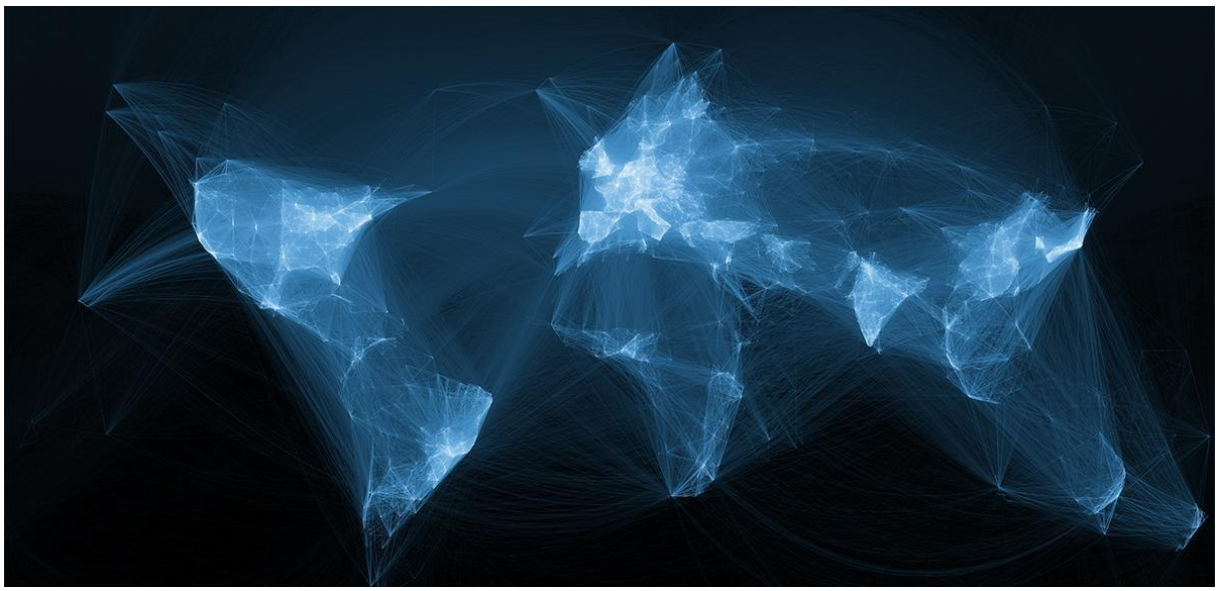
VEILLE TECHNOLOGIQUE :



LE BIG DATA

SITA Joseph Billy

2020-2022

I. LE BIG DATA

1. Présentation du Big Data

A- La définition du concept de Big Data

Le Big Data ne possède pas de définition précise à proprement parler, car c'est un « objet complexe polymorphe ».

On peut cependant le définir comme un ensemble massif de données, qui est inexploitable via un outil classique de gestion des données et de l'information. Les données de cet ensemble sont variées et préviennent de diverses sources (signaux GPS, messages, recherches web etc…).

Ces données sont stockées dans des data-centers

B- Les « 5 V » du Big Data

Le Big Data est donc soumis aux « 3 V » :

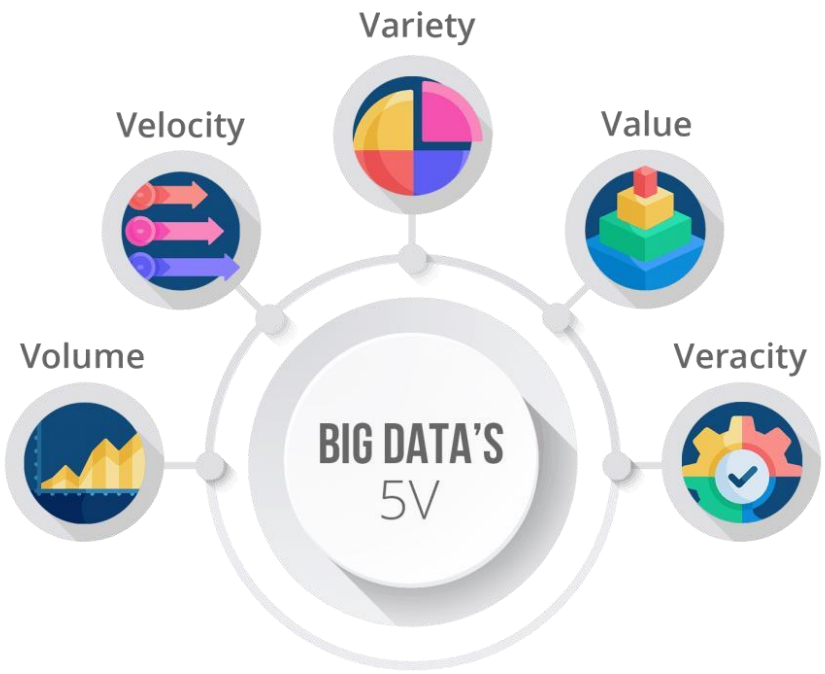
-Le Volume de données, considérable et en constante augmentation, trop volumineux pour être traité par des SGBD classiques.

