

Big data

J'ai choisi le thème de la big data à étudier puisque c'est un sujet qui m'intéresse énormément, j'aimerais me tourner vers ce domaine dans ma vie professionnelle. Je n'avais pas de réelles connaissances sur le sujet, c'est donc un avantage pour moi d'acquérir des connaissances qui me seront utiles pour ce que je veux faire. De plus, cela me permettra de m'encrer dans le sujet et de pouvoir mieux apprivoiser mon environnement par la suite.

J'ai accompli ma veille grâce au site lebigdata.fr qui recense toutes les actualités de la big data, j'ai suivi des comptes twitter qui font de la veille quotidiennement sur le sujet et des vidéos Youtube et je peux grâce à un outils de listage organiser mes recherches. J'ai bien sur complété avec le site de CNIL pour le côté juridique -qui est très important sur ce sujet-, des moyens audiovisuels et je reste à l'écoute de mon environnement quotidien pour déceler des informations utiles.

SOMMAIRE :


- I. Définition
- II. Historique et phénomène
- III. Utilisation
- IV. Intérêts et apports
- V. Avenir
- VI. Problèmes juridiques
- VII. Conclusion

I - Définition

Le terme big data est très vaste, en effet, il n'existe pas de définition officielle. Les réponses que j'ai pu trouver sont parfois conséquentes et compliquées, mais j'ai choisi de reprendre la définition du Youtuber Cookie Connecté qui me paraît la plus simple et la plus complète.

Le big data est donc un moyen de répondre à une volumétrie massive de données. Avec un enjeu principal qui est de valoriser ces données quel que soit le volume et leur nature, grâce à des outils technologiques.

Puis s'ajoute à cette définition le consensus qui s'accorde sur les **trois V (volume, vitesse, variété)** auxquels **deux V supplémentaires (véracité, valeur)** qui sont communément admis :



DATA

Big Data : Les 5V

- ▷ **Volume**
Enormément de données
- ▷ **Véracité**
La problématique de la fiabilité
- ▷ **Vélocité**
Toujours plus de données
- ▷ **Valeur**
La problématique du tri des données
- ▷ **Variété**
Des sources et formats

VENDEZ.PLUS

Un peu plus de détails :

- Volume : En 2000 seulement 20% des données étaient numérique, à ce jour 90% des données sur Terre ont été produites pendant les 24 derniers mois.
- Vélocité : L'augmentation fulgurante est dû à l'élargissement des individus observés (statistiques) et la fréquence des observations (appareils connectés en permanence)
- Variété : Les données ne sont pas forcément structurées environ 80% ne le sont pas comme celles des réseaux sociaux
- Véracité : elle est menacée à cause des multiples points de collecte, différents formats (redondance), formulaires faux et même les faux profils voire les robots d'internet
- Valeur : Surcharge informationnelle qu'il faut gérer

Pour stocker les données destinées à la big data les bases de données NoSQL sont préférables, elles assureront une scalabilité horizontale par l'abandon de la structure de donnée relationnelle au profit de la donnée qui porte elle-même sa cohérence. Ces systèmes permettent de réaliser des contrôles d'intégrité en temps réel et garantie, en même temps, la cohérence, la disponibilité et la résistance au morcellement des données.

II - Historique et phénomène

Avant de continuer, nous allons nous pencher sur l'histoire et la venue du terme big data, à quel moment la notion de "big" est apparue et pourquoi.

