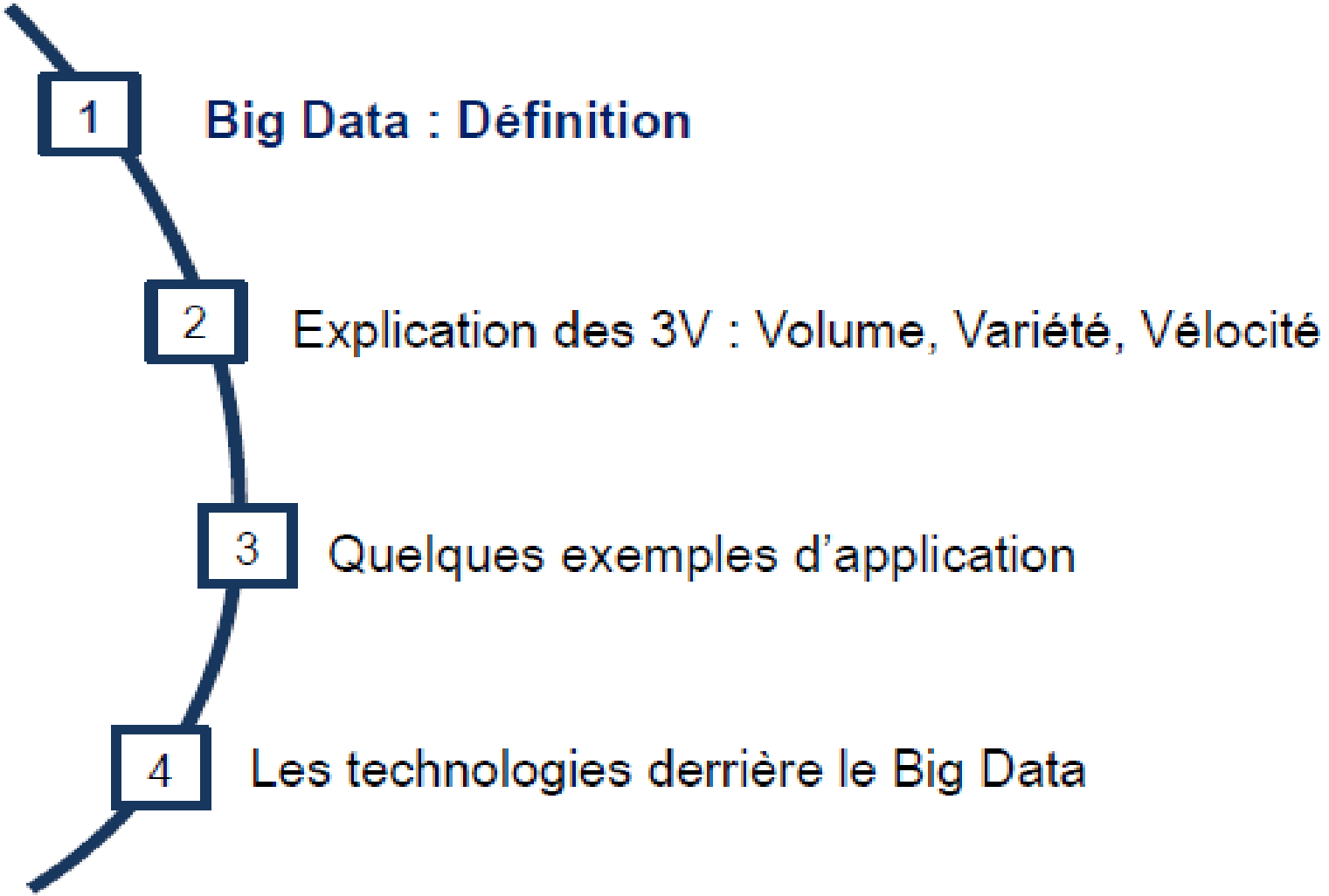
Introduction aux Big Data

Dr. Houda EL BOUHISSI

M1 GL

SOMMAIRE


1
Big Data : Définition
2
Explication des 3V : Volume, Variété, Vélocité
3
Quelques exemples d'application
4
Les technologies derrière le Big Data

Introduction au Big Data

- 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 (domaine Big Data in Education)