-La Vitesse (ou Vélocité) de génération et de collecte des données. Avec l'évolution des technologies, leur propagation au grand public et leur multiplication (smartphone, ordinateur, objets connectés…), les données arrivent de plus en plus rapidement, et leur exploitation doit se faire en temps réel pour être rentable.

On peut ici faire le parallèle avec la vitesse actuelle de la publication d'un fait divers ou d'une information, remplacé toujours plus rapidement par un autre.

-La Variété des types de données, venant de supports toujours de plus en plus diverses (comme vu précédemment), ces données peuvent être des postes sur un réseau social, des données GPS, des fichiers audios, des messages, des images, des vidéos, des transactions, des données météo etc…

C’est cette variété qui rend difficile leur exploitation, et donc qui nécessite la création d’un logiciel spécialisé pour les jeux de données composant le Big Data.



Mais l’on peut aussi rajouter « 2 V » de plus en plus importants dans un monde générant de plus en plus de données :

-La Valeur de ces données. Il est désormais nécessaire de dénicher les informations les plus importantes dans des téraoctets de données et de s’y concentrer.

-La Véracité de ces données. La fiabilité des données est également d’une importance croissante car elles peuvent être générées par de faux profils, des robots ou encore être simplement des « fake news ».

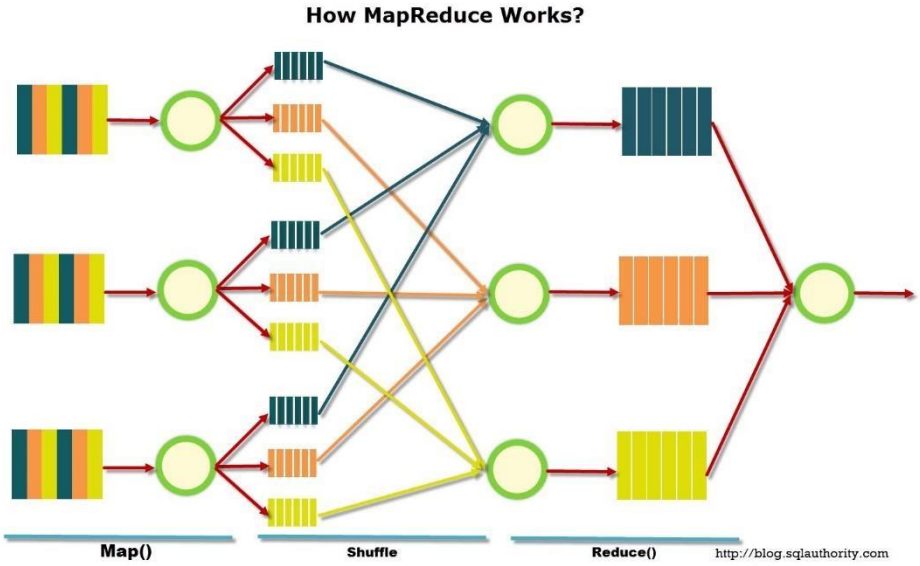
2. Les outils de gestion du Big Data

Comme vu précédemment, les SGBD classiques ne sont pas adaptés pour travailler avec un volume et une variété de données aussi importants.

De nouveaux outils spécialisés ont été développés face à la croissance du Big Data :

A- Hadoop

Hadoop est un framework Java développé en 2005 par Apache, traitant des ensembles de données de Big Data au moyen du modèle de programmation MapReduce (pattern d'architecture crée par Google), ce qui permet d'effectuer des calculs en parallèle afin d'améliorer la vitesse de traitement des données.



Il a également la particularité de voir ses données divisées en « chunks » de même taille, et séparés dans des serveurs a l'architecture simple afin d'en stocker un volume énorme et évolutif tout en ayant un cout de maintenance réduit.

Hadoop est utilisé par un grand nombre de géants du web comme IBM, Google, Intel, Microsoft, ou encore Amazon, et il est considéré comme le leader et l'outil le plus performant dans son domaine.

