Université Sidi Mohamed Ben Abdellah
Faculté des Sciences Dhar El Mahraz
Fès

Cours Big Data

Master Recherche en Informatique Décisionnelle et Vision Intelligente (MIDVI)

Préparé par :

Pr. Noura AHERRAHROU

Mise en contexte et problématique

Aujourd'hui, on parle beaucoup d'un terme actuellement à la mode « ***Big Data*** » et du fameux « *Framework* » de calcul « ***Hadoop***».

Faits

En une minute sur l'Internet, 4,3 Million vidéos sont vues sur Youtube, 973,000 utilisateurs rejoignemnt Facebook éditent 3,4 millions de statuts et génèrent 4 GB de données digitales, Google répond à 3,7 Millions recherches et reçoit 126 heures de vidéos et pas moins de 700 nouveaux utilisateurs rejoignent Twitter. Au même moment, 480 000 tweets sont générés et 38 Millions de messages envoyés via whatsapp

Faits

2,4 Million de snaps sont générés. 187 Millions emails envoyées. D'un autre côté, dans le monde commercial, 862,823 dollars de commandes faites en ligne. Ces statistiques prouvent que le monde devient le théâtre d'un accroissement de données sans précèdent.

Ces données proviennent d'où???

Actuellement, le volume des données en circulation connaît une démultiplication permanente

90%
des données du monde a été créé au cours des deux dernières années

80%
Des données du monde est aujourd'hui non structuré

20%
des données disponibles peuvent être traitées par les systèmes traditionnels

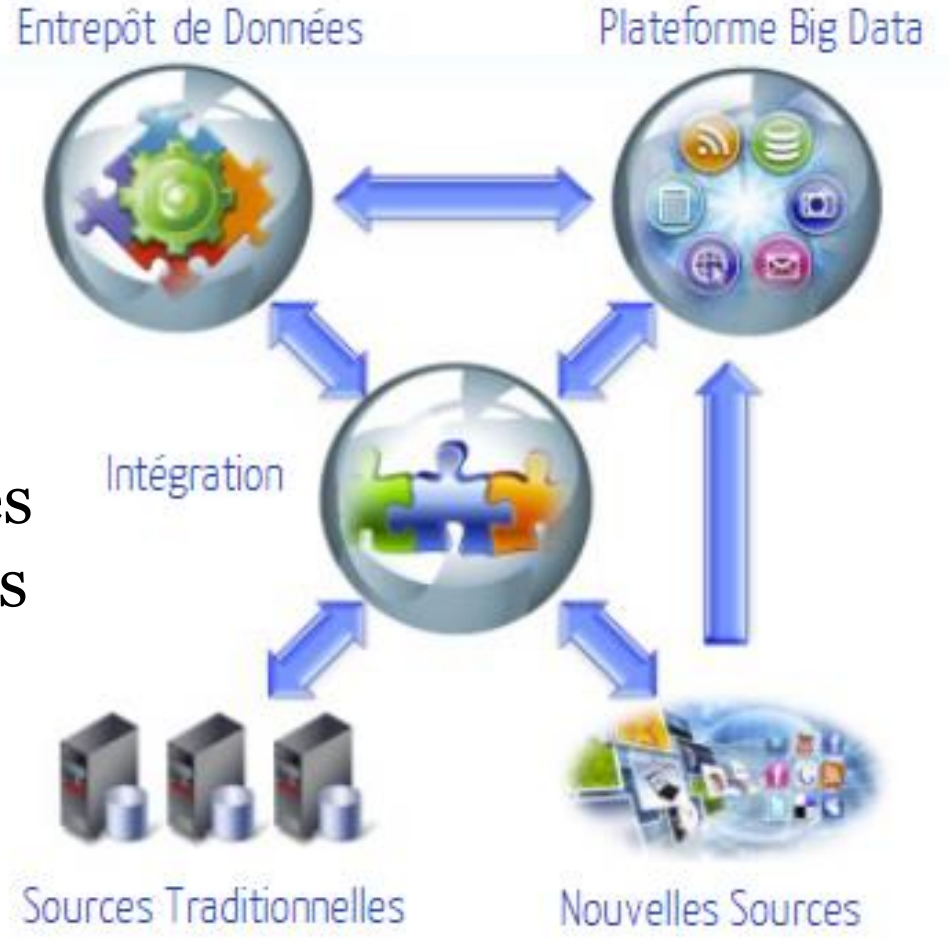
1 in 2
les chefs d'entreprise n'ont pas accès aux données dont ils ont besoin

1/3
Chefs d'entreprises ne font pas confiance aux informations qu'ils utilisent pour prendre des décisions.

