

Les contraintes du Big data

Qu'est-ce que le Big data ?

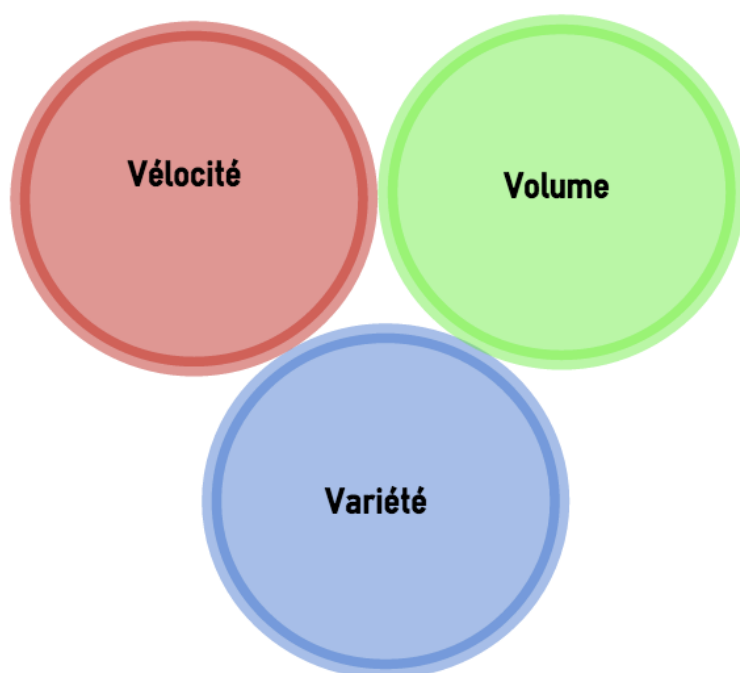
Selon Wikipédia le « Big Data » est une expression utilisée pour signifier un volume massif de données structurées et non structurées qui est si grande qu'il est difficile à traiter en utilisant les bases de données traditionnelles et les techniques logicielles

Enjeu :

Le Big Data représente un gros challenge pour les entreprises. Tout d'abord, les entreprises vont gagner un temps énorme en études de marché, segmentation et publicité. En effet, une entreprise ciblant un profil qu'il pourra analyser en temps réel leurs déplacements à différentes heures de la journée, établir des tendances et adapter la position de son magasin et de sa publicité en conséquence. Les entreprises vont employer énormément de moyens pour offrir des produits et services les plus personnalisés possibles. Beaucoup d'entreprises utilisent déjà Big Data Analytics pour mettre au point des produits anticipant nos désirs. Par exemple : Google grâce à son cloud, Facebook en sauvegardant chaque commentaire, chaque like et chaque photo et vidéo posté.

Nous allons vous parler des 3V qui représente le volume de données traités par jour ainsi que la vitesse à laquelle sont traités les données et la manière dont les données sont triées selon leur variété.

Les 3 v du Big Data



Les 3V :

Volume :

C'est le V le plus associé au Big Data, car le volume peut par définition être massif. Ce dont nous parlons ici, c'est de quantités de données qui atteignent des proportions presque inimaginables, on parle de donnée pouvant être en Pétaoctets voir en Yottaoctets (10^{24}).

Exemple : Sur tweeter il y a environ « 326 millions d'utilisateurs actifs par mois » ce qui représente environs 500 000 000 par jours de tweet qui doivent être stockées en plus du nombre de tweet permanent émis par les Twittos.

Vélocité :

La vitesse mesure la rapidité à laquelle les données affluent.

Exemple : Sur Facebook il y a 334 milliards d'images par an stockés mais les utilisateurs de Facebook y transfèrent plus de 900 millions de photos par jour.

En raison de l'augmentation des cyberattaques, de la cybercriminalité et du cyber espionnage, des charges utiles malveillantes peuvent se cacher dans ce flux de données qui transitent via le pare-feu.