B- Apache Spark

Spark est aussi un framework open-source, de la fondation Apache, dont la première version est parue en 2014. Spark est aussi spécialisé dans le traitement massif de données mais aussi dans le machine learning (c'est-à-dire l'apprentissage afin d'effectuer des prédictions à partir de données, mais aussi la découverte de patterns permettant à une machine d'apprendre de son expérience sur des données structurées).

Son avantage principal est que les données sont traitées directement en mémoire et non sur le disque ce qui le rend plus rapide que ses concurrents.

Spark est notamment utilisé par Yahoo, TripAdvisor, Tencent ou encore Amazon.

C- Cassandra

Cassandra a été initialement développé par Facebook en 2008. C'est un SGBD(Système de gestion de base de données) NoSQL en open-source dont une des qualités principales est de résister lors de montée en charge de la demande de traitement de données, il est aussi peu défaillant.

Il est avec MongoDB, un des deux principaux SGBD NoSQL à être utilisé dans le cadre de traitement de volumes massifs de données.

Cassandra est notamment utilisé par Reddit, Twitter, Netflix, Spotify ou encore eBay.

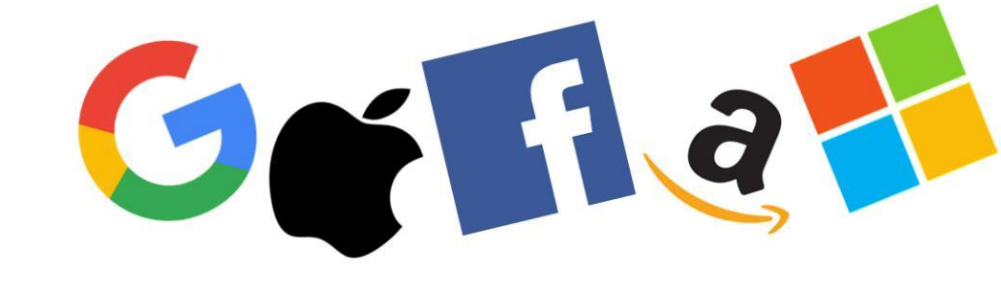
3. Les domaines d'utilisation du Big Data

Une de ses applications les plus visible consiste en la collecte des cookies sur le web a but commercial et plus spécialement dans le secteur du E-Commerce.

Utilisés par les navigateurs web afin d'optimiser et de cibler les publicités proposées, ces cookies peuvent également provenir de réseaux sociaux ou l'on divulgue

beaucoup de données personnelles, tels que Facebook ou Instagram, et ou l'actualisation en temps réel des publicités est visible.

La vente même de données personnelles à des annonceurs est également devenu un commerce, Google et Facebook en sont les leaders.



Le Big Data a révolutionné le monde de la finance et particulièrement le trading puisque ce sont désormais des intelligences artificielles qui achètent et revendent en quelques millisecondes avec une prise de risque et un bénéfice optimisé.

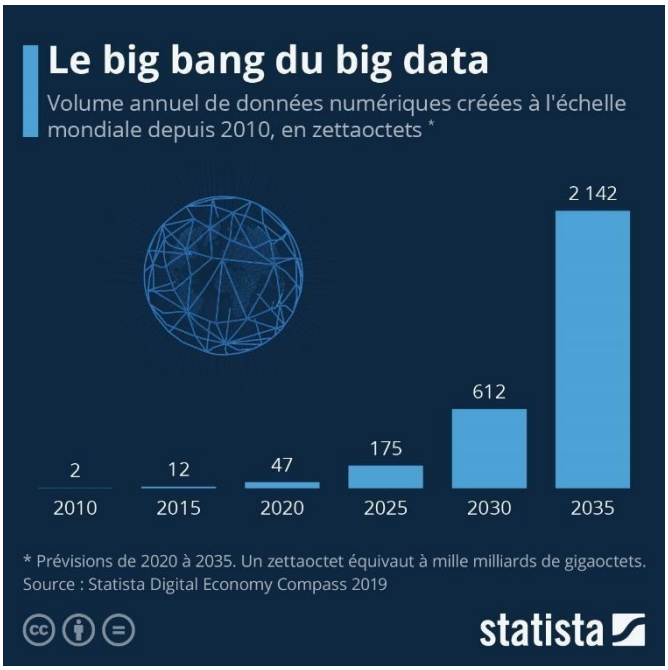
D'autres secteurs sont touchés par le Big Data :

- La santé (afin de détecter d'éventuelles épidémies)
- L'éducation (afin d'améliorer la détection d'élèves en difficultés)
- La recherche scientifique (afin d'évaluer les investissements)

En résumé on peut déterminer plusieurs utilisations du Big Data dans plusieurs et différents domaines

II. L'évolution récente du Big Data

A- La croissance du Big Data :



(Statista est un portail en ligne allemand offrant des statistiques issues de données d'instituts, d'études de marché et d'opinion ainsi que de données provenant du secteur économique.)