5.4X
plus susceptibles que les plus performants utilisent l'analyse commerciale

Pourquoi Adopter l'approche Big Data?

L'échelle de cette croissance des données surpasse la capacité raisonnable des **technologies traditionnelles** et plus précisément celle des Systèmes de gestion de bases de données relationnelles (SGBDR) – ou même la configuration matérielle typique permettant l'accès à ces données.



Mise en contexte

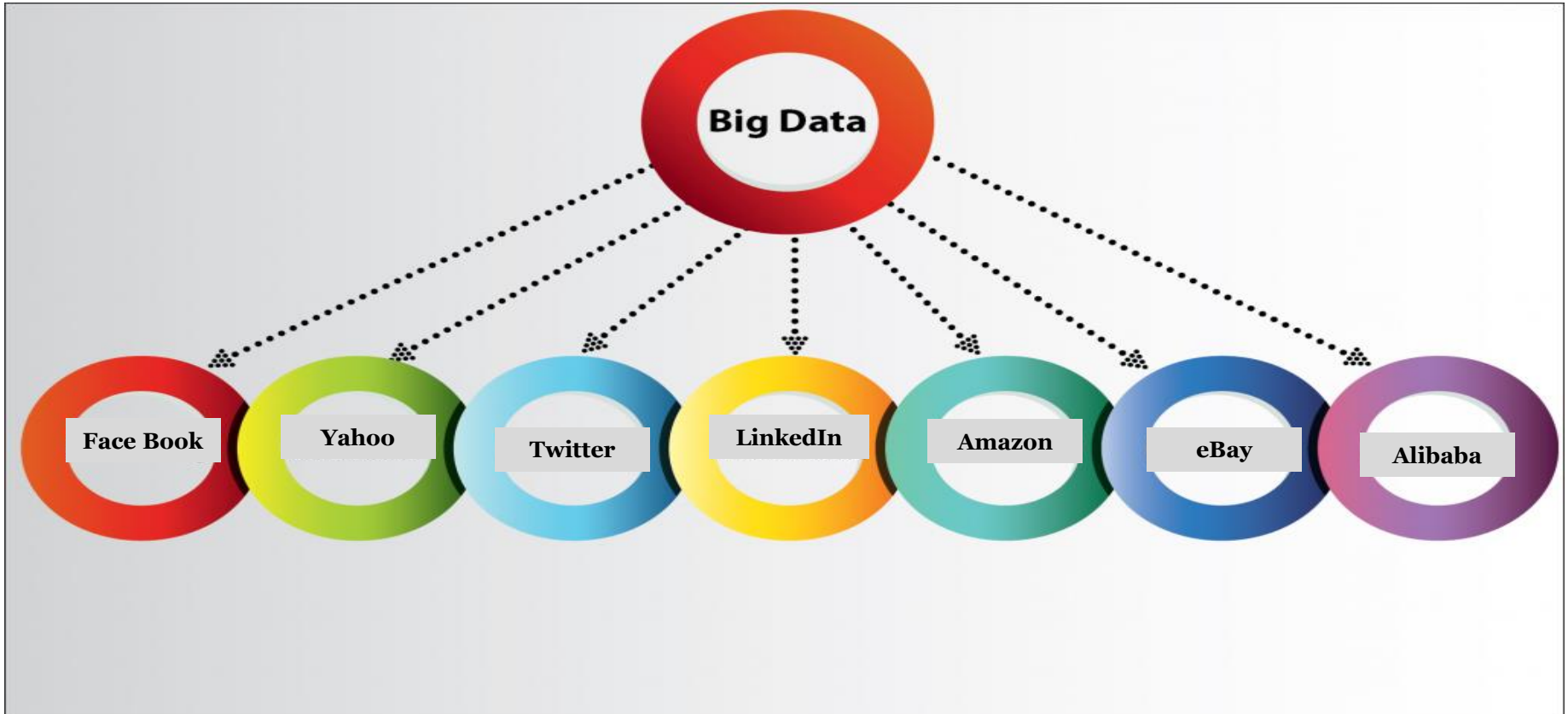
Les entreprises doivent trouver le moyen de maîtriser et traiter efficacement ces données pour continuer à servir fidèlement leur clientèle et rester compétitives. Google fait partie des entreprises qui ont très tôt ressenti le besoin de gérer efficacement les gros volumes de données liés aux requêtes des utilisateurs.

