM (Information Lifecycle Management)