Pour éviter la compromission, ce flux de données doit être examiné et analysé pour détecter les anomalies et les schémas de comportements qui constituent des signaux d'alarme. Or, cette opération se complique, car de plus en plus de données sont protégées à l'aide du cryptage. En même temps, les individus malveillants dissimulent leurs charges utiles de programmes malveillants à l'intérieur de paquets cryptés. Ce flux de données est le vecteur de la vitesse.

Variété :

Wikipédia, nous explique que le volume des big data met les centres de données face à un réel défi : la variété des données. Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées, voire non structurées (cependant, les données non structurées devront, pour utilisation, être structurées). Ce sont des données complexes provenant du web (Web mining), au format texte (text mining) et images (image mining). Elles peuvent être publiques (open data, Web des données), géo-démographiques par îlot (adresses IP), ou relever de la propriété des consommateurs. Ce qui les rend difficilement utilisables avec les outils traditionnels. La démultiplication des outils de collecte sur les individus et sur les objets permet d'amasser toujours plus de données. Et les analyses sont d'autant plus complexes qu'elles portent de plus en plus sur les liens entre des données de natures différentes.

Par exemple : Les différents fichiers que l'on peut trouver sur le Web sont les images JPEG, Tableurs Excel, des messages et d'autres encore.

Le 4e V :

Le 4e V permet de repérer les informations et les classés selon leur qualité . Il s'agit de la véracité.

Véracité :

Les chercheurs font face à un problème de taille : l'élaboration de logiciels suffisamment puissants traiter et analyser ce flux immense d'informations déstructurées. Afin d'en exploiter au maximum le potentiel, ces logiciels devront être capables de scanner les vidéos, images, en interpréter le sens, faire le lien entre énormément de sources différentes. De plus, ils devront être capables de détecter l'ironie dans le flux de messages de réseaux sociaux pour éviter des interprétations erronées. Si l'on base nos décisions sur des données il faut être sûr qu'elles apportent une « vérité » vraie, ou tout du moins utile. Cette validation de la véracité est cruciale pour ne pas être induit en erreur sans le savoir. Que peut-il arriver en fait, quelles sont les erreurs à éviter ? La plus banale des erreurs est celle d'utiliser des données trop anciennes pour créer des profils comportementaux. Les modèles seront précis sur des comportements qui ne se reproduiront plus. Les plus récentes sauf que certaines de ces données n'ont aucune valeur pour comprendre ou modéliser par exemple un comportement d'attrition. Les plus techniques. La multiplicité des données ne nous empêche pas de valider la pertinence de celles-ci en tant que variable, il faut même attacher plus d'importance à connaître son intérêt ou non

Pour faire simple il s'agit d'un système « anti fake news ».

Conclusion :

Pour conclure, les 3V et 4V sont plutôt des contraintes auxquelles on cherche toujours comment augmenter la vitesse de traitement des données, la sécurisation des données en vue d'empêcher les hackers de voler les données confidentielles, la manière de trier les différentes données et la façon de sauvegarder des informations de qualité en vue d'aider les entreprises ou les utilisateurs.

Glossaire :

Twittos : personne postant un tweet sur twitter.

Fakes news: Fausse information.

Sources:

<http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>

<https://www.linkedin.com/pulse/20140925030713-64875646-big-data-the-eye-opening-facts-everyone-should-know?trk=mp-author-card&trk=mp-author-card>

<https://www.zdnet.fr/actualites/volume-vitesse-et-variete-comprendre-les-trois-v-du-big-data-39841460.htm>

<https://www.culture-informatique.net/cest-quoi-big-data-partie-2/>

<https://www.lebigdata.fr/definition-big-data>

<https://www.marketing-professionnel.fr/tribune-libre/big-data-marketing-info-intox-veracite-donnees-201509.html>

From:

<https://wiki.sio.bts/> - **WIKI SIO : DEPUIS 2017**

Permanent link:

<https://wiki.sio.bts/doku.php?id=1c>

Last update: **2020/07/26 16:27**