L'approche conceptuelle de Google

Pour répondre à ces défis, l'idée de Google était de développer une approche conceptuelle consistant, d'une part, à **distribuer le stockage** des données et, d'autre part, à **paralléliser le traitement** de ces données sur plusieurs noeuds d'une grappe de calcul (un cluster d'ordinateurs).

A l'origine des Big Data

L'approche conceptuelle qui a été adoptée par Google au paravent a envahit l'ensemble des grands acteurs du web actuellement.



C'est quoi selon vous le Big Data ??

Définition de Big Data

Il s'agit d'un ensemble de technologies, d'architecture, d'outils et de procédures permettant à une organisation de très rapidement capter, traiter et analyser de larges quantités et contenus hétérogènes et changeants, et d'en extraire les informations pertinentes à un coût accessible.

Définition de Big Data

Le terme Big Data réfère à la croissance exponentielle des données, au traitement de ces dernières ou de manière plus générale à toutes les étapes entrant en jeu dans le processus d'extraction d'informations utiles à partir de l'énorme lot de données brutes [Tudoran, 2014]

Définition de Big Data

L'expression « big data » a été proposée par le Gartner Institute : « le big data est un fort volume

de données très variées et produites très rapidement, qui exigent des techniques innovantes et rentables de traitement d'information pour une meilleure prise de décision ».

Définition de Big Data

Selon la Commission générale de terminologie et de néologie française, le big data, fait référence aux « données, structurées ou non, dont le très grand volume
requiert des outils d'analyse adaptés. »

