Introduction aux Big Data

Dr. H. EL BOUHISSI Epouse BRAHAMI

M2 IA

SOMMAIRE

Introduction

- Chaque jour, nous générons 2,5 trillions d'octets de données
- 90% des données dans le monde ont été créées au cours des deux dernières années.
- 90% des données générées sont non structurées.
- Source multiples: sites, bases de données, téléphones, serveurs:

- Détecter les sentiments et réactions des clients.
- Détecter les conditions critiques ou potentiellement mortelles dans les hôpitaux , et à temps pour intervenir.
- Prédire des modèles météorologiques pour planifier l'usage optimal des éoliennes.
- Prendre des décisions risquées basées sur des données transactionnelles en temps réel.
- Identifier les criminels et les menaces à partir de vidéos, sons et flux de données.
- Étudier les réactions des étudiants pendant un cour, prédire ceux qui vont réussir, d'après les statistiques et modèles réunis au long des années.

Challenges

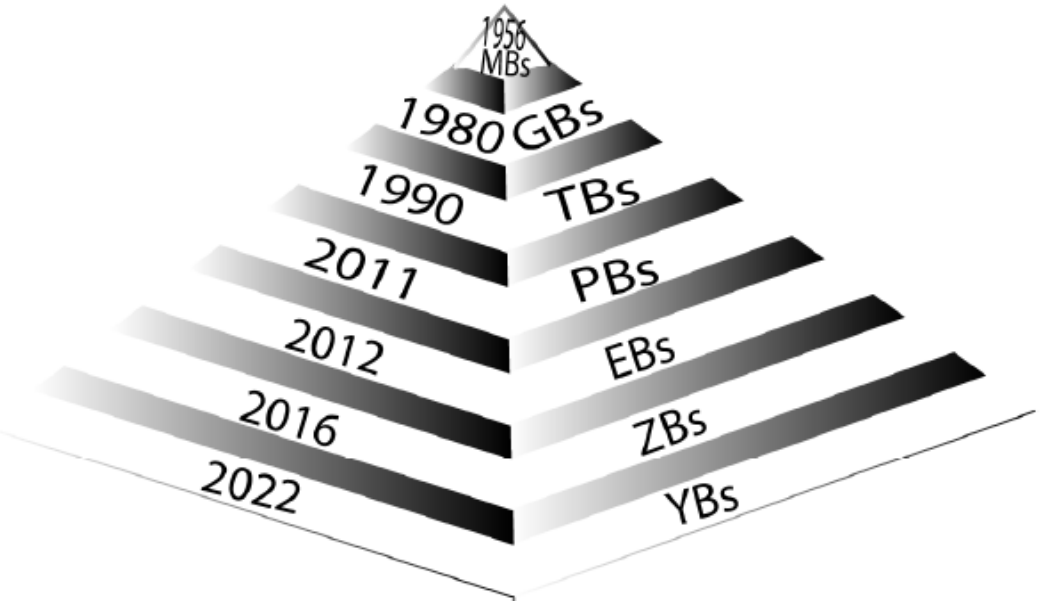
- Réunir un grand volume de données variées pour extraire de nouvelles connaissances.
- Capturer des données créées rapidement.
- Sauvegarder toutes ces données.
- Traiter ces données et les utiliser.
- Visualiser ces données.

Données appelées Big Data ou Données Massives

Historique

Le terme « ***Big data*** » a été inventé avant le développement de la technologie des bases de données, en raison de la nécessité de trouver des solutions pour faire face à l'afflux massif de données et l'insuffisance de l'espace de stockage (allant du mégaoctet (Mo) dans les années 1970 à YottaByte (YBs) en 2022).

A partir de l'an 2010, le volume de données augmente de façon exponentielle dicté par la variété et de nombreuses sources de données numérisées.



1 kilobyte	1,000
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

Big Data : Définition

Le terme Big Data se réfère aux technologies qui permettent aux entreprises d'analyser rapidement un volume de données très important et d'obtenir une vue générale.

Le Big data est un terme populaire utilisé pour caractériser le développement exponentiel, la disponibilité et l'utilisation de l'information, à la fois structurée et non structurée ", selon SAS [1].