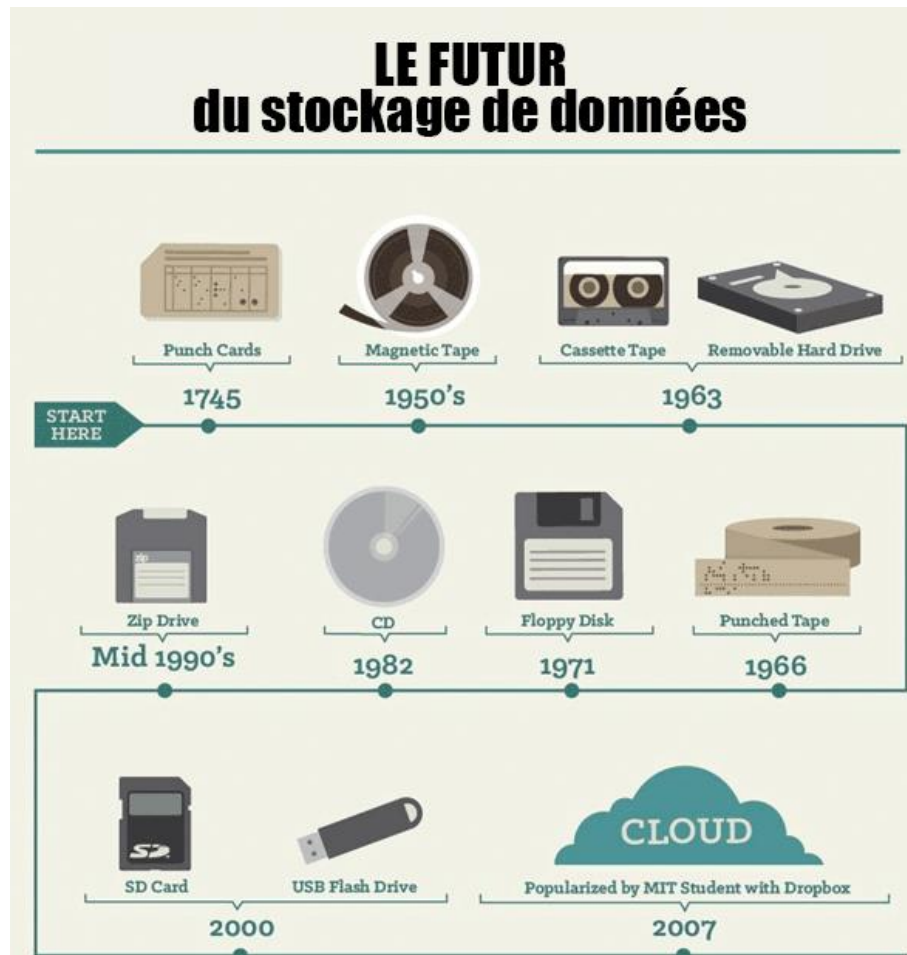
1. Chronologie

- 1950 : Etudes de statistiques de quelques centaines de personnes recueillis en laboratoire dans un cadre de recherches scientifiques et humaines.
- 1960-1980 : Apparition de l'informatique qui permet l'analyse statistique et d'ainsi élargie les études à des milliers de personnes.
- 1980-1990 : Apparition du data mining (exploration des données) grâce au développement des outils informatique de plus en plus accessibles à tous, qui permettent l'automatisation ou semi-automatisation de l'extraction des données. C'est alors que l'analyse statistique s'est déplacée dans un cadre professionnel, notamment dans l'aide à la décision et aux prévisions.
- 2010 : Apparition de la big data (mais terme utilisé pour la première fois en 1997 par la NASA) caractérisé par son volume conséquent et ses sources diverses. On n'analyse plus uniquement les données interne mais on prend en compte également les données externes. Ce qui apporte des centaines de millions de données = big data !

Rapport de la NASA :

https://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/parallel_volviz/paging_outofcore_viz97.pdf

2. Les évolutions des moyens de stockage des données



3. L'explosion du volume des données

L'expansion du volume des données et de l'ordre du pétaoctet (10^{15} octets). Cela est dû à la disparité de celles-ci dont on retrouve les principaux types qui sont :

- Les données *sociodémographiques* : âge, situation matrimoniale, ...
- Les données *comportementales* : sms, carte bancaires, usage d'une application, ...
- Les données *CRM* (Customer Relationship Management) : données clients
- Les données *open data* : données publiques, administratives, transports, santé, ...
- Les données *externes capteurs industrielles* : trafic routier, données climatiques et électriques, RFID, NFC, empreintes digitales, reconnaissance faciale, ...
- Les données de *géolocalisation* : GPS, IP, ...