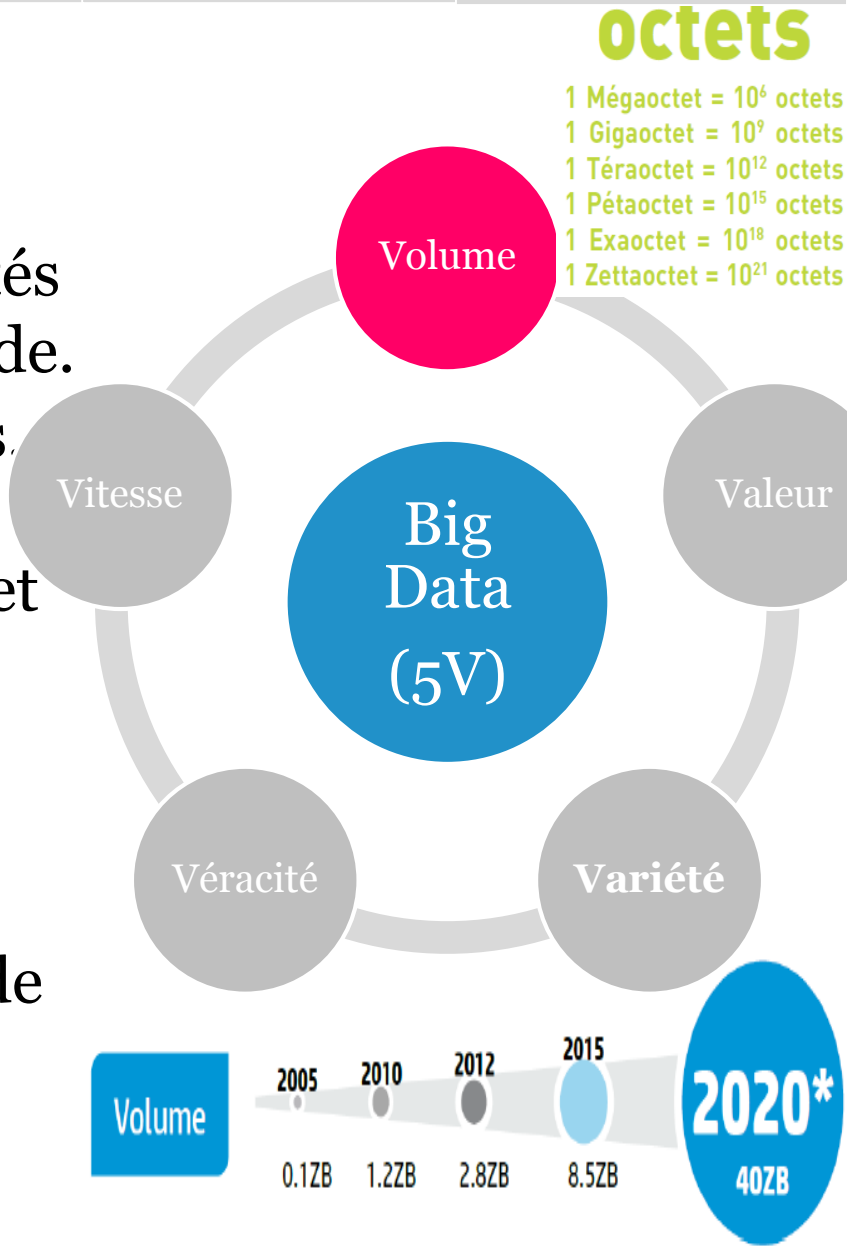
Les 5 v

Même si les définitions diffèrent, elles s'articulent autour de certaines caractéristiques que partagent les données. Il s'agit originellement des « 5V » du Big Data.

Tirer parti des Big Data implique d'intégrer à la fois volume, vélocité, variété, véracité et valeur des données considérées.

Volume

- Fait référence aux énormes quantités de données générées chaque seconde.
- Il suffit de penser à tous les e-mails, tweets, photos, vidéos, les données des capteurs que nous produisons et partageons chaque seconde.
- Sur Facebook seulement, nous envoyons 10 millions de messages par jour, « Likons » 4,5 millions de fois et téléchargeons 350 millions de nouvelles photos chaque jour.