Le domaine du Big Data est en pleine croissance et la pandémie actuelle de Covid-19 a eu un impact non négligeable sur le secteur.

Celle-ci en combinaison aux confinements ont indirectement fait générer une masse de données bien plus grande qu'en temps normal ce qui a des conséquences directes sur les dépenses sur les data centers, liées directement au big data.

Les dépenses consacrées aux data centers des 20 sociétés leaders de l'hyperscale ont grimpé de 16% en 2020 avec 100 milliards de dollars de dépenses en 2020 dont 37 au 3^{e} trimestre, porté par les 4 géants du secteur : Amazon (avec Amazon Web Services), Google, Facebook (deux géants générant une masse de données gigantesque) et Microsoft (avec Microsoft Azure).

Cette croissance s'explique également par la croissance de la vente en ligne (avec Amazon en tête) ainsi la numérisation des entreprises entre autres, qui a pour but de mettre la donnée au centre de l'entreprise et non plus cantonnée au fonctions IT afin qu'elle soit exploitée a plein potentiel.

Pour finir, le confinement a d'ailleurs mis en lumière notre dépendance vis-à-vis des GAFAM via l'utilisation plus forte des réseaux sociaux, des moteurs de recherche ainsi que de logiciels comme Microsoft Teams.

B- L'importance du Big Data à travers la pandémie

Le Big Data a montré toute son utilité et sa puissance en cette période de pandémie, et permis la naissance de nouveaux outils du Big Data dans différents secteurs et notamment celui de la santé.

Dans le domaine de la santé, le Big Data a tout d'abord permis à la société canadienne BlueDot d'anticiper la propagation de la pandémie via sa plateforme de surveillance sanitaire portée par le Big Data et l'IA, en exploitant des millions de données médicales fournies par les actualités mondiales, les compagnies aériennes ou encore grâce aux rapports sur les maladies affectant la faune et la flore. Le Big Data avait également servi au début de la pandémie, à retracer les trajets et les contacts des premiers infectés.

III. Les enjeux futurs du Big Data

A l'image des « 5 V », le Big Data possède des enjeux majeurs sur lesquels il faut se questionner afin de mieux prévoir les problèmes et obstacles auxquels les acteurs du secteur ainsi que les entreprises vont devoir faire face, puisque le Big Data lui-même est en train de prendre une importance non-négligeable dans la société actuelle.

1- La sécurité des données

Les données ayant de plus en plus de valeur, elles suscitent les convoitisent de la part de hackers. Les cyberattaques visant d'importantes organisations, parfois des Etats, possédant des données sensibles, se multiplient.

Plusieurs exemples d'importantes cyberattaque, récentes, nous le montrent :

- Cyberattaque sur la Banque Centrale Néo-Zélandaise, touchant des données personnelles sensibles, ainsi que des informations commerciales
- Cyberattaque sur l'Agence Européenne des Médicaments, touchant des données liées au vaccin fourni par Pfizer/BioNTech, publiées récemment par les hackers
- Cyberattaque sur l'entreprise Américaine SolarWinds, et son logiciel Orion, qui a touché plus de 80% des entreprises du Fortune 500 et des Etats dont le Gouvernement Américain et ses agences fédérales

Il est donc primordial que les données soient protégées au mieux contre ces attaques pouvant toucher tout le monde à n'importe quel moment.

De plus, il faut miser sur la mise en place de nouveaux règlements à l'image du RGPD de l'Union Européenne afin que les données que l'utilisateur fourni à des sociétés ne soient pas ensuite utilisées de manière malveillante.

2- L'humanisation des données

Les données doivent être utilisées à des fins d'optimisation et de personnalisation d'un service client et non à des fins malveillantes comme vu plus haut. Elles doivent créer du lien entre le client et l'entreprise, c'est pour cela que cette dernière ne doit pas voir son client comme une masse de données qu'il va pouvoir exploiter afin de faire du profit, il faut utiliser ces données intelligemment.

De plus, un questionnement se pose sur le plan éthique puisque nous rentrons dans un monde de plus en plus régis par la donnée donc il est important pour les acteurs majeurs du secteur et les autorités de mener l'utilisation et le traitement de ces données vers une issue tournée vers l'humain.