- Les données de *tracking* : cookies, visites, surveillance, ...
- Les données *partagées et opinion* : réseaux sociaux, photos, vidéos, blogs, articles, musique, commentaires, ...

III - Utilisation

Nous comprenons donc grâce à l'historique de la big data que le traitement et l'analyse d'informations massives n'est pas nouveau et ne se résume pas à internet. Des entreprises tel que des supermarchés ou même des administrations géraient déjà des milliards de données bien avant internet. Bien sûr grâce à l'informatisation et l'évolution des technologies on peut maintenant utiliser les données dans tous les domaines comme :

- Les *transports* : fixation instantanée des prix, trafic en temps réel, visualisations des places vacantes, ...
- *Marketing* : Promotions, publicités ciblées, suggestions d'activité et d'achat grâce à la géolocalisation, ...
- *Grande distribution* : Croisement compte client avec les promotions en cours, fidélité, analyse des tickets de caisse pour des promotions personnalisés, ...
- *Ressources humaines* : Analyse des candidatures, trie par filtre, ...
- *Sciences* : Météorologie, épidémiologie, prévisions environnementales, ...

Actuellement, le modèle économique le plus rencontré est la fourniture de services contre des données personnelles. Certains GAFAM comme Facebook ou Google utilisent un système d'exploitation complet des données personnelles qui permet à des entreprises tierces d'y accéder et d'interagir en retour comme le cas des publicités ciblées.

Article de lesechos.fr : "Vendre ses données personnelles : un business controversé" de Fabiola Dor datant du 11 décembre 2020 <https://start.lesechos.fr/societe/culture-tendances/vendre-ses-donnees-personnelles-un-business-controverse-1273442>

IV - Intérêts et apports

L'intérêt principal du de la big data est de repenser la façon dont sont analysés les données, ce qui n'est plus une simple amélioration analytique mais une multitude de possibilités nouvelles afin d'augmenter les performances d'une entreprise.

Comparaison avec une analyse classique des données :

Tableau 1.1 Comparaison du Big Data et de l'analytique classique

	Big Data	Analytique classique
Type de données	Formats non structurés	En lignes et colonnes
Volume de données	100 téraoctets à plusieurs pétaoctets	Dizaines de téraoctets ou moins
Disponibilité des données	Flux constant	Pool statique
Méthode d'analyse	Apprentissage automatique	À base d'hypothèses
Objectif premier	Produits orientés données	Support aux décisions internes et services

Source: *Stratégie Big Data – Thomas Davenport – Edition Pearson*

https://www.pearsonelt.ch/download/media/9782744066177_SP_01.pdf

Les apports de la big data se font dans plusieurs domaines en voici quelques exemples :

- **Marketing** : Elle permet une connaissance approfondie des centres d'intérêts clients et donc d'adapter les offres surtout lorsque l'entreprise est en relation directe avec le client final, notamment en B to B.

Elle augmentera les parts de marché et de fidélisation, elle aidera à la décision, améliora les processus opérationnels et logistiques et permettra d'améliorer ses services.

- **Finances** : Grâce aux applications bancaires nous pouvons maintenant détecter des risques à la fraudes, suivre ses comptes en temps réel, détecter

des risques de surendettement, ... Et plus largement de détecter les variations brutales afin de d'anticiper et mettre en place une stabilisation.

Cela permet de réduire les risques, assurer une sécurité à ses clients ainsi qu'être efficace et rapide en termes de prise de décision.