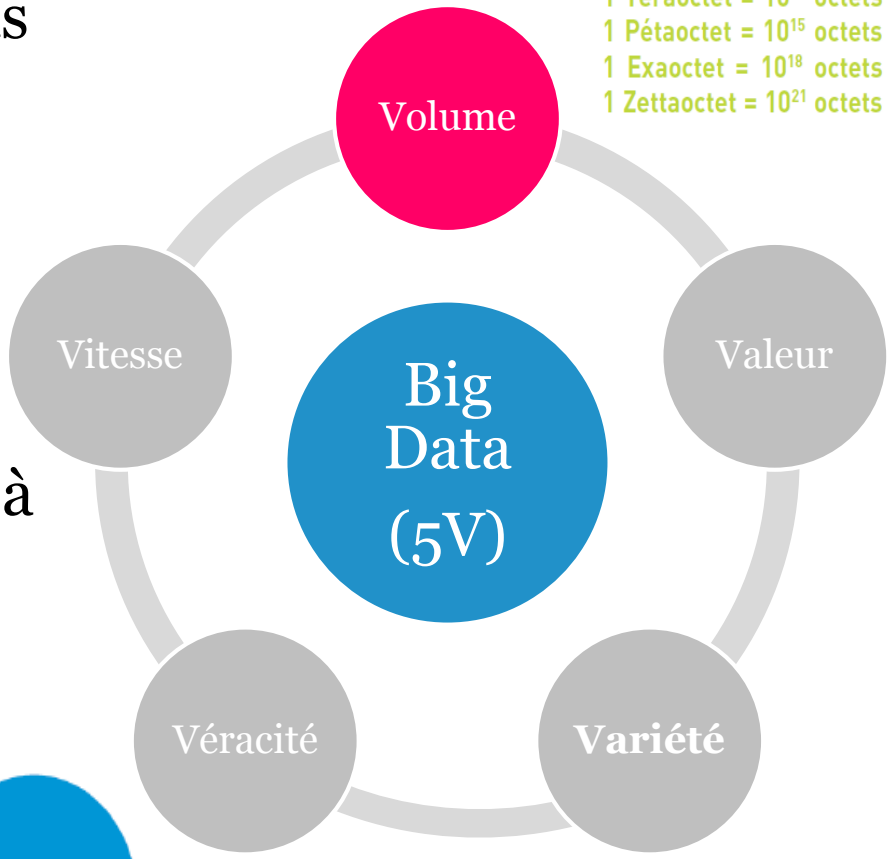


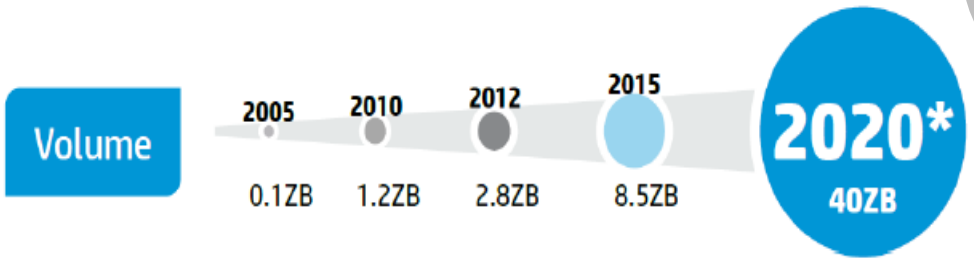
Volume

- Chaque jour, 2.5 trillions d'octets de données sont générées.
- 90% des données créées dans le monde l'ont été au cours des dernières années.
- Prévision d'une croissance de 800% des quantités de données à traiter d'ici à 2020.
- On estime qu'en 2020, 40ZB seront générés.

octets

1 Mégaoctet = 10^6 octets
1 Gigaoctet = 10^9 octets
1 Téraoctet = 10^{12} octets
1 Pétaoctet = 10^{15} octets
1 Exaoctet = 10^{18} octets
1 Zettaoctet = 10^{21} octets