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

Intérêts

1/3	• Chefs d'entreprise prennent fréquemment des décisions basées sur des informations en lesquelles ils n'ont pas confiance, ou qu'ils n'ont pas
1/2	• Chefs d'entreprise disent qu'ils n'ont pas accès aux informations dont ils ont besoin pour faire leur travail
83 %	• Des DSI (Directeurs des SI) citent : « L'informatique décisionnelle et analytique » comme faisant partie de leurs plans pour améliorer leur compétitivité
60 %	• Des PDG ont besoin d'améliorer la capture et la compréhension des informations pour prendre des décisions plus rapidement

Big Data : Définition

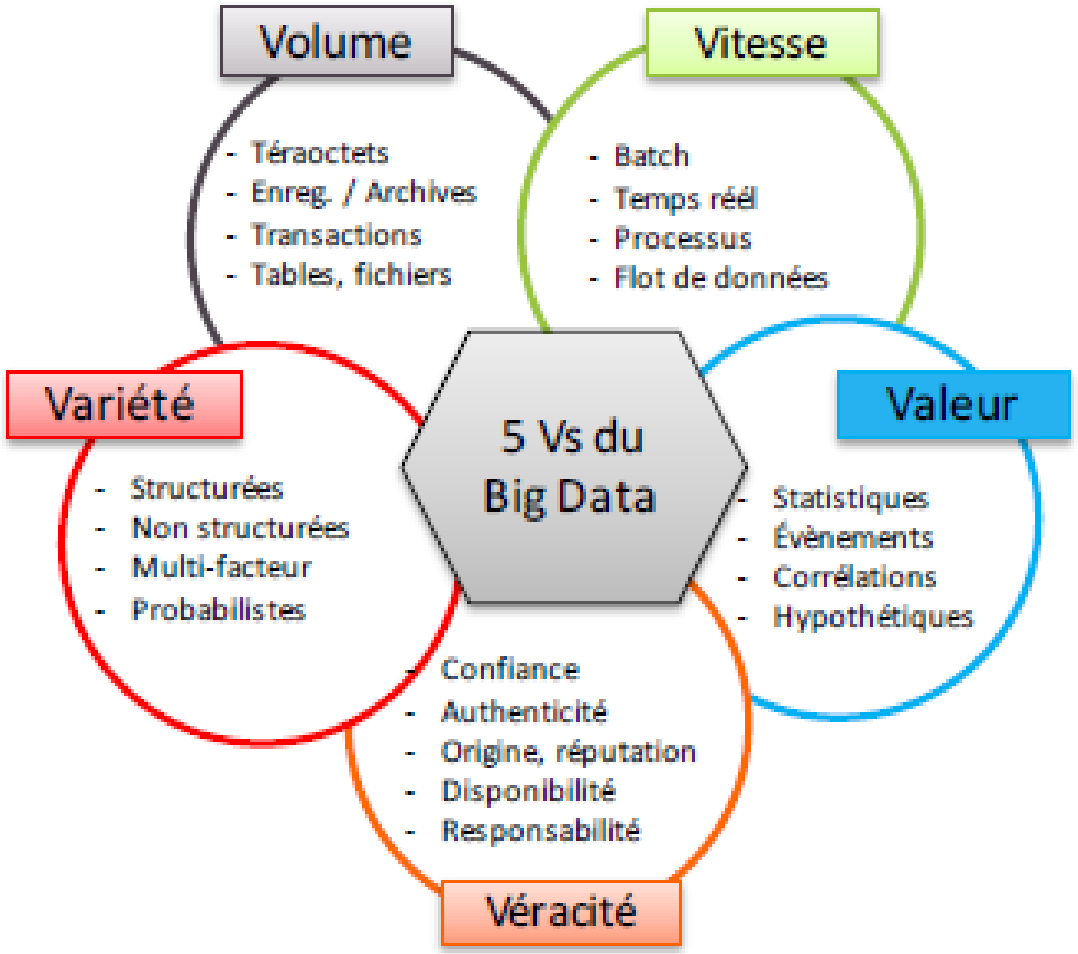
Le terme Big Data se réfère aux technologies qui permettent aux entreprises d'analyser rapidement un volume de données très important et d'obtenir une vue générale.

En mixant intégration et stockage de données, analyse prédictive et applications, le Big Data permet de gagner en temps, en efficacité et en qualité dans l'interprétation de données.

Big Data : Les 3 étends à 5 V

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)



Volume

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

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

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)

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…)

Vélocité

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 ?