- **Industrie** : L'industrie se sert de la big data à des fins de prévoyance, surveillance et business. Les compteurs linky par exemple permettent la prédiction en temps réel de la consommation électrique, la SNCF développe des capteurs spécifiques sur des matériels productifs afin de prévenir les interruptions, casses, ...

Tout cela permet de connaître les besoins clientèle, de réduire les opérations de maintenance, limiter les pannes et casses, tout en réduisant les dépenses.

- **Utilisateur** : Pour lui c'est une multitude de nouveaux services à disposition dans tous les domaines (aide à la conduite, informations en temps réel, déplacements administratifs réduits, ...).

C'est une source d'aide, voire d'assistanat, cela limite les cas d'erreurs, ...

- **Médecine** : Il est possible d'atteindre des avancées bénéfiques grâce à des analyses comme celle des mots-clefs de recherche Google qui permettrait d'anticiper d'une semaine les veilles sanitaires épidémiques.

Cela permet une prévoyance, une limitation des risques et des contagions tout en ne perdant pas de temps.

Source : Pubmed.gov "Detecting influenza epidemics using search engine query data" par Jeremy Ginsberg datant du 19 février 2009 <https://pubmed.ncbi.nlm.nih.gov/19020500/>

Et bien d'autres domaines encore tel que la **politique** (statistiques des votes), la **croissance d'une société** (amélioration urbaines), ...

V - Avenir



1. Un secteur porteur

La big data est en carence d'effectifs dans tous les pays, en effet, une **prise de conscience trop tardive** du potentiel des données et la **dévalorisation des métiers associés**, comme data scientist qui se voit à ½ demande sur 1 offre sur le secteur d'Annecy. Les entreprises doivent maintenant augmenter les salaires et baisser le niveau d'études (accessible bac +3 au lieu d'un +5) afin d'avoir des demandeurs. Les formations spécialisées tel qu'une licence BDD comme à Tétraz sont, elles aussi très récentes (environ 5 ans) mais tendent à se généraliser.

Néanmoins 2M d'emplois devrait continuer à être créés, 34 plans industriels ont été lancés par le gouvernement français en 2013 et le stockage dans la big data devrait augmenter considérablement entre 2021 et 2027.

Source : Journal l'action régionale "Stockage dans le Big Data devrait connaître un taux de croissance relativement plus élevé en 2021-2027 | Google, Microsoft Corporation, Amazon Web Services" par Sagar le 8 avril 2021. bit.ly/3mwAMZO

Fait impensable mais vrai la COVID-19 aura été bénéfique au secteur de la big data, elle lui a permis de s'accroître notamment en médecine en termes de statistiques, prévisions, collecte des informations des hôpitaux en temps réel, partage de connaissances mais aussi pour retrouver et tracer les personnes contagieuses, ...

Les ingénieurs se sont alors penchés dessus ainsi que sur l'IA pour développer des entrepôts de données et des technologies.

Source : lebigdata.fr "Coronavirus : comment l'IA et le Big Data aident à lutter contre le Covid-19" par Bastien L le 28 février 2020. <https://www.lebigdata.fr/sars-cov-2-ia-big-data>



2. Enjeux informatiques

La finalité de la big data n'est pas seulement la puissance de stockage, le traitement et le type de stockage qui tend plus vers le cloud. Mais un aspect **sécurité** est aussi primordiale quand autant de données sensibles et personnelles sont stockées.

Les données doivent être à l'abri des vols, pertes, détériorations, piratages, attaques, ... Des webinar, smart building (bâtiments connectés) ont été attaqués récemment par manque de cyber sécurisation, il suffit d'une faille, une entrée dans l'architecture réseau du bâtiment pour que des milliers de données très personnelles soient dérobées, ou même que le contrôle des objets connectés soient pris en main à distance ce qui peut être dans ce cas allant jusqu'à des risques mortels (chauffages, verrouillage des portes, ...). C'est pour cela que les données personnelles sont si précieuses pour les entreprises et ont un aspect financier, puisque qu'une notion de confiance est en amont.