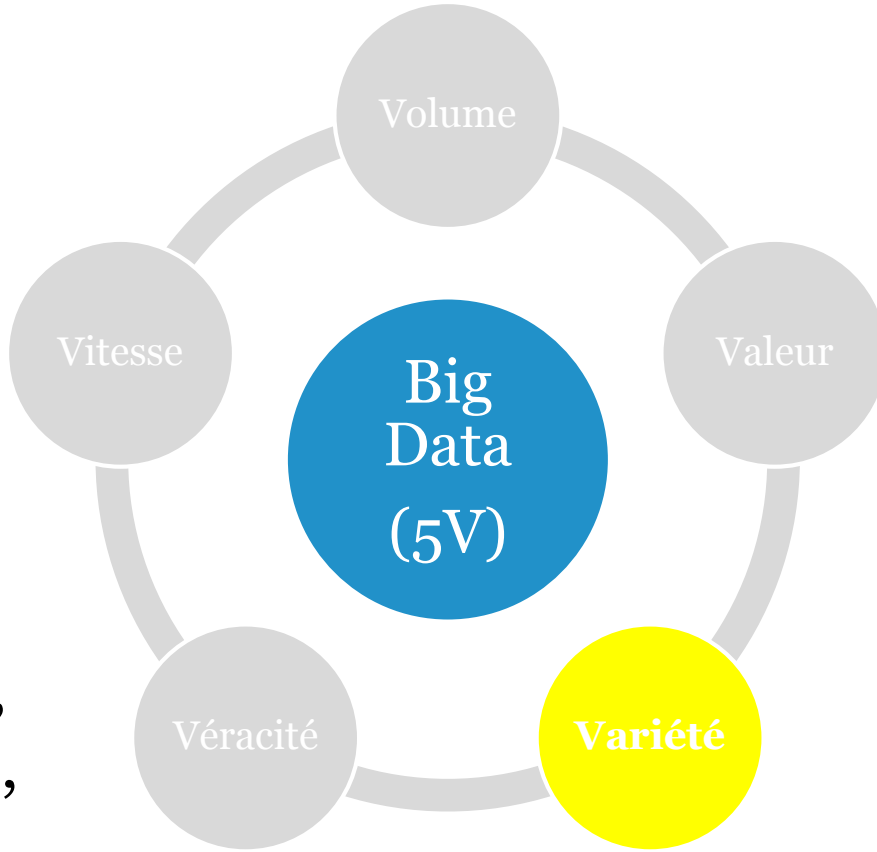


Variété

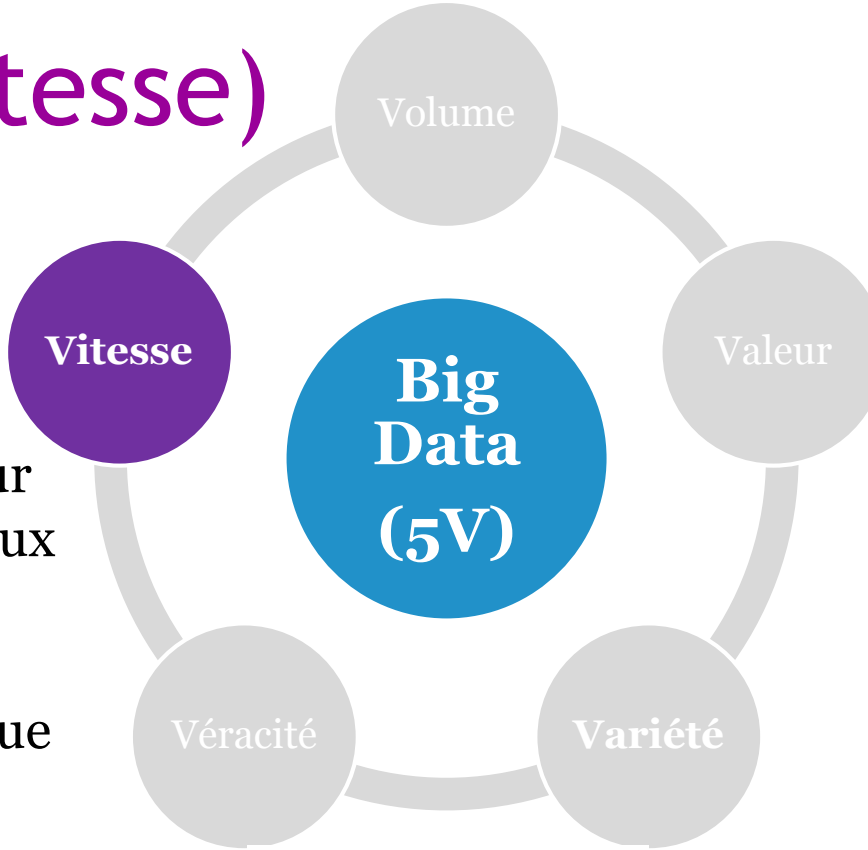
Fait référence aux différents types de données que nous pouvons utiliser.

ces données sont très diversifiées. Autant par leurs provenances (Réseaux sociaux, emails, historique de navigation Internet, échange vocaux,...), que par leurs formats (texte, images, vidéos, ...), domaines auxquels elles sont liées, structurées ou

non structurées....