IV. Sources exploitées pour la veille

Outils utilisés pour la veille :

Moteur de recherche : Microsoft Edge

Définition de la veille technologique :

https://fr.wikipedia.org/wiki/Veille_technologique

https://jobphoning.com/dictionnaire/veille-technologique# ***Veille***

active et veille passive :

https://megancortesblog.wordpress.com/veille-technologique/

http://jacques.breillat.fr/veille-strategique/veille-active-ou-veille-passive ***Le Big Data :***

https://fr.wikipedia.org/wiki/Big_data https://www.lebigdata.fr/definition-big-data

https://www.lebigdata.fr/marche-big-data-atteindrait-67-milliards-de-dollars-2021 ***Histoire***

du Big Data :

https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-bigdata/?sh=2707375e65a1#

Les 5 V du Big Data :

https://www.journaldunet.com/solutions/analytics/1102057-les-3-v-du-big-data-volumevitesse-et-variete/

https://www.definitions-marketing.com/definition/5v-du-big-data/

https://www.oracle.com/fr/big-data/what-is-big-data.html#link5 ***Les***

outils du Big Data :

https://superdatacamp.com/big-data/top-12-des-outils/ https://www.lebigdata.fr/top-7-outils-big-data-0712 https://fr.wikipedia.org/wiki/Scalability

https://fr.wikipedia.org/wiki/Partitionnement_de_donn%C3%A9es

https://www.jedha.co/blog/la-vraie-difference-entre-machine-learning-deep-learning

Les domaines d'utilisation du big data :

https://www.kdnuggets.com/2018/11/top-5-domains-big-data-analytics.html

https://www.la-croix.com/Economie/Economie-et-entreprises/Facebook-Google-geants-ventedonnees-personnelles-2019-12-29-1201068909

https://www.lcl.fr/mag/tendances/big-data-definition-enjeux-et-applications

https://www.lebigdata.fr/trading-en-ligne-comment-le-big-data-est-en-train-derevolutionner-le-monde-de-la-finance

https://www.lafinancepourtous.com/decryptages/finance-et-societe/nouvelleseconomies/big-data/les-applications-du-big-data/

https://www.thalesgroup.com/fr/big-data-et-aeronautique

https://www.thalesgroup.com/fr/marches-specifiques/systemes-dinformation-critiques-etcybersecurite/news/meteo-lere-du-big-data https://inventiv-it.fr/2017/11/17/10-cas-d-usage-big-data-dans-le-domaine-de-la-finance/ https://www.lebigdata.fr/big-data-service-de-leducation

L'évolution récente du Big Data :

-La croissance du big data en 2020 :

https://www.lebigdata.fr/depenses-data-center-explosent ***-Numérisation des entreprises et big data :***

http://www.economiematin.fr/news-donnees-croissance-entreprise-strategieinternet-dubois

Big Data contre covid-19 : https://www.rfi.fr/fr/podcasts/nouvelles-technologies/20201121-big-data-contre-covid-19 https://www.lebigdata.fr/sars-cov-2-ia-big-data https://www.lebigdata.fr/bluedot-ia-coronavirus-wuhan

https://www.lebigdata.fr/google-relance-tourisme-avec-big-data

https://www.lebigdata.fr/union-europeenne-hub-big-data-covid-19 ***L'évolution d'Apache Spark***

https://www.programmez.com/actualites/sortie-de-apache-spark-30-30691 https://big-data.developpez.com/actu/306530/Apache-Spark-la-version-3-0-du-frameworkopen-source-de-traitement-big-data-disponible-avec-une-amelioration-des-API-Python-unemeilleure-compatibilite-ANSI-SQL-et-est-deux-fois-plus-rapide/

L'importance du Big Data dans l'Avenir :

https://www.talend.com/fr/resources/future-big-data/

Les enjeux du Big Data :

- ***La sécurité des données :***

  https://www.lemonde.fr/international/article/2021/01/10/nouvelle-zelande-labanque-centrale-victime-d-un-piratage-informatique_6065772_3210.html
  https://www.lebigdata.fr/vaccin-pfizer-covid-donnees-hackers
  https://www.lebigdata.fr/russie-derriere-solarwinds-hack
  https://www.lebigdata.fr/solarwinds-cyberattaque-historique-usa

- ***L'humanisation des données :***

  https://blog.outscale.com/fr/les-5-grands-enjeux-du-big-data

- ***Le traitement des données :***

  https://blog.outscale.com/fr/les-5-grands-enjeux-du-big-data

- ***Le Big Data dans l'entreprise :***

  https://pic.digital/blog/les-enjeux-du-big-data-pour-les-entreprises/