Il faut aussi trier régulièrement la masse de données afin de garder une certaine **qualité et pertinence** de celles-ci.

Source : [lebigdata.fr](https://www.lebigdata.fr/batiments-connectes-cyberattaques-webinar-eaton) "Bâtiments connectés et cyber-attaques !" par Bastin L, le 7 avril 2021
<https://www.lebigdata.fr/batiments-connectes-cyberattaques-webinar-eaton>

VI - Problèmes juridiques



De nombreux scandales ont éclaté ces derniers temps concernant l'utilisation de nos données personnelles à des fins pécuniaires afin de développer des applications, robots intelligents (décryptent le discours des banques centrales) ou encore de technologies (voitures intelligentes, surveillance) collectant des données personnelles sans le consentement des usagers, dépassant donc la liberté individuelle. Il n'y a pas de véritables preuves puisque les scandales sont vite passés sous silence mais le sujet devient de plus en plus sensible mondialement.

Source : [lesechos.fr](https://www.lesechos.fr/tech-medias/intelligence-artificielle/le-mariage-explosif-de-nos-donnees-et-de-lia-1031813) "Le mariage explosif de nos données et de l'IA" par Rémy Demichelis le 24 juin 2019
<https://www.lesechos.fr/tech-medias/intelligence-artificielle/le-mariage-explosif-de-nos-donnees-et-de-lia-1031813>

Source : [lesechos.fr](https://www.lesechos.fr/finance) "Intelligence artificielle : quand des robots décryptent le discours des banques centrales" par Guillaume Benoit, le 28 juin 2019
[https://www.lesechos.fr/finance-](https://www.lesechos.fr/finance)

[marches/marches-financiers/intelligence-artificielle-quand-des-robots-decryptent-le-discours-des-banques-centrales-1033962](#)

Tout ce qui est réalisable n'est donc pas légalement, ni éthiquement possible. Toutes les données ne sont donc pas exploitables et leur utilisation pourrait se révéler improductive. De plus, même anonymisées, les données sont dépendantes de la loi n° 78-17 du 6 janvier 1978 « Informatique et Libertés » de la Commission nationale de l'informatique et des libertés (CNIL) liée à la collecte des données à caractère personnel car l'anonymisation n'offre pas une protection totale, chaque internaute peut être associé à un profil précis.

Les sanctions prévues par le code pénal vont jusqu'à 1.5 millions d'euros et 5 ans d'emprisonnement pour la personne morale ou physique qui traite des données personnelles illégalement.

Les sanctions prévues par la CNIL vont jusqu'à 150 000 € et le double en cas de récidive. De plus, la CNIL dispose du pouvoir de publier les sanctions qu'elle prononce, portant ainsi atteinte à l'image et à la réputation.

VII - Conclusion

La big data n'a pas encore atteint sa maturité, elle émerge doucement mais n'a pas encore de cadre légal sérieux, nous sommes donc encore dans l'attente de résultats finaux.

Pourtant d'allure très prometteuse avec les données comme matière première, un taux de croissance de plus de 30%, un marché de plus de 24 milliards d'euros en expansion et de nombreuses multinationales qui investissent. Les PME, elles, sont en difficultés pour se projeter par manque de moyens et de subventions.

La big data est assimilée à l'avenir et à la réussite des entreprises. Néanmoins il ne faut pas en oublier les échecs qui ont un taux plus élevé que ceux des projets traditionnels.

En bref, la big data est beaucoup trop jeune pour tirer des conclusions net et précise totalement positive ou négative.

Source : zdnet.fr "Big Data : peu de projets, le plus souvent des échecs" par Christophe Auffray datant du 20 janvier 2015 <https://www.zdnet.fr/actualites/big-data-peu-de-projets-le-plus-souvent-des-echecs-39813301.htm>