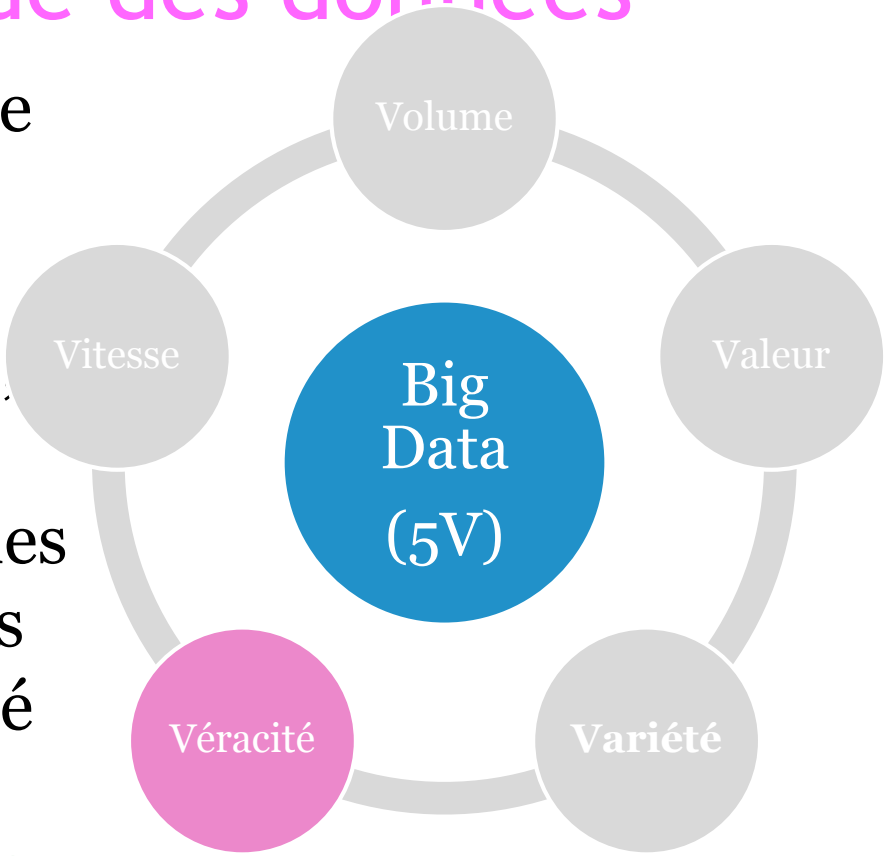
Vélocité (Vitesse)



- Fait référence à la vitesse à laquelle la nouvelle donnée est générée et traitée.
- Un système big data optimisé doit apporter la bonne réponse au bon moment ; Pensez juste aux messages sur les réseaux sociaux qui deviennent viraux en quelques secondes, les transactions bancaires frauduleuses détectées en quelques minutes ou encore le temps que prennent les logiciels pour analyser les réseaux sociaux et capter les comportements qui déclenchent l'achat.
- Le Big Data nous permet aujourd'hui d'analyser les données pendant qu'elles sont générées, sans avoir à les analyser dans des bases de données.

Véracité: la certitude des données

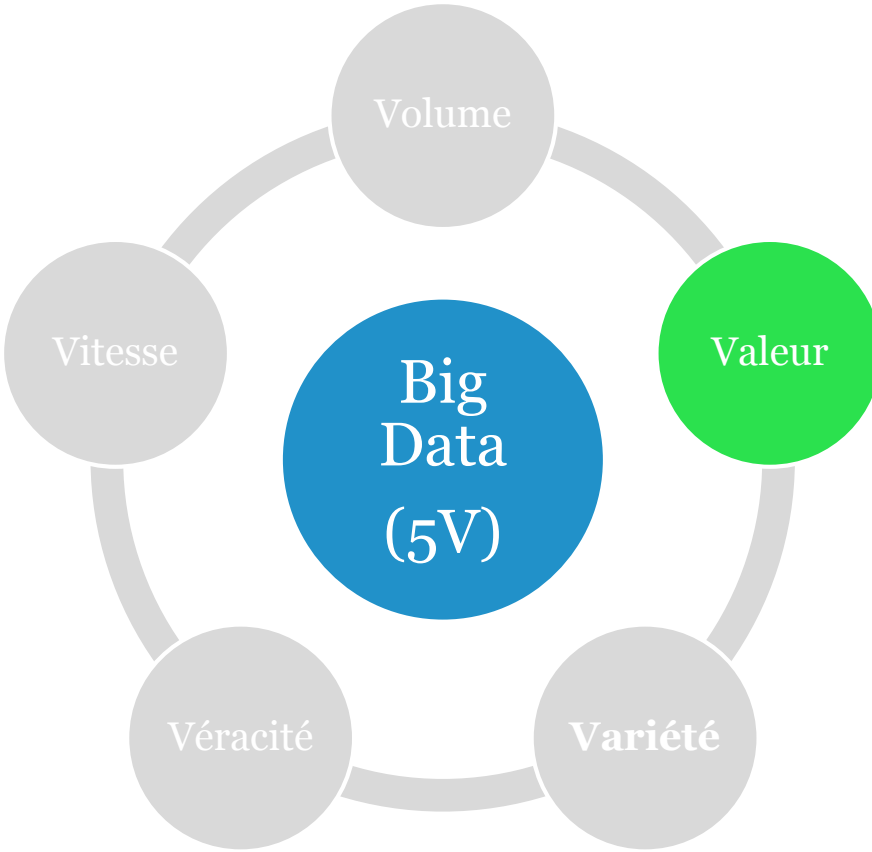
- c'est l'un des enjeux majeurs de l'exploitation des Big Data.
- Faux profils sur les réseaux sociaux, fautes d'orthographes, fraudes, ...
- Il est nécessaire de multiplier les précautions pour minimiser les biais liés au manque de fiabilité du Big Data.