Il ne suffit pas de savoir quel article un client a acheté ou réservé
Si on sait que vous avez passé plus de 5mn à consulter un article dans une boutique d'achat en ligne, il est possible de vous envoyer un email dès que cet article est soldé.

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…)

Question :

- 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 ?

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 :

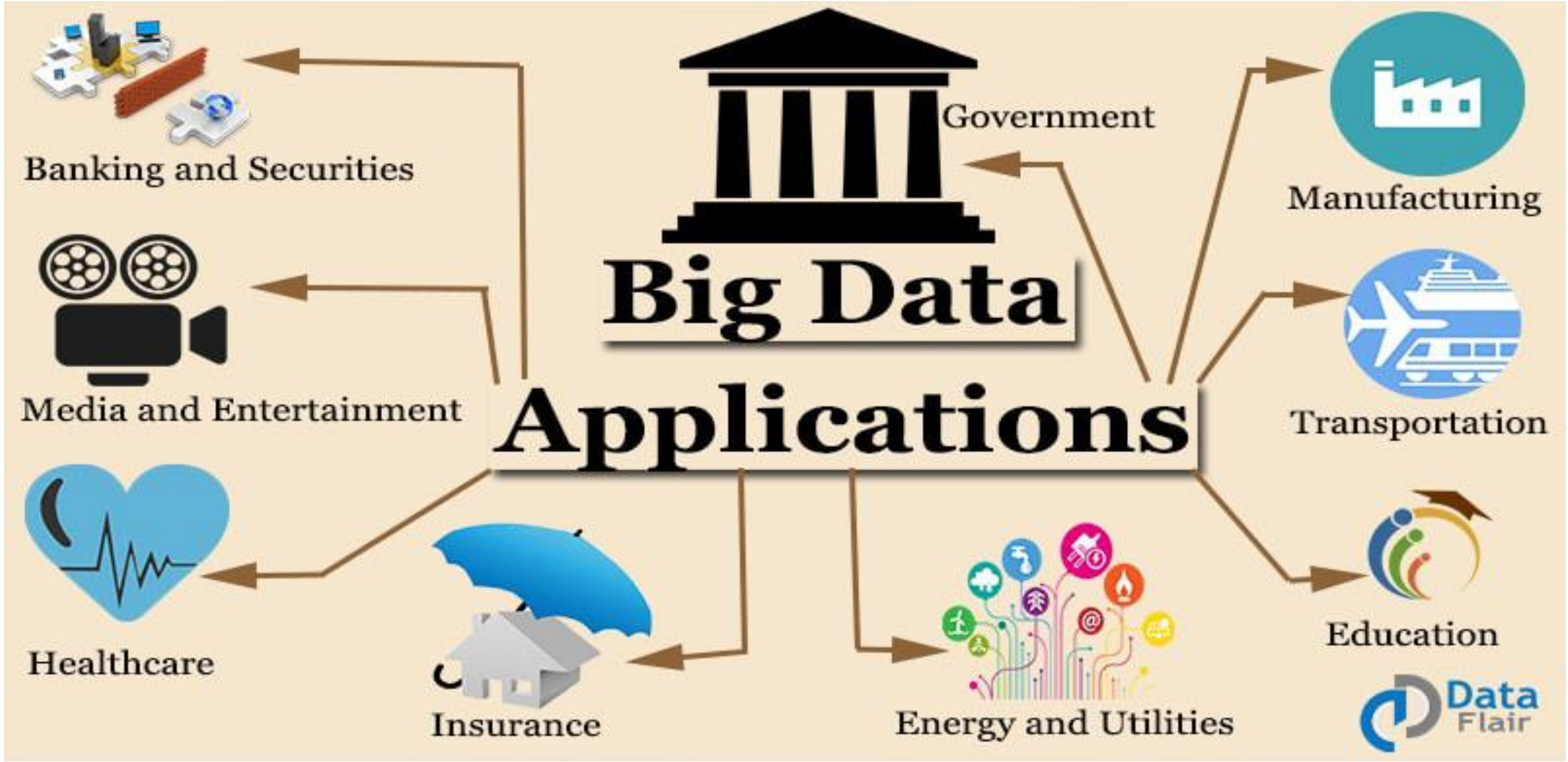
- Comment déterminer dans le déluge d'informations (infobésité) ce qui est utilisable ?
- Comment transformer les données en valeurs exploitables ?

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

- L’écosystème Hadoop
- Bases de données NoSQL