Le big data est un domaine qui traite des méthodes d'analyse, d'extraction méthodique d'informations ou de traitement des volumes de données qui sont trop importants ou trop compliqués pour les logiciels d'application de traitement de données typiques. Les données comportant de nombreux champs (colonnes) ont une plus grande puissance statistique, mais les données comportant de nombreux attributs ou colonnes ont un taux de fausse découverte plus élevé [2].

[1] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[2] https://www.datakwery.com/techniques/big-data/

Big Data : Pourquoi

- Augmentation exponentielle de la quantité de données non structurées ○ Email, chat, blog, web, musique, photo, vidéo, etc.
- Augmentation de la capacité de stockage et d'analyse (L'utilisation de plusieurs machines en parallèle devient accessible).
- Les technologies existantes ne sont pas conçues pour ingérer ces données (Base de données relationnelles (tabulaires), mainframes, tableurs (Excel), etc.
- De "nouvelles" technologies et techniques d'analyse sont nécessaires.

D'où le "Big Data": pas strictement plus de data...

Big Data : Evolution

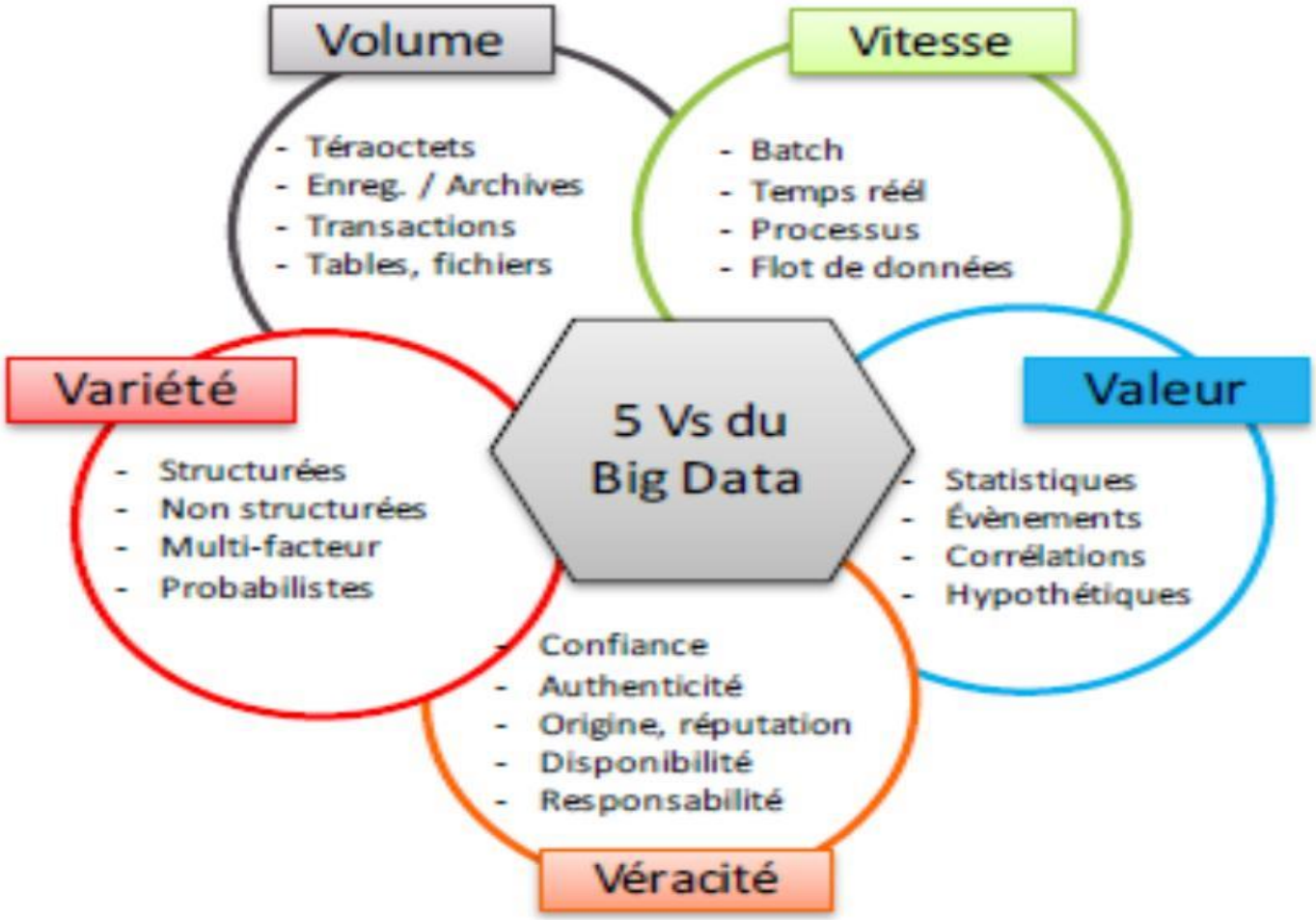
Le cadre d'évolution du Big Data passe par trois phases distinctes