Valeur

Sûrement c'est le point le plus important des 5.

Les technologies de stockage et d'analyse des Big Data n'ont de sens que s'elles apportent de la valeur ajoutée.



Challenges des Big Data

- Réunir un grand volume de données variées pour trouver de nouvelles idées
- Capturer des données créées rapidement
- Sauvegarder toutes ces données
- Traiter ces données et les utiliser

Challenges des Big Data

- Les données recèlent de trésors de connaissance. Nous pouvons extraire de ces données des informations, des tendances voire des prédictions, afin d'anticiper et suivre des phénomènes.
- Le traitement, le stockage et l'analyse de ces mégadonnées sont donc de vrais enjeux. Ainsi les données sont décrites comme l'or noir du numérique, ou l'or d'aujourd'hui tout court.

Challenges

Big Data Exploration

Trouvez, visualisez, comprenez toutes les données volumineuses pour améliorer la prise de décision

Vue à 360 ° améliorée du client

Étendez les vues client existantes en intégrant des sources de données internes et externes supplémentaires

Extension sécurité / renseignement

Réduction du risque, détection de la fraude et surveillance de la cybersécurité en temps réel

Analyse des opérations

Intégrez les capacités Big Data et Data Warehouse pour augmenter l'efficacité opérationnelle

Augmentation de l'entrepôt de données

Analyser une variété de données de machine pour améliorer les résultats de l'entreprise

Les applications du Big Data

Monitoring : contrôle, veille, domotique, ou surveillance

Technologies mettables: réalité virtuelle ou augmentée

Réseaux sociaux: Facebook, Twitter, Amazon, etc.

Processus: « Smart Cities », Détection de pannes, etc.

Amélioration de l'expérience: client (marketing), sport ou santé, etc.

Ciblage: Marketing, Risque, Fraude, etc.

Analyse des **flux**, consommation énergétique, etc.

Analyse de **comportements** (géolocalisation, indicateurs corporels, etc.)

Text-mining et détection de **besoins** « différents » de ce que l'existant permet

LES BIG DATA TROUVENT UNE APPLICATION DANS DE NOMBREUX DOMAINES D'ACTIVITÉS :

Sciences & recherche : au CERN (Comité Européen sur la Recherche Nucléaire) pour les calculs de données générées par le LHC (Large Hadron Collider), séquençage de l'ADN

Développement durable : paramétrage des éoliennes après collecte de données météorologiques

Education en ligne : activité des élèves, façon de suivre les programmes, pour amélioration des modes d'enseignement.

Transports : amélioration des horaires et des trajets desservis après collecte des mouvements des usagers.

Politique : L'analyse de Big Data ajoué un rôle important dans la campagne de ré-élection de Barack Obama, notamment pour analyser les opinions

Santé : Analyse des données globales des patients et des résultats pour comparer l'efficacité des différentes interventions.
Prédire et prévenir des