Phase 01	Phase 02	Phase 03
Période 1970-2000	Période 2000-2010	Période 2010-présent
Contenu structure, basé sur DBMS (data base management system) : -RDBMS (Relational DBMS) & warehousing. -Charge de transfert d'extraits. -Traitement analytique en ligne. -Tableau de bord et analyse statistique.	Contenu non structuré, basé sur le web : -Recherche et extraction d'informations. -Exploitation des opinions. -Réponse aux questions. -Analyse et intelligence web. -Analyse des media sociaux. -Analyse des réseaux sociaux. -L'analyse spatio-temporelle.	Contenu mobile et basé sur des captures : -Analyse de localisation. -Analyse centré sur la personne. -Analyse contextuelle. -Visualisation mobile. -Interaction homme-machine.

Big Data : Les Vs

Extraction d'informations et décisions à partir de données, caractérisées par les 5 V:

1. Volume (Volume)
2. Variété (Variety)
3. Vitesse (Velocity)
4. Véracité (Veracity)
5. Valeur (Value)



Big Data : Volume

Capacité à traiter des Pétaoctets, des Exaoctets voire des Zettaoctets de données.

Questions :

. Quels sont les coûts et les outils de stockage et de traitement ?
. Comment gérer des données qui sont en croissance exponentielle ?

- ✓ 250 milliards de mails par jour
- ✓ 40 000 recherches sont analysée sur Google chaque seconde, soit plus de 3,5 milliards par jour ! (Source : Google Search, Statistics)
- ✓ 100 heures de vidéo sont en moyenne téléchargées sur YouTube chaque minute (Source : YouTube)
- ✓ 30 milliards d'objets connectés en 2010 (plus que d'humains)

Big Data : Variété

Capacité à traiter des données sous différents formats (texte, image, vidéos...), structurées ou non structurées.

A - Variété des sources

- Données internes de l'entreprise
- Données externes (OpenData, Météo, indicateurs économiques…)
- Données comportementales clients (géolocalisation, réseaux sociaux,…)

B - Variété des contenus

- Données structurées : informations que l'on trouve dans les bases de données
- Données semi-structurées : contenu composé d'éléments s'adressant à un humain à d'éléments s'adressant à une machine (emails, page web)
- Données non-structurées : contenu ne comportant pas de "balises" structurées lisibles par une machine (enregistrement audio, vidéo…)

Big Data : Velocité

Capacité à traiter des informations en temps réel.

- Rapidité d'arrivée des données
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des Téraoctets par jour, Sinon, risque de perte d'informations

Questions :

- Comment intégrer ces données en temps réel dans les schémas actuels conçus pour être alimentés en temps différé ?
- Comment canaliser ce déluge d'informations dans des flux maîtrisés ?
- Comment faire parvenir la bonne information au bon moment et au bon destinataire ?

Big Data : Véracité

Capacité à déterminer la fiabilité des données.

Cela fait référence au désordre ou la fiabilité des données. Avec l'augmentation de la quantité, la qualité et précision se perdent (abréviations, typos, déformations, source peu fiable…)

Questions :

- Comment s'assurer de la qualité et de la précision des données avec l'augmentation de la quantité ?
- Quelles techniques pour collecter, recouper, croiser et enrichir les données pour lever l'incertitude, créer la confiance, garantir la sécurité et l'intégrité des données ?

Big Data : Valeur

Capacité à se concentrer sur les données ayant une réelle valeur.

Le V le plus important

- Il faut transformer toutes les données en valeurs exploitables: les données sans valeur sont inutiles
- Atteindre des objectifs stratégiques de création de valeur pour les clients et pour l'entreprise dans tous les domaines d'activité

Questions :

- Comment déterminer dans le déluge d'informations (infobésité) ce qui est utilisable ?
- Comment transformer les données en valeurs exploitables ?

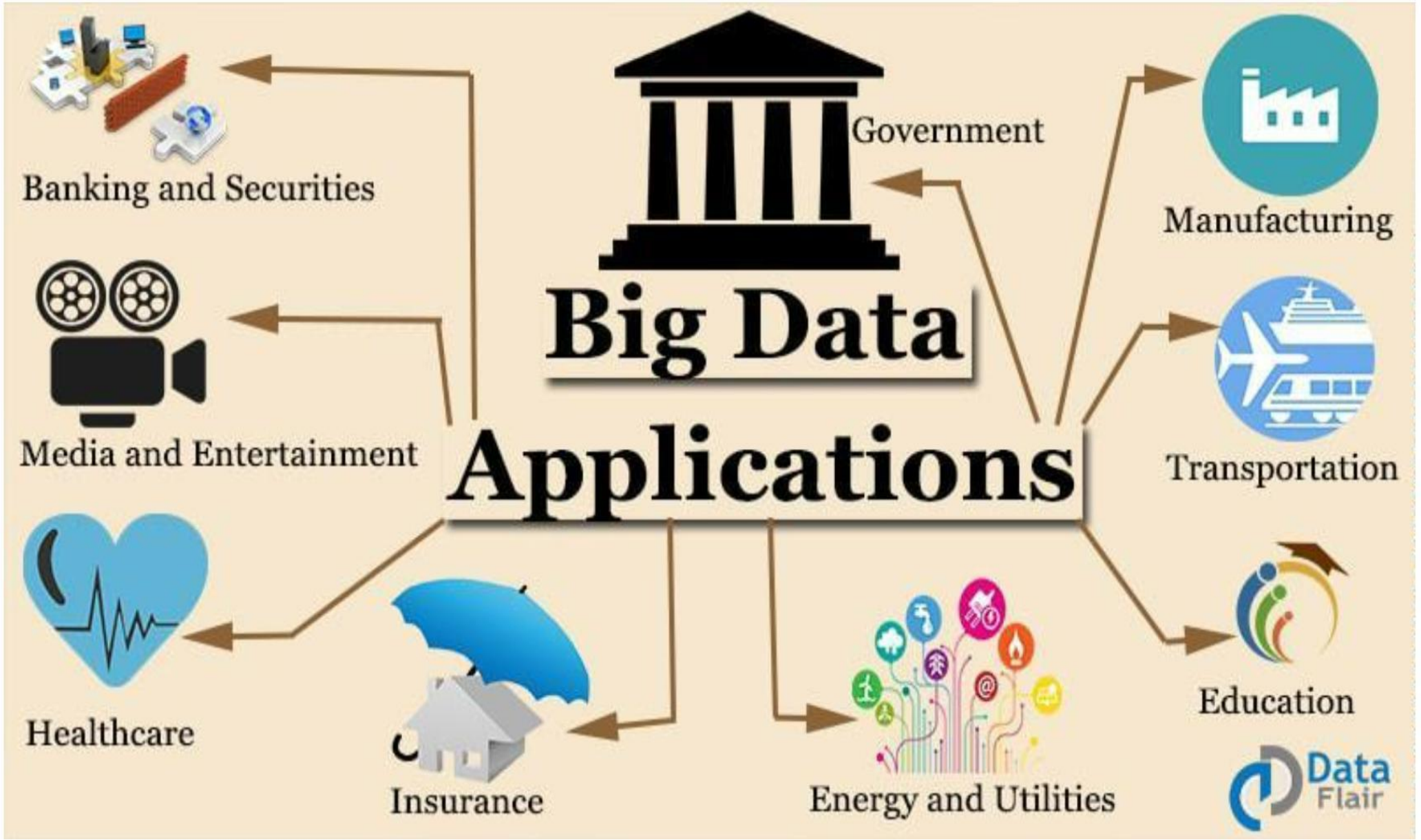
Big Data : Visualisation

Capacité à visualiser et rendre accessible les données collectées et traitées.

Question :

Comment obtenir une visualisation optimale et adaptée en un temps record ?

Exemples d'application



Technologies du Big Data

- Bases de données NoSQL (Cours 2)
- L’écosystème Hadoop (Cours 3)